2016

The Impact Of The Choice Of The Item Response Theory Model Used In The Analysis Of Student Response Data From Statewide Educational Assessments

Maureen O'Gorman Petkewich
University of South Carolina

THE IMPACT OF THE CHOICE OF THE ITEM RESPONSE THEORY MODEL USED IN THE ANALYSIS OF STUDENT RESPONSE DATA FROM STATEWIDE EDUCATIONAL ASSESSMENTS

by

Maureen O'Gorman Petkewich

Bachelor of Science
University of Notre Dame, 1994

Master of Science
University of South Carolina, 2003

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

2016

Accepted by:

Christine Distefano, Major Professor

Brian Habing, Major Professor

Tammiee Dickenson, Committee Member

Robert Johnson, Committee Member

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

ACKNOWLEDGEMENTS

I decided to pursue my Ph.D eight years ago while serving as a full-time faculty member in the Department of Statistics and as a parent to two young daughters. I approached this goal with the knowledge that it would be a slow process and that it would require sacrifices from my entire family. I would like to thank my incredibly supportive and patient husband, Matthew, who graciously took care of our daughters when I attended evening classes or worked on my dissertation during weekends. Furthermore, I am grateful to him for impromptu reviews of my writing, technical tips with formatting, and for helping me find both humor and determination during some frustrating moments. I would like to thank my oldest daughter, Madeline, who grew into a bright and capable young woman during the years I worked on my Ph.D. I would like to thank my youngest daughter, Elizabeth, who transitioned from a toddler to a loving and enthusiastic ten year old girl during this time. I appreciate that both of my daughters adapted well to the circumstances, witnesssed the challenges and rewards of continuing education, and continually inspired me to be a worthy role model for them!

I would also like to thank my wonderful EDRM classmates who collaborated with me on numerous course projects, preparation for the comprehensive exams, and the dissertation research process. I thoroughly enjoyed learning with them and from them. I thank those who finished their dissertations before me for inspiring me to complete my degree as well.

Finally, I would like to thank my dissertation committee. I am grateful to Robert Johnson for maintaining high expectations for research and writing as well as asking thought-provoking and insightful questions. I would like to thank Tammiee Dickenson for contributing valuable and practical insight to my research due to her extensive knowledge and involvement in current issues and practices in statewide education. I am especially thankful to my dissertation chairs, Brian Habing and Christine Distefano. Brian guided me though multiple IRT applications and solutions and was subject to my daily spur-of-the-moment questions about research and programming quandaries. Christine helped me to develop and connect ideas, extensively reviewed my writing and was instrumental in helping me to mold my dissertation into a cohesive body of work.

ABSTRACT

The practical significance of the item response theory model (IRT) choice on the results of a statewide assessment was investigated at multiple decision making levels: the examinee level, school and district summary levels, and in terms of impact to subgroups. Data for the study included the student response matrix for South Carolina's 2014 Palmetto Assessment of State Standards (PASS). The Rasch model, used with PASS and in nearly half of PASS-like multiple-choice statewide assessments in other states, was compared to another popular IRT model used in similar statewide assessments: the 3PL model.

Model fit checks indicated that the 3PL had a better person-fit than the Rasch model for PASS. Results centered around the impact of PASS summary scores reported for schools and districts on state and federal report cards showed that for most schools and districts, percentage in PASS performance level and PASS means are largely unchanged by the choice of 3PL or Rasch model. However, for some small schools and districts, the IRT model would have striking effects on percentage in performance level featured on report cards. Furthermore, at the examinee level, examinees near the lower end of the score distribution are sensitive to the change in IRT model. Decisions for some examinees at this level, such as selection for various support programs or even for retention based on PASS scores, might be redistributed due to the change in model. The subgroup with individualized education plans (IEPs) showed the most change because this subgroup, on average, had scores near the lower end of the score distribution. With

regard to grade and subject areas, 8th grade Math, as compared to 3rd grade ELA, 3rd grade Math, and 8th grade ELA, was the most impacted. The 3PL model's estimated guessing parameter was higher for 8th grade math than the other grades and subjects.

In addition to analyzing the student response matrix from the actual administration of PASS, a small simulation study on the most impacted group, the 8th grade Math IEP subgroup, was performed based on the ability parameter and item parameter estimates of the actual examinees. The fit and misfit models accurately estimated the modeled true PASS scores except in the case where 3PL was the true model and Rasch was the misfit model used for estimation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

As members of the Information Age's data rich society, professionals and government officials are increasingly turning to data-based decisions to guide and advance the cause of their organizations. One monumental source of data used to aid decision making in today's society is the data collected from assessments. Assessments in this setting refer to tests administered on a large scale that are designed to evaluate concepts such as ability or aptitude.

Usage of assessments can be found in practically every field, such as, testing to satisfy licensure requirements or to meet admissions criteria for acceptance into professional programs. Most certainly, assessments are used extensively in the field of education. In education, testing is used at many levels such as when a teacher constructs a classroom exam and uses the results to frame his or her own interpretations about student learning and responding course of action. However, large scale assessments are also administered at the state or national level. These tests, and the decisions resulting from them, can have an immense impact on society. Results from large scale educational assessments are used not only to measure student learning but to assess the effectiveness of teachers, principals, schools, districts and states as well. They inform decisions about future curriculum, instruction, and program funding. Thus, large scale educational assessments could be considered high stakes with far reaching effects.

While students have been tested in schools historically, modern assessments go well beyond the scope of classroom testing; schools are charged with preparing students for large scale assessments as demanded by government policy. The implementation of large scale educational assessment was mandated by the the No Child Left behind Act of 2001 (NCLB). The NCLB established that all states would identify statewide annual measurable objectives and administer annual academic assessments which would be used as the chief measure for state and district annual review.

In order to better understand the far reaching impact of the conclusions drawn from these statewide educational assessments, consider one statewide assessment in particular: South Carolina's Palmetto Assessment of State Standards (PASS). PASS scores provide students, parents, and teachers with information about student achievement but PASS results influence many other decisions as well. (Although PASS was selected as an example of a statewide assessment for this study, it should be noted that the purpose of the study is not to examine PASS specifically. Rather, the purpose of the study focuses on the utilization and scoring of statewide assessment results in general with PASS being used as an illustration. While statewide assessments change from time to time, the intent is for the goals addressed in this study to be applicable to any statewide assessment.)

Every year, South Carolina publishes report cards for every school and district in the state to satisfy state accountability requirements. PASS scores are the only achievement data used for the state report cards for elementary and middle schools. PASS results are also used to construct federal report cards for each school and district as well in order to satisfy federal accountability requirements. In addition to satisfying legal

requirements, the report cards can have a strong bearing on the reputation of a school or district.

The school and district report card results are also used to guide school and district curriculum plans. Decisions for school renewal and strategic planning as well as the writing and revision of curriculum draw from report card results in South Carolina. Funding provided for the programs initiated by school renewal and strategic planning is therefore indirectly impacted by the report card results. Schools and districts may also use report card results to request state or federal funding based on low performance or to support other school or district grant proposals.

Standardized test results such as PASS scores are used in the accreditation process as well. AdvancedED is an accreditation agency used in many states including South Carolina. The agency performs a comprehensive internal and external review of a school. The evaluation includes a component on student performance data, including standardized test results, which serves in part to create a quality improvement plan for the school (AdvancedED, 2015). Maintaining accreditation through a respected accreditation agency such as AdvancedED is essential to a school's reputation.

Statewide assessment results such as PASS influence the performance evaluation of principals in South Carolina (SCDE, 2015). Principals are rated in the area of student growth based on statewide assessment results and these evaluations are then used to determine a professional development plan for the principal (SCDE, 2015). In addition, PASS results may be incorporated as part of a teacher's evaluation "score" which could impact improvement plans for the teacher as well.

According to various South Carolina school district officials, PASS scores are a component used in addition to formative interim assessment to identify students for placement in Multi-Tiered Systems of Support (MTSS). MTSS is a three-tier system aimed at improving student academic achievement by providing quality instruction to all students at Tier I, interventions to targeted groups at Tier II and intense intervention to individuals at Tier III (SCDE, 2015).

Furthermore, scores on PASS assessment could lead to retention for 3rd grade students. South Carolina's Read to Succeed Act indicates that a 3rd grade student who scores at the lowest achievement level on PASS "substantially fails to demonstrate third-grade reading proficiency" and is mandated to be retained in 3rd grade beginning in the 2017-2018 academic year (Read to Succeed Act, 2014).

Because the results of state assessment have such a great bearing in many areas of the educational system, it is crucial for states to implement quality testing systems. In fact, NCLB specifies that the quality of the assessments shall be held to "nationally recognized professional and technical standards." Indeed, professional councils have emerged to guide standards for assessment. The American Educational Research Association (AERA), American Psychological Association (APA), and The National Council on Measurement in Education (NCME) are all highly regarded national organizations committed to the implementation of high quality educational assessments. Together, these councils published guidelines that set the bar for educational assessment. The original *Standards for Educational and Psychological Testing* was published in 1966 and the most recent edition was released in 2014. The *Standards* "represents the gold

standard in guidance on testing in the United States and many other countries" (APA, 2016).

The *Standards* recognizes that "Educational and psychological assessments are among the most important contributions of cognitive and behavioral sciences to our society" (AERA, APA, NCME, 2014). Educational assessment results are used to guide educational policy and to make decisions regarding teaching and learning; however, the results are not only used to evaluate the performance of individuals taking the exams but also schools, school districts, states and even nations (AERA et al., 2014).

A key concept in the field of evaluating education assessments is the concept of *validity*. Validity is described as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" and is "the most fundamental consideration in developing and evaluating tests" (AERA et al., 2014). An assessment without evidence of validity for the intended use of its results is not really useful and is potentially damaging. As a very simple example, consider an exam that asks students to define vocabulary words used in a high school Algebra course. The exam may be valid for assessing student knowledge of vocabulary words but not for assessing mathematical reasoning used in Algebra. Interpreting student scores as an indication of mathematical reasoning would be extremely misleading. Validity is a complex concept and the collection of validity evidence draws from many aspects of a testing system. Judging the validity of assessment results is not a matter of concluding that an assessment is valid or not valid but rather an examination of the strength of the validity evidence. Examples of sources supporting validity include evidence of appropriate test content or evidence of appropriate scoring of an assessment. "Ultimately, the validity of an intended

interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system" (AERA et al., 2014).

A major technical aspect of the testing system is the method used to analyze student responses on an assessment. In educational assessment, student ability is often the variable of interest. However, true student ability is a latent trait that is not directly observed or measurable. Instead, measurable outcomes such as student responses on an assessment are used to estimate the unobservable latent trait of interest. Item response theory (IRT) is an approach typically used with large scale assessments to model the relationship between student responses and student ability.

IRT models can be used with various item types but in the statewide testing setting, they are often applied to multiple choice items. IRT uses different models to estimate student ability. But, what influence does the model have on scores? Moreover, does the IRT model impact the critical decisions that are made from assessment results? Could it be that the IRT model selection affects funding provided to a school or which student is selected to participate in targeted programs such as MTSS? Might the IRT model affect state school and district report cards to the extent that strategic renewal planning would differ? A general goal of this study is to investigate methods used in the analysis of data resulting from statewide educational assessments with the intention of acquiring knowledge to increase the likelihood that appropriate conclusions are drawn from assessment results.

Data collected from State Department of Education websites for the 50 states showed that about 60% of state assessments use one type of IRT model to estimate student ability, whereas about 40% of states use a different model. The far reaching

impact of PASS scores for South Carolina has been discussed as just one example of how large scale educational assessments are used in the nation. Meanwhile, the assessment community is split on the selection of the IRT model implemented to analyze student responses on statewide assessments. Collection of evidence supporting the validity of the interpretations of statewide testing data includes examining methods such as the IRT model used to analyze student response data.

Statement of the Problem

Considering that high stakes decisions are based on statewide assessments, it is of interest to investigate IRT model selection. The object of this study is to investigate the impact of IRT models used in the analysis of student response data from statewide educational assessments with the intention of acquiring knowledge to increase the likelihood that valid interpretations are drawn from assessment results.

The study focuses specifically on South Carolina's Palmetto Assessment of State Standards (PASS) 2014 which utilized an IRT model for scoring student response data. However, a different IRT model might be a better fit for the response data. How would ability estimates and PASS scores change if a different IRT model was used and how would this affect decisions made from those scores? More generally, if another IRT model was used, how would the results be impacted? How does the IRT model contribute to the evidence of validity for the assessment?

Research Questions

The study addresses the following questions using recent PASS data for ELA and Math for grades 3 and 8:

1. If a different IRT model were used to score student responses on PASS, how would state school and district reports cards be affected?
2. If a different IRT model were used to score student response on PASS, how would federal school and district report cards be affected?
3. Is the impact of the IRT model different among age groups?
4. Is the impact of the IRT model different among subgroups (including a subgroup of students who received modifications or accommodations)?

CHAPTER 2

LITERATURE REVIEW

The implementation of any large scale assessment, such as those required by NCLB, is an extraordinary execution relying on the expertise and involvement of multiple offices. Typically, content specialists are responsible with ensuring that test item content is appropriate to meet the objectives of the assessment. Psychometricians consider the statistical properties of the test items and appropriate methods for scoring the assessment. Meanwhile, other offices oversee the cost and logistics of administration. This chapter will address various facets of large scale assessment beginning with fundamental concepts including the definition of a latent trait and considerations in the selection of test item formats. Then, different approaches for scoring state-wide assessment will be presented along with their advantages and disadvantages. Background on the concept of validity and the connection between validity and the scoring approach will be established. Previous studies investigating various scoring approaches and the impact of the scoring approach on validity will be included. Finally, the role of these elements in the development and implementation of the PASS statewide assessment will be presented along with proposed research questions.

Psychometric Components

The following sections introduce general elements of large scale assessment that are related to the psychometric functioning of the assessment. These elements contribute to the estimation of a latent trait which is the goal of educational assessment. First, a discussion of the concept of a latent trait is provided. Then, the validity associated with a test designed to measure a latent trait along with the many aspects of validity are reviewed. Next, the roles of test reliability and item formats in latent trait estimation are discussed.

Latent Trait

The principal objective in psychological or educational measurement typically is to measure an unobservable variable of interest. Concepts such as happiness or intelligence, may be considered as examples of unobservable variables. Such concepts are often referred to as a latent trait. In the academic setting and for statewide educational assessments such as PASS, the latent trait of interest is usually student ability in areas measured by content standards (e.g., mathematics, English language arts, writing, science, social studies).

Although aspects of latent traits can be described, the latent variable cannot be measured directly because it is a concept. This is different from physical dimensions. For example, physical dimensions such as distance can be determined in a straightforward manner with the use of a ruler or similar tool (Baker, 2001). Conversely, the approach to measure a latent trait is involved and multi-faceted. For a statewide testing program, this includes many steps such as: operationally defining a construct,

creating a test blueprint, determining test item formats and constructing test items containing content specific to the latent trait. A team of content experts is hired to build a table of test specifications and to write test items. Another team critiques test items. Test items are further reviewed through pilot studies by administering the items to a sample of examinees. Then, the items are examined by psychometricians in terms of their statistical properties and selected to be included or removed from the final assessment. The final assessment is administered to the examinees such as students taking the PASS.

After administration, statistical models are used to relate student performance on the assessment to the latent trait of interest. Student performance is transformed into a scaled score on the assessment. The student's placement on the latent construct (i.e., ability in a given subject area) is inferred based on his or her test performance.

Validity

A primary consideration regarding the estimation of a latent trait, such as student ability on content standards, is that the resulting scores on the assessment are valid for their intended uses. Messick (1995) describes validity as "an evaluative summary of both the evidence for and the actual as well as potential consequences of score interpretation and use." In statewide assessment such as PASS, score usage is widespread and used as a measure of not just student ability but to evaluate teachers, principals, the school, district, and the state as well. There are many facets to the examination of validity and these are discussed in this section.

Historically, types of evidence for validity have been categorized as content validity, criterion validity and construct validity (Messick, 1980). Content validity refers

to the degree to which items on an assessment represent the construct being measured (Crocker & Algina, 2006). Evidence of content validity is typically collected at the time of test development and often relies on the expertise of content specialists. Criterion validity indicates how well the assessment predicts performance behaviors (Crocker & Algina, 2006). Use of the assessment for predictions of future performance is referred to as predictive validity while concurrent validity describes how well the assessment correlates with performance at the same time of the assessment (Crocker & Algina, 2006). A validity coefficient measuring the correlation between the assessment and a measure of a future or concurrent performance is one source of evidence for criterion validity. Evidence for the interpretations of test scores as estimates of a theoretical construct falls under the category of construct validation (Kane, 2009). Messick (1980) finds construct validity as "the unifying concept of validity that integrates criterion and content considerations into a common framework."

According to Messick (1995), construct validity can be further delineated. He warns that the delineation is useful in terms of recognizing the complexities of construct validity rather than an attempt to oversimplify the concept or treat any one of the aspects as evidence of construct validity as a whole. Messick (1995) defines six interrelated areas of validity that are important to consider: content relevance, the substantive aspect, the generalizability aspect, the external aspect, the structural aspect, and the consequential aspect. Content relevance relates to "determining the knowledge, skills and other attributes to be revealed by the assessment tasks" (Messick 1995). The substantive aspect refers to evidence that performance on the assessment reflects engagement of the theoretical processes. The generalizability aspect of validity entails

the generalizability of test scores to the population of interest as well as to tasks and settings. The external aspect encompasses both convergent and discriminant validity. A convergent validity coefficient measures the correlation between an assessment and other measures of the same construct and is expected to be high (Crocker & Algina, 2006). A discriminant validity coefficient measures the correlation between and assessment and a measure of a different construct and is expected to be low (Crocker & Algina, 2006). The structural aspect refers to "the extent to which the internal structure of the assessment reflected in the scores . . . is consistent with the structure of the construct domain at issue" (Loevinger, 1957, as cited in Messick, 1995). Evidence for the consequential aspect of validity includes "rationales for evaluating the intended and unintended consequences of score interpretation" (Messick, 1995).

Reliability

An area related to validity and part of the evidence collected in a validity study is the reliability of the test. Reliability refers to a measure of the reproducibility or consistency of the test scores (Crocker & Algina, 2006). A review of technical reports for statewide assessments indicate that reliability is one of the main statistical properties considered by psychometricians on statewide assessments. Reliability is important because it indicates how much variability can be expected in the test score if the test were repeated. This form of reliability is often called test-retest reliability and is a form of reliability that would be relevant for an assessment like PASS. A test with low reliability would not be very useful because if the test were repeated, a substantially different score might be obtained. It would be very difficult to estimate true ability within a reasonable margin of error on an assessment with low reliability. A simple analogy is a bathroom

scale used to measure a person's weight: a scale that varies greatly in measured weight when a person repeatedly steps on the scale has low reliability, whereas a scale that repeatedly gives the same weight or very close to the same weight has high reliability. A person who is trying to maintain weight would not be well served by a scale that varies by ten pounds, for example, if he repeatedly steps on the scale. It would be very difficult to estimate true weight within a reasonable margin of error on a scale with low reliability. Clearly, an assessment that does not yield reliable scores would not be valid for most interpretations, especially those used in statewide assessments.

This study does not focus on reliability directly but is related because the study compares two different measurement models for scoring an assessement. The method for determining the reliability of an assessment depends on the measurement model.

Item Formats

As mentioned previously, the administration of statewide tests is an extremely complex operation. Beyond expert input and review regarding the content validity of the assessment and item selection based on statistical properties such as reliability, there are financial and logistical challenges as well. Content experts must be trained on item writing and potentially grading rubrics as well. Also, a very large number of students must be tested in a relatively short amount of time. Given that schools and districts depend on the results for decision making and potentially funding, there is a demand for a fast turnaround of results. The type of item format utilized is a factor in all of these areas: content validity, statistical properties, timeliness, and cost.

There are two main types of item formats used in large scale assessment: constructed response (CR) and multiple choice (MC). For MC items, examinees select an answer from a list of available options, where only one option is the correct answer. The format makes the scoring of MC items clean and clear; the response is scored as either correct or incorrect. This type of scoring is called objective scoring because a rater does not have an effect on the score (Haladyna, 2004). Objective scoring can be performed inexpensively using machines, score templates, or an untrained observer (Haladyna, 2004).

For CR items, examinees are presented with an item stem and then must construct an original response (typed or handwritten). Given that an original response is provided from each student, the grading of CR items is more involved. Rubrics are needed to judge what is acceptable for a correct response. This process is known as subjective scoring and requires human graders, training, and consideration of partial credit. Even with thorough training, human graders may arrive at different scoring decisions resulting in a potential threat to the structural aspect of construct validity called rater effect.

For statewide testing programs, MC has obvious advantages over CR; it saves time and money. MC can be graded quickly with the aid of technology and saves the cost of training and paying human graders. Also, examinees can answer MC more quickly allowing an increased number of items on the exam. With an increased number of items on the exam, more content can be covered allowing more comprehensive coverage of the domain defined by the latent trait of interest (Lissitz & Hou, 2012). Longer test also have an advantage in terms of statistical properties: tests with more items have higher test reliability (Crocker & Algina, 2006).

Though MC has advantages, there are concerns over the usage of MC items. Campbell (as cited in Lissitz, Hou & Slater, 2012) noted that MC does not generally tap into higher order thinking processes. This issue is of particular concern when the construct of interest is of an abstract nature such as writing ability where an MC format may not provide a high-fidelity measure of the construct. Fidelity refers to the plausible connection between the criterion and the criterion measure (Haladyna, 2004). Clearly, if writing ability is the construct of interest, it is more plausible to judge writing ability by actually requiring the examinee to write than to have the student respond to multiple choice items. However, if two items have strong proximity (a measure of the relationship between two items with varying fidelity), it is practical to choose the item format that measures more efficiently (Haladyna, 2004). For example, if we can show that the responses to the multiple choice questions representing writing ability can predict how well the examinee can respond to a writing prompt, then it is more practical to utilize the multiple choice format because it is more efficient to score. A cocern though, with this approach, is that curriculum might be shifted to focus on writing skills rather than direct writing (Haladyna, 2004).

Another concern regarding MC items is the opportunity for guessing because the correct option is presented to the student along with distractor options. With MC, guessing may lead to lower reliability for lower ability students (Cronbach, 1988). Furthermore, guessing introduces a threat to validity known as construct-irrelevant variance resulting from a tendency to respond to an item in a manner that is unrelated to the interpreted construct (Messick, 1995). The resulting estimate of ability for the latent trait of interest will be contaminated by variation produced by the effects of guessing. A

discussion of scoring methods and the treatment of guessing utilized by those methods will be addressed later in the chapter.

Regardless of some of the disadvantages, MC remains a very popular format for large scale tests. Many statewide assessments, including South Carolina's PASS, continue to use the MC format for most subject areas. For the 2014 administration of PASS, MC items were used exclusively for Math and ELA subject area exams. Ultimately, the goal of statewide assessments such as PASS is to measure student ability in various subject areas. The following sections provide background on utilizing MC items for measuring student ability.

Measurement Models

As discussed previously, latent traits such as student ability, cannot be measured directly. Instead, an assessment is constructed with content specific to the latent trait with care taken to provide evidence of content validity. *Measurement* of the latent trait occurs when a quantitative value is given to the sample of results collected from the assessment (Crocker & Algina, 2006). There are many challenges to the latent trait measurement process including the following: results from the assessment only provide a sample of the student performance in the content area, there will always be some degree of error in the measurement even with reliable assessments, and a scale must be constructed for the latent trait of interest (Crocker & Algina, 2006). A measurement model provides a statistical approach for latent trait estimation and addresses challenges presented by the measurement process. The traditional measurement model of Classical Test Theory and the more modern approach of item response theory are discussed in the

next sections. Both of these models have been used historically with the multiple choice item format which is often used in large scale assessment.

Classical Test Theory and Classical Item Analysis

Historically, classical test theory (CTT) was the statistical method used in educational measurement to analyze student response data. In CTT, the term "true score" is typically used as opposed to the term "latent trait". In CTT, a linear relationship is used to model the relationship between the true score and the total observed score (total number correct) on the exam. The model has the form,

$$X_j = T_j + E_j \tag{2.1}$$

where X_j is the observed sum score, T_j is the true score and E_j is the random error for examinee j. The true score T_j, can be thought of as the mean observed score obtained by examinee j on the assessment if the assessment was repeated a large number of times. CTT has many desirable properties: it is mathematically simple, conceptually uncomplicated to understand, and it relies on minimal assumptions making the model largely useful in practice (Le, 2013).

Although the classical test theory model has no item level statistics, classical item analysis is often applied in conjunction with classical test theory. Classical item analysis measures item difficulty by the proportion of examinees who answer the question correctly. Item discrimination, which refers to the capability of an item to distinguish between low and high ability level students, is measured by the correlation between the item score and the total test score (Abedalaziz & Leng, 2013). Classical item analysis has limitations; these item measures depend on the sample of examinees and do not

characterize properties of the test (Le, 2013). Also, examinee scores depend on the test because examinees may achieve better scores on less difficult exams and lower scores on more difficult exams (Le, 2013). Le (2013) also points out that with CTT, test items cannot be linked with ability levels. "The major limitation of CTT can be summarized as circular dependency: (a) The person statistic (i.e., observed score) is (item) sample dependent, and (b) the item statistics (i.e., item difficulty and item discrimination) are (examinee) sample dependent" (Abedalaziz & Leng, 2013).

This dependency can be problematic because item parameters change depending on the sample of examines taking the test which would make it difficult to create equivalent testing forms in a large scale testing situation. Also, because CTT focuses on test level information as opposed to item specific information, it is difficult to select individual items from the test to construct other assessments aimed at certain ability groups (Abedalaziz & Leng, 2013). Finally, ability estimates are determined by the particular test and therefore ultimately determined by the group of examinees taking the test. This issue leads to concerns of reliability because the ability estimates would change in repeated administrations of the exam (Abedalaziz & Leng, 2013). Finally, ability levels of student responding to different items cannot be compared.

Modern Measurement

More recently, measurement models have been developed that overcome the major limitations of CTT. These mathematical models are grounded with strong test theory and vigorous assumptions. They provide measurement that is free of sample or examinee dependency. Modern methods, such as item response theory (IRT) focus on

analyses at the item level; the approach allows examinees to be compared even if they take different tests and also for the item analysis to be relevant to examinees with different ability levels than the examinees used for the item analysis (Crocker & Algina, 2006).

Item Response Theory

Unlike CTT, item response theory (IRT) focuses on the responses to individual items on the assessments instead of the raw score or sum of correctly answered items. While IRT models can accommodate a variety of item formats, binary MC items are usually used with IRT; the item is either marked correct and scored as a "1" or marked incorrect and scored as a "0". These types of items are known as producing dichotomous or binary data. IRT uses a probability model to relate item responses to the latent trait. The general form of the probability model is given below where *e* is the base of the natural logarithm.

$$P_i(\theta) = \frac{e^x}{1+e^x} \tag{2.2}$$

Here, $P_i(\theta)$ represents the probability that an examinee with ability θ will answer item i correctly. A distribution is established for θ, typically with a mean of 0 and a standard deviation of 1 resulting in a general range from -3 to 3. Lower levels of ability correspond to a smaller probability of answering the question correctly and higher levels of ability correspond to a higher probability of answering the question correctly.

Item Parameters and the Item Characteristic Curve

A graphical representation of the IRT logistical function, known as an item characteristic curve (ICC; or item response function, IRF) shows the probability of a randomly selected examinee from a subpopulation of examinees with the same ability correctly responding to an item. The mathematical model for the ICC is given by formula 2.5 and is discussed in a later section. The ICC is defined by three parameters. The item difficulty parameter is denoted by *b* and is the level of ability at the inflection point on the curve. For example, if the inflection point occurs at $b = 2$ and probability = .50, this means that 50% of the population of all examinees with an ability level of 2 can answer the item correctly. Item discrimination, denoted by *a*, measures how well the item distinguishes between students with an ability level below the item difficulty versus those with ability level above item difficulty. Item discrimination is proportional to the slope of the curve. Items with steeper slopes discriminate better than items with less steep slopes. The third parameter on an ICC is the guessing parameter, denoted by *c*, which is the lower asymptote of the curve. The lower asymptote shows the probability that low ability students will answer the questions correctly just by chance. The inflection point on the curve occurs at $(1 + c)/2$ on the probability scale. Figure 2.2 provides an illustration. There is some debate over the inclusion of the guessing parameter in IRT models as well as allowing the item discrimination parameter to vary; this discussion will be addressed later in the chapter.

Lord (as cited in Crocker & Algina, 2006, p. 340) "specifically recommends against interpreting the probability of responding correctly as the probability that a specific examinee answers a specific item correctly." Instead, the probability of

answering correctly refers to the probability that a randomly selected individual from a subpopulation of examinees with the same ability will answer a specific item correctly (Crocker & Algina, 2006). Alternatively, the probability can be interpreted as the probability that a specific examinee correctly answers a randomly selected item from a subset of items with the same difficulty level. The significance of the interpretations is that with IRT, examinees can be compared even if they do not encounter the same items provided that the items are addressing the same latent trait. This desirable property is

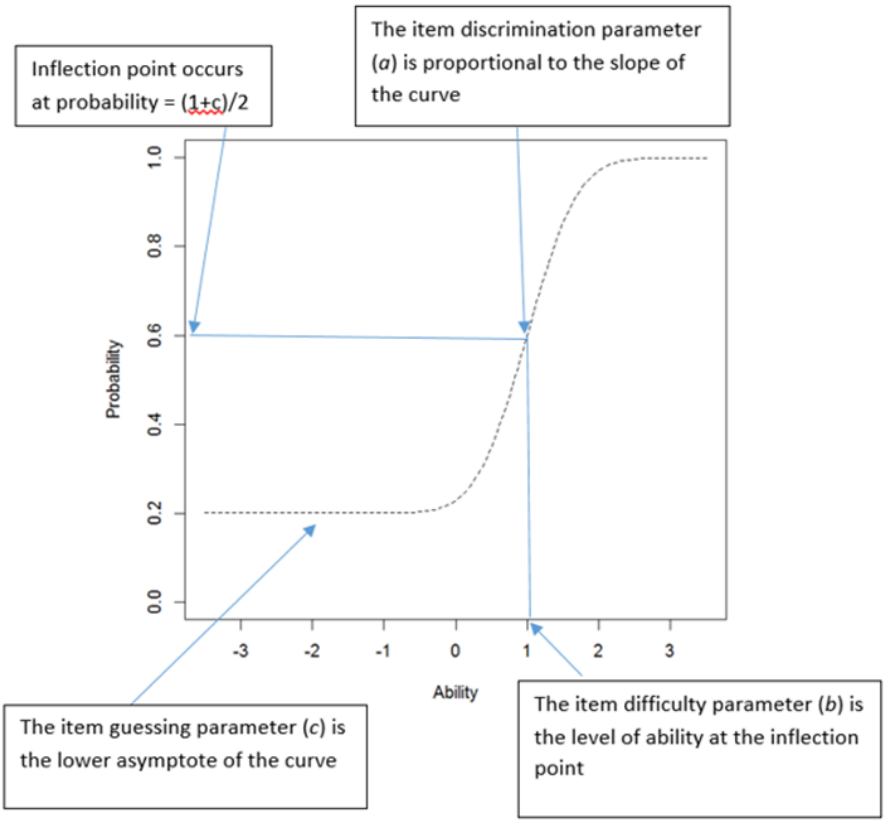


Figure 2.1. Relationship between ability and probability of correctly answering an item

called test-free measurement. Recall that this is not the case with CTT, as ability levels are test dependent; therefore, examines who do not respond to the same items cannot be

compared. Additionally, in an IRT framework, item parameters are independent of the sample and this is known as person-free item calibration (Crocker & Algina, 2006). This allows items to be more easily utilized in terms of constructing equivalent forms or creating assessments geared at specific ability groups.

Assumptions in IRT

Unlike CTT, most IRT models are limited by three major assumptions unidimensionality, local independence, and monotonicity. Unidimensionality means that all items on the assessment measure the same latent trait. In PASS testing, unidimensionality means that the Math assessment for 3rd graders, for example, only measures ability on the 3rd grade Math standards and not other abilities. An exam question on the Math portion of PASS that required a high ability in reading due to complex vocabulary or sentence structures, for example, would be a violation of the unidimensionality assumption because the question measures the second dimension of reading ability.

Secondly, IRT models are based on laws of probability and mathematically assume local independence between test items. Local independence means that after conditioning on the latent trait, performance on one item of the exam is independent of performance on another item on the exam. In other words, if a student answers question 1 correctly, the probability of her answering any other question on the exam does not increase or decrease after conditioning on the latent trait. This property is important because the framework of IRT focuses on the information obtained from responses to

individual items on the assessment and this structure would be contaminated if responses on one item influence responses on other items. There are methods available for checking that the unidimensionality and local independence assumptions are upheld though some experts question the reality of the assumptions fully being met in practice.

A third assumption for IRT models is the monotonicity assumption. Monotonicity means that as ability level for the latent trait increases, the probability of correctly responding to the item measuring the ability increases:

$$\theta_1 > \theta_2 \rightarrow P[U_i = 1|\theta_1] > P[U_i = 1|\theta_2] \quad (2.3)$$

Unidimensionality, local independence and monotonicity are the three assumptions of IRT models.

Scaling, Calibration, Equating and Scoring with IRT Models

As indicated in many statewide assessment technical reports, the IRT model is utilized for scaling, calibrating and equating (or linking). Scaling is a broad term that refers to transforming values to a common scale. IRT models are used to scale examinee abilities. This means the model estimates student ability, typically denoted as θ, as a location on the theoretical latent trait scale which usually ranges from -3 to 3. The IRT model is also used to estimate item parameters such as item difficulty, item discrimination and item guessing, if applicable. The item difficulty parameter is placed on the same scale as examinee ability. Estimating the items parameters is part of the scaling process but is often referred to as item calibration.

Scaling can also be used to equate two test forms so that the resulting scores are on the same scale. While the two test forms may be similar, one form may be slightly more difficult and a transformation is necessary for a fair comparison in a high-stakes assessment. One approach for equating is to employ a set of items that appear on both forms of the exam to serve as a common basis for the equating process. These items are referred to as the anchor test items. Anchor test items were utilized in PASS testing to equate test forms from one year to another.

After examinee ability is estimated on the latent trait scale, it is transformed to a more readable and reportable score for the particular assessment. On the SAT, for example, the ability estimate will be transformed to a reported score on the SAT scale, somewhere between 200 and 800 for one subject area.

The next two sections describe the two most popular IRT models used for scaling, calibration and equating in statewide assessment: the Rasch model, and the 3PL model.

The 1PL Model

The Rasch model is mathematically equivalent to the most basic IRT model, the one parameter logistic (1PL) model. The 1PL model is given by the following formula

$$P_i(\theta) = \frac{e^{Da(\theta - b_i)}}{1 + e^{Da(\theta - b_i)}} \qquad (2.4)$$

Here again, $P_i(\theta)$ is the probability that an examinee with ability θ will answer item i correctly. For the 1PL model, the item difficulty parameter, b_i, varies for each item. The item discrimination parameter, a, is the same for each item. D is a constant typically set

to 1.7 or 1 so that "$P_g(\theta)$ for the normal and logistic ogives will not differ by more than .01" (Lord &Novick, 1968, p. 399).

An important property for the 1PL model is that the total score is a sufficient statistic for the latent trait of interest. That is, all of the information regarding the ability of the examinee is contained in the total score. We will later see that in the other logistic models used in IRT, the total score is *not* a sufficient statistic. Information about ability is obtained from the pattern of responses in other IRT models.

Distinction between Rasch and 1PL

While the Rasch model (Rasch, 1960) is *mathematically identical* to the 1PL model, the two models have completely *different philosophical approaches*. The logic behind the 1PL model is that it should only be used if the model fits the data. If the model does not fit, and any very egregious items have been removed, then consider a different model. Meanwhile, the approach with Rasch is that student response data that is appropriate for educational measurement should fit the Rasch model. The Rasch model is viewed as an ideal measurement model and in practice, a means for determining if a data set has met ideal measurement requirements (Engelhard, 2013). Rasch is a very popular model used in statewide assessments such as PASS. More comprehensive viewpoints on the Rasch versus the logistic models will follow this basic introduction of the models.

The 3PL Model

Another very popular IRT model used in statewide assessments is the three parameter logistic model (Birnbaum, 1968). Like the 1PL, the three parameter logistic model (3PL) model includes the difficulty parameter b_i for each item. However, on the 3PL, the discrimination parameter, a_i, varies for each item. Finally, the 3PL includes a third parameter, c_i, which accounts for guessing. The mathematical model is given by equation 2.5:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1+e^{Da_i(\theta - b_i)}} \quad (2.5)$$

As stated previously, the 3PL model obtains information about examinee ability based on the pattern of responses as opposed to just the total score.

Rasch Model History

During the 1950s, a Danish mathematician named Georg Rasch became involved in the field of psychometrics and through his work with reading assessment data, discovered a probabilistic function that enabled separation of the text parameters used in the assessment from person parameters (Fischer, 2007). This separation allowed for the difficulty level of the texts to be compared independently of the examinees and also for the ability levels of the examinees to be compared independently of the difficulty level of the text (Fischer, 2007). Rasch developed a concept he called "specific objectivity" based on the idea of invariant comparisons between items and persons (Fischer, 2007). Guided by this concept, Rasch formulated a probabilistic formula for latent traits known as the Rasch model (Fischer, 2007). Rasch measurement models, based on a quest to

achieve invariant measurement, are viewed by many theorists as an ideal type of measurement (Engelhard, 2013).

Engelhard describes invariant measurement as a measurement process that upholds these five requirements: the measurement of persons is independent of the items used to measure the person, the measurement of items is independent of the people responding to the items, persons with higher ability are more likely to respond correctly to items than persons with lesser ability, any person is more likely to correctly respond to an easy item than to a more difficulty item, item difficulty and person ability must be measured on the same scale.

The properties are very important in high stakes assessment because student abilities can be compared regardless of which items the students responded too and also, item difficulties can be compared regardless of the sample of students who took the assessment.

Furthermore, the invariant measurement properties of Rasch create a data structure where the total score is a sufficient statistic for student ability. That is, the total number of correct answers contains all of the necessary information to estimate an examinee's ability. In other models the response pattern adds information to the ability estimation so two examinees with the same total score may have a different estimate of ability based of the pattern of their correct answers which can be difficult to interpret for laypeople. Therefore, the total score as a sufficient statistic is very attractive to practitioners because it provides a simplistic interpretation of scores which is easier for stake-holders to understand. With the invariant measurement properties allowing for easy comparisons among items and persons and also more interpretable scores, the Rasch

model has remained very popular and is still implemented in a very large number of statewide assessments.

Rasch Philosophy

One of the key principles with Rasch measurement is that assessments are constructed to meet these ideal properties of invariant measurement. The model is determined a priori and the data structure is expected to fit the model (Engelhard, 2013). This is achieved in part by analyzing the psychometric properties of assessment items in advance with Rasch item fit statistics. Fit statistics identify assessment items that are functioning appropriately and also items that produce response patterns anomalous to the required data structure. Items that are identified as aberrant are then carefully reviewed and either discarded or modified. The review process continues until all items generate a data structure that fits the Rasch model and achieves invariant measurement. This process is quite different from statistical approaches that will be discussed later where the model is selected based on the data structure. With Rasch, the model receives priority.

A recent study on the application of Rasch compared the advantages of the Rasch model over CTT for obtaining information about examinations used in an Anatomy course (Royal, Gilliland, & Kernick, 2014). Royal, Gilliland, and Kernick recognize that sophisticated models such as Rasch have been commonly used in high stakes assessments but not often applied in the classroom setting. They indicate that for exams with moderate implications for examinees such as the Anatomy assessment, CTT is most often used. However, modern technology makes IRT software readily available to instructors and the authors explored how utilizing Rasch analysis might improve the psychometric functioning of the Anatomy assessment.

One of the outcomes of the study included a finding that 10% of the 69 items were detected by Rasch fit statistics as potentially too easy or susceptible to either guessing or careless mistakes. Thus, these items would not fit the data structure required by Rasch. Furthermore, the results of the Rasch analysis provided a variable map enabling instructors to make a connection between person abilities and item difficulties since these properties were measured on the same latent trait scale. They could then compare the content of items to the measured item difficulty to determine if the results were logical. For example, if an instructor noticed that an item that appeared easy in terms of content but registered as highly difficulty on the analysis of responses, then the item should be reviewed. This type of result might also inform teaching. In general, Royal et al. found that the Rasch analysis was more useful than CTT as CTT is limited by sample dependency results that are potentially distorted and irreproducible. Rasch transcended these limitations and provided an opportunity "to produce examinations that are both fair for students and capable of producing valid and reliable scores that are legally defensible" (Royal, Gilliland, & Kernick, 2014).

3PL Background

Another popular IRT model was introduced by American Statistician Allan Birnbaum. Birnbaum introduced the approach of employing a cumulative logistic distribution model to describe the relationship between items and responses (Lord, 1980). One form of the model contains an item difficulty parameter and an item discrimination parameter. With these two parameters, the model is referred to as the 2 parameter logistic or 2PL model. However, a lower asymptote was added to the item characteristic

curve which came to be known as the "guessing" parameter (Lord, 1980). The guessing parameter represents the likelihood that an examinee with low ability will answer the item correctly (Lord, 1980). This model is sometimes referred to as the "Birnbaum" model but more typically is called the 3 parameter logistic, or simply, the 3PL model. The mathematical model is given by formula 2.5:

$$P_i(\theta) = c_i + \frac{(1 - c_i)e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

With the 3PL model, the focus is on the response pattern of the examinee rather than simply the total score.

3PL Philosophy

The approach with 3PL is to use the 3PL model if the model is the best fit to the response data. Recall that this is opposite of the approach with the Rasch model where the model dictates. Proponents of the 3PL advocate selecting the statistical model that is the best fit to the data. Many testing agencies firmly believe in this strategy and the 3PL model is also widely used in high stakes assessment.

For example, CTB/McGraw Hill is contracted by many state agencies to analyze high stakes statewide assessments. The company explains that the accuracy of test scores depends on selecting a model that best explains the relationship between ability and item responses (CTB McGraw Hill, 2008). This relationship is impacted by the reality that guessing takes place in the real world. "Empirical evidence indicates that students guess on multiple-choice items that they find too difficult or do not have the motivation to consider carefully" (Lord, 1980 as cited in CTB McGraw Hill, 2008).

CTB McGraw Hill (2008) also explains that when items vary in their discrimination level, the addition of a discrimination parameter increases the accuracy of the information obtained from the tests. Furthermore, item pattern scoring generates more accurate scores for examinees than total correct scoring (CTB McGraw Hill, 2008). The company asserts that it makes sense that more information can be obtained from looking at which items an examinee answered correctly rather than simply how many items he got right.

In addition, CTB McGraw Hill (2008) points out that the Rasch or 1PL model is a special case of the 3PL model. A 3PL model with a guessing parameter equal to zero and an item discrimination parameter equal to 1 is the equivalent of the 1PL model. Therefore, if the Rasch model fits, the 3PL model will take on the Rasch model form. In this sense, the 3PL has the capability to take on the advantage of the Rasch model in terms of having the total score as a sufficient statistic when the data fits the model.

3PL Advantages/disadvantages

The 3PL model has the advantage of flexibility; it is used as a model to adapt to the fit of the data. Its parameters can change such that it becomes the 1PL model if that is the best fit for the data. It also takes into account the reality of student guessing and the reality of items discriminating differently from each other. By including these additional parameters, the 3PL model can produce better estimate of student ability.

While the flexibility of the 3PL is viewed as advantageous to statisticians, many measurement theorists find its flexibility to be misguiding. Those who oppose the 3PL suggest that data structures that require a guessing parameter result from poorly worded

items. Opponents also maintain that an assessment with items varying in discrimination levels are the result of unintended dimensions and violate the assumption of unidimensionality. Furthermore, it can be difficult to fit and interpret the 3PL model.

Rasch versus 3PL Debate

Psychometricians have debated the use of the Rasch model versus the use of the 3PL model in the analysis of assessment data for more than two decades. The American Educational Research Association (AERA) Annual Meeting included a debate on the topic between Ron Hambeton and Ben Wright back in 1992.

Wright (1992) defended the Rasch model explaining that the Rasch model was "derived to define measurement". "Rasch is the one who made the deduction of the necessary mathematical formulation and showed that it was both sufficient and necessary for the construction of linear, objective measurement" (Wright, 1992). With the Rasch model, the total score is a sufficient statistic to estimate student ability. The Rasch model does not allow for item discrimination or guessing. "In practice, guessing is easy to minimize by using well-targeted tests" (Wright, 1997). Item discrimination is viewed as a result of item bias (Wright, 1992). With the Rasch model, if the data does not fit the model then the solution is to get better data (Wright 1992). "The Rasch model is derived a *priori*, to define the criteria which data must follow to qualify for making measures" (Wright, 1992). Wright (1992) explains that the philosophy with the 3PL model is exactly the opposite: "The Birnbaum (3PL) model has loose standards . . . because it's adjusted to adapt to whatever strangeness there is in the data."

Arguments for the 3PL model are based on empirical results rather than on theory. Lord (1980), in a study on the verbal section of the College Board Scholastic Aptitude Test, found that "in actual practice, low-level examinees do less well than if they responded at random . . . at low ability levels the effect of random guessing becomes of overwhelming importance."

Several studies have been conducted comparing the fit of the logistic models to determine which fits better in reality. For example, Bergan (2010) conducted a study assessing the fit of different IRT models to data illustrating the "empirical approach to model selection" with the goal of selecting the IRT model to fit the data being analyzed. Bergan (2010) examined data from a 5th grade math assessment administered to 3098 students and employed a chi-squared test comparing the fit of the 1PL, 2PL and 3PL models. Conclusions showed that the 3PL fit the data better than the 1PL and also provided more information about the items by providing an estimated guessing parameter and item discrimination parameter for each item (Bergan 2010).

Another study by Jiao and Lau (2003) conducted a simulation study to determine the impact of employing a misfit IRT model on a computerized classification test. Jiao and Lau (2003) simulated 1PL, 2PL and 3PL data and then examined the data with a misfit model. The results were analyzed to determine when examinees were not classified correctly as passing or failing the exam. (Because the data was generated with known simulated abilities, it could be determined when the misfit model placed the examinee in the correct or incorrect category of passing or failing. Simulated examinees placed into the wrong category were considered false classifications.) Conclusions indicated that when the 1PL was the true model and the 2PL or 3PL model was

employed, the error rates were not too serious. However, when 3PL was the true model, using the 1PL model had a large impact on false classification rates (Jiao and Lau, 2003). This result suggests that the 3PL model is a safer choice than the 1PL model if model fit is in question.

Study relating Validity to IRT Method

Beyond looking at model fit, other researchers have explored how model selection affects scores overall for statewide assessment. Sinharay et al. (2014) analyzed data from a state assessment for three subject areas using the 1PL model that was originally used to equate a new form to an old form and also a restricted 3PL model. Sinharay et al. defined a restricted 3PL where the guessing parameter is constant for all items.

Using a generalized residual analysis method based on residuals falling outside of a confidence band, their results show that neither model is a good fit for the data but that the 3PL is better fit than the 1PL. The authors clarify that misfit in general is not surprising in the study due to the large sample size and thus narrow confidence bands. The study examines the practical significance between the two models by determining the disagreement in student classification as proficient versus not proficient between the two models. For two subject areas they found no disagreement but in one subject area, 2.4% of students changed classifications based on the model. Sinharay et al. point out that although many researchers have investigated model fit, studies regarding the practical significance of model misfit are rare. Meaning, few researchers have examined the effects of the IRT on real assessment data to determine the practical impact on resulting

scores. Furthermore, the authors recognize that "assessment of the practical significance of misfit may involve several layers of analysis." The article provides the example of the results of a teacher certification exam: beyond the effect of the model on simply the pass/fail outcome of the assessment, how will the success of students of those teachers be affected? They explain that "*assessment of the practical misfit is a never ending process . . . similar to validation that is a never ending process (e.g., Messick, 1980).*" Sinharay et al. call for further studies investigating the practical misfit of IRT models and specify that their study did not consider the effect of the IRT model on pattern scoring which may have more significance than the effect of the IRT model on equating.

IRT and the Modern Framework of Validity

As Sinharay et al. pointed out, there is a connection between the ongoing study of model misfit and the continuous collection of validity evidence. Messick (1996) explains that evidence of validity is never complete but rather a means of constructing the most plausible case to inform the usage of modern assessments and to guide understanding of what test scores mean.

Evidence for validity comes in many forms and one source is the selected IRT method used to score the assessment. Because the IRT model that is utilized affects the score, and many decisions are made on score interpretations for high stakes assessment, the IRT model ultimately impacts decisions made from high stakes assessment. The concept of validity revolves around appropriate interpretations and usage of assessment scores and therefore the IRT model contributes to validity evidence.

Usage of statewide assessments results such as PASS have been discussed previously and include reviews of the principal, school, district and state. They provide a basis for informing teaching and identifying students for targeted programs. Ideally, the assessment results would appropriately guide these decisions and contribute positively to both teaching and learning. These types of positive results would support the consequential aspect of construct validity (Messick, 1996). The concern though, is that sources of test invalidity can produce adverse consequences and result in a negative impact on individuals or groups (Mesick, 1996). This type of validity evidence is sometimes referred to as consequential validity.

It is logical to conclude that an IRT model that produces the best estimate of student ability would support valid interpretations of test scores. Meanwhile, a poorly selected IRT model would contribute to misleading interpretations of test scores and potentially adverse consequences for individuals or groups. However, as Sinharay et al. explained, little research has been conducted to examine the role of the IRT model on practical consequences in high-stakes exams.

In summary, we know that large scale statewide assessments have high stakes implications. Collection of validity evidence is an ongoing process especially in the high stakes setting. Validity is broad concept and evidence of test validity draws from all aspects of an assessment including the IRT model used to calibrate and scale the assessment. We know that there are two popular IRT models that are widely used: the Rasch model and the 3PL model. Rasch theorist advocate the approach of developing an assessment that is fit well by the Rasch model; such an assessment would represent sound measurement. However, simulation studies have shown that the 3PL model fits Rasch

data better than the Rasch fits 3PL data if a misfit model is used. We know that more research is needed on the practical impact of using one model over another. A recent study explored the practical impact of the IRT model used for equating on a statewide assessments at the state level overall. However, we do not know what the practical impact of the IRT model is on large statewide assessment results *especially at the level where many decisions are made*: the school and district level. We do not know if there are potential consequences to decisions made at the school and district level based on the IRT model.

This study will continue research on the contribution of the IRT model to consequential validity evidence in high-stakes assessment and focus on South Carolina's PASS assessment. In order to better understand the setting for the research, the next section provides further details on the development, usage and technical aspects of the PASS assessment.

Laws surrounding PASS and Major Uses

According to the *Technical Documentation for the 2012 Palmetto Assessment of State Standards or Writing, English Language Arts, Mathematics, Science, and Social Studies (2012)*, PASS was established to satisfy the requirements of The Education Accountability Act of 2008 (EAA). The South Carolina Code of Laws Section 59-18-110 describe the objectives of the accountability system mandated by the EAA.

The system is to:

(1) use academic achievement standards to push schools and students toward higher performance by aligning the state assessment to those standards and linking policies and criteria for performance standards, accreditation, reporting, school rewards, and targeted assistance;

(2) provide an annual report card with a performance indicator system that is logical, reasonable, fair, challenging, and technically defensible, which furnishes clear and specific information about school and district academic performance and other performance to parents and the public;

(3) require all districts to establish local accountability systems to stimulate quality teaching and learning practices and target assistance to low performing schools;

(4) provide resources to strengthen the process of teaching and learning in the classroom to improve student performance and reduce gaps in performance;

(5) support professional development as integral to improvement and to the actual work of teachers and school staff; and

(6) expand the ability to evaluate the system to conduct in-depth studies on implementation, efficiency, and the effectiveness of academic improvement efforts.

Section 59-18-310 of the EAA calls for the Department of Education to

develop or adopt a statewide assessment program to promote student learning and to measure student performance on state standards and:

(1) identify areas in which students, schools, or school districts need additional support;

(2) indicate the academic achievement for schools, districts, and the State;

(3) satisfy federal reporting requirements; and

(4) provide professional development to educators.

PASS was implemented as the statewide assessment program to address the objectives of the EAA with PASS test results serving as the bases for local, district and state accountability (SCDE, 2012).

PASS Development

The *Technical Documentation* (2012) states the development of PASS included input from an Education Oversight Committee (EOC) which included members from state government, business and educations. According to the report, the EOC evaluated PASS for alignment with state standards, assessed the level of difficulty, reviewed the assessment for evidence of content validity, and determined achievement standards. The documentation notes that the Technical Advisory Committee (TAC), comprised of local, state and national specialists, who advised the South Carolina Department of Education (SCDE) on technical issues including the IRT model. An outside contractor, Data Recognition Corporation (DRC) provided test administration, scoring and reporting services (SCDE, 2012).

According to the *Technical Documentation* (2012), items selected for PASS underwent extensive content review by content experts including teachers and curriculum specialists as well as a sensitivity review by social service agency staff. These types of reviews support content validity of the assessment. Additionally, the report indicates that items were field tested and statistics were collected regarding item difficulty level and item discrimination. Presumably, this analysis included the analysis of Rasch fit statistics. The documentation notes that items were reviewed for differential item functioning (DIF) between ethnic groups and gender. Content and technical advisors then determined if items were accepted or rejected as PASS items or potentially modified and field tested again (SCDE, 2012).

Determination of PASS Scores

The *Technical Documentation* (2012) indicates that the Bookmark method (Lewis, Mitzel & Green, 1996) was used to determine performance standards for PASS. To employ the Bookmark method, the standards setting committee was provided with an ordered item booklet (OIB) containing test items in order of increasing difficulty. With the Bookmark method, items are typically ordered by item difficulty as measured empirically through IRT calibration (Karantonis & Sireci, 2006). The ordering facilitates comparison of items and the selection of items that would likely be answered correctly by examinees at different proficiency levels (Karantonis & Sireci, 2006). Research supports utilizing a correct response probability of .67 as the measure of whether the student will likely answer the question correctly (Karantonis & Sireci, 2006).

Bookmarks were inserted by committee members between items that divide achievement levels and several rounds of judgements were made before reaching a consensus (SCDE, 2012). Achieving agreement among judges for the cut scores is a source of internal validity for the PASS assessment. The *Technical Documentation* defines the achievement levels as follows:

> **Not Met –** the student did not meet the grade level standard,
>
> **Met –** the student met the grade level standard, and
>
> **Exemplary –** the student demonstrated exemplary performance in meeting the grade level standard.

Cut scores for the performance levels were then translated to a Rasch ability scale (SCDE, 2012). The cut scores for Rasch ability are the same from year to year.

Rasch abilities for the examinees are found empirically each year and then translated to the PASS scale score (SCDE, 2012). This means that the Rasch model is applied to student response data each year to determine the estimate of student ability, called Rasch ability, on the theoretical Rasch ability scale. Finally, the Rasch abilities are converted to a PASS scale that is easier to read and report than Rasch ability. The PASS scale ranges from 300 to 900. The PASS scale score and proficiency level for each subject are reported for examinees.

Consideration of another IRT Model for PASS

As discussed earlier, the Rasch model does not account for item discrimination or for guessing. Perhaps, item selection for the assessment successfully removed all items

that would be subject to guessing or that would pose a high level of item discrimination. However, due to the call for continuous collection of validity evidence for high stakes assessment, it is of interest to investigate the impact of using another IRT model to estimate student ability and consequently on the many areas affected by PASS scores such as state and federal report cards.

Summary

Statewide education assessments are high stakes exams and the results of these assessments are used not only to evaluate students individually but to evaluate schools districts and states as well. States are mandated by law to implement quality annual assessments measuring academic standards and to report the results of the assessment on annual report cards. The results influence many decisions regarding curriculum, professional development, funding and placement of students in targeted programs.

The administration of statewide assessment is complex and multifaceted. Multiple governing bodies collaborate to determine appropriate test items formats and appropriate item content while considering financial and logistical demands. The collection of evidence for valid interpretations of test scores begins with the development of the assessment and continues indefinitely with decisions made from test results having far reaching and long lasting effects. There are many elements in the collection of validity evidence.

One source of validity evidence is the technical aspect of the assessment. Psychometricians study the statistical properties of the assessment items and on most

modern day high stakes assessments, employ an item response model to scale, calibrate and equate student response data from the assessment. There is a division among practitioners regarding the two most popular IRT models: the Rasch model and the 3PL model. The selected model could have a significant impact on examinee scores as well as school and district report cards and therefore affect the consequential validity of the assessment.

The arguments for the Rasch model are mainly grounded in a philosophical theory of measurement with the goal of utilizing assessments that represent sound measurement. Such assessments would produce a student response data structure that would fit the Rasch model. Assessment items are reviewed in advance in order to remove items that would be aberrant to the Rasch structure. Practitioners also promote Rasch because total score is a sufficient statistics for ability and therefore easy for laypeople to interpret. Thus, the Rasch model scores can be easier to explain and defend to stakeholders.

Proponents of the 3PL argue that the Rasch does not account for student guessing that is a reality in assessments. Furthermore, the flexibility of the 3PL allows it to fit the actual data structure that is present and if the structure is in fact Rasch, the 3PL model will estimate the guessing parameter and item discrimination parameter accordingly. Simulation studies have shown that when the data structure is actually 3PL and the 1PL model is applied, inaccurate estimates of student ability result; however, when the data structure is actually 1PL and the 3PL model is applied, student ability estimates are not as greatly affected (Jiao and Lau, 2003). Nationwide, about 60% of states utilize the Rasch model for their statewide assessment. How would assessment results change at the school and district level if another IRT model was employed? PASS is an example of a

statewide assessment that uses the Rasch model for scaling and calibration. If the 3PL model were used instead of Rasch, how would state and federal report cards from the school and district be impacted? Furthermore, is the impact substantial enough that it affects decisions surrounding curriculum planning, program funding or personnel evaluations that are made from report card results?

While some studies have explored model fit and compared results of examinee ability based on the IRT model, few have utilized actual data from statewide assessments. Recent studies that have utilized real data, examined impact of the IRT on the overall population of examinees. A review of the literature does not show any research studies investigating the impact of the IRT model on assessments results at the school, district or state level. However, multiple decisions are made based of state and federal report cards which are reported at the school and district level.

Proposed Research

The current study proposes to determine the change in achievement level for students on PASS test results based on IRT model for each school and district in South Carolina. This analysis could provide further insight than a percentage change overall for the state because it may capture significance to a particular school or district and report card as these levels are required by law. A school with an unusually high number of low achieving students or students with accommodations may be more sensitive to the inclusion or exclusion of the guessing parameter, for example. Finally, the current study will compare results for two grade levels to determine if the impact differs among grade levels.

This study is significant because the study continues the collection of validity evidence for a high stakes statewide assessment. The practical impact of the IRT model selection on school and district results ultimately impacts far-reaching decisions regarding schools and districts such as curriculum revision and qualifications for grant funding. While the data are limited to South Carolina's statewide assessment, the study is applicable on a national level because the nation is and has been historically split on the use of the Rasch versus 3PL model for statewide assessments. Although many studies have explored the question of model fit, few have addressed the practical significance of model misfit (Sinharay et. al, 2014). This research delves beyond overall results for an assessment by examining the impact of the IRT model at the school and district level for each school and district in the state. In other words, the study extends to the next "layer" of practical significance and contributes to validity evidence as an "evaluation of evidence and consequence" (Messick, 1980).

CHAPTER 3

METHODOLOGY

Purpose

The purpose of this study was to investigate the impact of the choice of IRT models used in the analysis of student response data from statewide educational assessments with the intention of acquiring knowledge to increase the likelihood that valid interpretations are drawn from assessment results. The study focused on the 2014 administration of South Carolina's annual Palmetto Assessment of State Standards (PASS) which utilized an IRT model for scoring student response data. In practice, the Rasch model was used to score and calibrate PASS scores. Recall that the Rasch model is a one parameter logistic model with an item difficulty parameter that varies for each item and an item discrimination parameter that is constant for each item. Meanwhile, the 3PL model is a three parameter logistic model which estimates the following parameters for each item: item difficulty, item discrimination and item guessing. Different IRT models used may lead to different estimates of student ability. This study examined the impact on validity that the choice between the Rasch and 3PL model would have on scoring and calibrating PASS.

Research Questions

The study addressed the following questions using 2014 PASS data for ELA and Math for grades 3 and 8:

1. If a different IRT model were used to score (i.e., calibrate and scale) student responses on PASS, how would state school reports cards be affected? Note that school and district report cards are based on the percentage of students scoring in the 'Not Met,' 'Met,' and 'Exemplary' category in each subject.

2. If a different IRT model were used to score (i.e., calibrate and scale) student responses on PASS, how would federal school reports cards be affected? Note that school and district report cards are based on the mean score for each subject.

3. Is the impact of the IRT model different among age groups? It could be that younger students are more sensitive to a change in IRT model or vice versa. Younger students may be more susceptible to guessing. On the other hand, because older students may be exposed to more difficult questions or higher order thinking problems, and thus, they may be more susceptible to guessing. (Studies pertaining to the relationship between age and guessing were not found in a review of the literature.)

4. Is the impact of the IRT model different among student demographic subgroups (including a subgroup of students who received modifications or accommodations)?

Data Description

The PASS data for this study were provided by the South Carolina State Department of Education (SCDE)[1] in the form of SAS data sets. All students, schools and school districts were de-identified by the SCDE for confidentiality purposes and the de-identified (false) IDs are referred to as student IDs, school IDs, and district IDs. In order to protect confidentiality, schools with a small number of student IDs were combined by the SCDE and represented by a single school ID. The SAS data sets included student response data from the 2014 PASS administration, which was the most recent data available at the time of the data request.

Student response data were obtained for all South Carolina students in the 3rd or 8th grade who attempted at least one question on either the Math or English Language Arts (ELA) portion of the regular PASS test form during Spring 2014. The core subjects of Math and ELA were selected for the study because in South Carolina, all students in grades 3 through 8 are tested in ELA and Math through PASS every year (SCDE, 2012). Also, it will later be established that Math and ELA contribute substantially to scoring components on school and district reports cards. Math and ELA were also selected over other subjects, in part, because the other subjects (writing, science and social studies) are not tested for all students every year (SCDE, 2012). Grades 3 and 8 were selected to include examinees with varying levels of development in test taking skills as well as subject area content with varying levels of complexity. Including a variety of age levels

[1] The use of South Carolina Department of Education records in the preparation of this material is acknowledged, but it is not to be construed as implying official approval of the Department of Education of the conclusions presented.

was important because one of the objectives of the study is to investigate whether certain age groups are more sensitive to the change in IRT model.

For each student ID, the following was provided: the vector of scored responses for Math and ELA PASS questions with '0' representing an incorrect response and '1' representing a correct response, the PASS numerical scale score for both Math and ELA, the PASS performance level ("Not Met," "Met," or "Exemplary"), school ID, school district ID, student gender, student ethnicity, student English speaking status, student free and reduced lunch status, student individualized education plan status (IEP), and student test accommodation status. For PASS, the SCDE assigned a score of '0' to missing student responses or items with multiple responses. Therefore, vectors of scored responses obtained from the SCDE did not contain any missing student responses. More details regarding the provided variables in the data set can be found in Appendix D. Table 3.1 provides counts of students, schools, and districts included in the SCDE provided data.

Table 3.1
Counts of 3rd and 8th South Carolina students taking at least the Math or ELA portion PASS in 2014 along with counts of schools and districts administering PASS to those students

Grade	Students	Schools	Districts
3rd	53,731	634	83
8th	54,906	301	83

Data Preparation

In order to establish confidence in the study results, the first step in analyzing the data included an attempt to replicate the student ability estimates reported by the SCDE.

The method to estimate student ability and corresponding PASS scale scores used by the SCDE is described in the next section.

Method used by SCDE

According to the PASS 2012 Technical documentation[2], Data Recognition Corporation (DRC), a company contracted by SCDE, used Winsteps software for item calibration (SCDE, 2012). These calibrations were run using representative samples from the first set of returns of the statewide administrations (SCDE, 2012). The samples included 20,000 or more students for the subjects and grades tested (SCDE, 2012). The calibrations produced a Rasch ability estimate, denoted by θ, for each possible raw score (SCDE, 2012). Raw score refers to the total number of correct answers. Recall that with the Rasch model, total score is a sufficient statistic for θ. The θs were then converted to a more readable PASS scale score. PASS scale scores range from 300 to 900. The PASS 2012 Technical documentation describes the scaling process as follow:

> For ease of interpretation, PASS abilities for each grade and subject were converted into scale scores. The anchor point for all scales was the met cut point which was set to a scale score of 600; the standard deviation of scale scores in the initial year was set to 50 for every grade and subject. Decisions on the scale score system were made by SCDE staff in consultation with Huynh Huynh of the TAC (Technical Advisory Committee). Calibration of PASS test forms yielded a value of the Rasch ability, theta (θ), corresponding to every possible raw score. Scale scores were calculated for every raw score for each grade and subject using the formula:
>
> $$[\text{unrounded}]\ \text{scale score} = 600 + ((\theta_{RS} - \theta_{Met}) / \sigma_{\theta}) * 50, \text{ where}$$
>
> θ_{RS} is the value of theta corresponding to that raw score, θ_{Met} is the value of theta at the met cutpoint, and σ_{θ} is the initial observed standard deviation of theta for

[2] The PASS 2012 Technical Documentation was the most recent PASS technical documentation available at the time of this study. However, SCDE officials confirmed that the technical methodology relevant to this study is the same for the year 2014.

the specified grade and subject. θ_{Met} is the value of theta at the met cutpoint, and σ_θ is the initial observed standard deviation of theta for the specified grade and subject. Values of θ_{Met} were obtained from the PASS standards setting. Values of σ_θ were computed based on empirical data from the 2009 PASS administration.

Replicating Results of the SCDE

Upon receiving the SCDE data, one data set was selected to determine if the raw score θ's reported by the SCDE could be reproduced. As an initial test, the data for 3rd grade ELA was analyzed using BILOG-MG software, specifying a Rasch model with the maximum likelihood estimation method (MLE). BILOG-MG was selected because it can handle both Rasch analysis and 3PL analysis whereas Winsteps, used by the SCDE, is used for Rasch analysis only. A rescaling option in BILOG-MG, utilizing the mean and standard deviation for the SCDE supplied ability estimates (i.e., θs), placed the BILOG-MG abilities on the same scale as the SCDE supplied theta abilities.

There are three main differences between the procedure used by SCDE and the BILOG-MG procedure:

1. SCDE used a sample of 20,000 students or more for calibration but this study used all student results for calibration (N =53,731 for 3rd Grade ELA).
2. SCDE used Winsteps software for the Rasch analysis but this study used BILOG-MG. In addition, Winsteps, used by the SCDE, uses a joint maximum likelihood estimation for item parameters but marginal maximum likelihood (MML) was used here.
3. Winsteps default values were used for zero and perfect scores (SCDE, 2012) but the BILOG-MG procedure provided estimates for the zero and perfect scores.

Regardless of these differences, the student ability estimates were extremely close under the 2 methods. The theta scores produced by the BILOG-MG matched the theta

abilities reported by the SCDE for 70% of examinees to the nearest hundredth. Note that Winsteps defaults to 5.09 for perfect scores and -4.9 for zero scores. In the data sets examined there were not any zero scores but there were 422 perfect scores for 3rd grade ELA. The BILOG value for a perfect score was 4.77 for 3rd grade ELA. Outside of this extreme, the largest difference between the thetas was .03. The correlation between θ estimates from the two data sets was very high, with r = .999. Due to the majority of ability estimates matching and the rest of the differences being within ±.03 logits (outside of the perfect score exception), and the nearly perfect correlation between the two sets of θs, this was thought to be sufficient to allow BILOG-MG abilities to be converted to PASS scale scores using the formula supplied by the SCDE. Similar results were found for all data sets, with the exception of 8th grade ELA and Math; both tests had significantly fewer numbers of perfect scores.

The BILOG-MG code and other details regarding the BILOG-MG options and estimation methods used can be found in Appendix H. Note that item parameters were estimated with the standard marginal maximum likelihood method in BILOG-MG. Item parameters were not obtained from the SCDE for comparison because ability estimates are the focus of this study.

Estimating Student Ability with BILOG-MG and the 3PL Model

In order to estimate student ability with the 3PL model, BILOG-MG was used. Again, the rescaling option was used to place the ability estimates on the same θ scale as the SCDE θs. However, using the MLE estimation method proved to be problematic with the 3PL model. The MLE method produced extreme values for ability estimates as well as unattainable standard errors for low ability examinees. For 3rd grade ELA, many

of the 1,300 examinees who answered between 1 and 7 items correctly received an estimated ability of -3.99 and an unattainable standard error. Similar results were observed for the other data sets. This issue is known to occur when using MLE estimation with the 3PL model and can be attributed to aberrant patterns, such as examinees correctly answering difficult and discriminating items but incorrectly responding to easier items (Hambleton, Swaminathan, & Rogers, 1991).

Bayesian estimation methods, which incorporate prior information about ability parameters, are able to overcome the estimation issues encountered with MLE and the 3PL model (Hambleton, Swaminathan, & Rogers, 1991). Therefore, Bayes expected a posteriori (EAP) estimation method was used for the 3PL model along with a corresponding Bayesian estimation method for item parameters called maximum marginal a posteriori estimation (MAP).

Note on EAP versus MLE Estimation Method

The decision to use EAP with the 3PL model raised the question of using EAP with the Rasch model as well. However, the focus of this study is to compare the Rasch model to the 3PL model, not to compare estimation methods. Ideally, the same estimation method would be used with both models. However, MLE was used originally the SCDE and did not present problems for Rasch as it does for the 3PL. For thoroughness, the EAP estimation method was compared to the MLE estimation method for Rasch. Model fit appeared to be about the same for both estimation methods with MLE fitting slightly better on extreme low and high ends for the Rash model. Therefore, it was concluded to continue with the MLE estimation with the Rasch model and the EAP

estimation method for the 3PL model. Details for comparing model fit for the two estimation methods can be found in Appendix I.

Data Checks

General investigations of the data were performed before addressing the research questions. The mean and standard deviation of the state supplied θs was obtained to determine a matching θ scale for the BILOG-MG analysis. Also, for each data set, the number of zero and perfect scores were obtained. This count was of interest because Winsteps assigns more extreme values for zero and perfect scores than BILOG-MG. Additionally, the data sets were examined for response strings of zeros at the end of the exam which might indicate guessing. These results can be found in Appendix J.

Assumptions

Chapter 2 described the rigorous assumptions for IRT models that are difficult to meet in practice: unidimensionality, local independence, and monotonicity. This analysis is in part a replication study of a current IRT model being used in practice with an existing educational assessment and therefore will be carried out regardless of assumption outcomes.

Model Fit Checks

Model fit checks were performed to compare the fit of the Rasch model to the fit of the 3PL model for these data sets. The focus if this study is on PASS scores computed from estimated student ability. Therefore, this section will focus on person fit.

However, item fit analyses were completed and details of these checks can be found in Appendix K.

Drasgow, Levine, & Williams (1985) introduced a goodness of fit index, z_h, to measure the degree to which the observed response pattern for each examinee agrees with the response pattern predicted by the item response theory model employed. The z_h index has an empirical distribution that is an approximately standard normal distribution (Drasgow et al, 1985). Furthermore, while z_h is not perfectly independent of ability, the effects of ability level on the index are slight (Drasgow et al, 1985).

For each of the PASS data sets, the z_h index was computed for both Rasch and 3PL using the Multidimensional Item Response Theory (MIRT) package in R (Chalmers et al., 2016). Quantile plots were constructed to compare the Rasch and 3PL results. The plots were constructed for all examinees as well as low, middle and high ability examinees separately to better ascertain where misfit occurred when detected. As discussed in Chapter 2, guessing could have more of an impact with low ability examinees (Lord, 1980). Therefore, the model-fit check was examined for the various ability groups.

In order to obtain more readable plots, the approximately standard normal z_h indices were squared and transformed into Chi-squared distributions with 1 degree of freedom. In the standard normal distribution, 95% of the distribution is between -1.96 and 1.96 while 99% of the distribution is between -2.576 and 2.576. Similarly, 95% of the Chi-squared distribution is below 1.96^2 or 3.84 and 99% of the distribution is below 2.576^2 or 6.64. The quantile plots were examined to see how well the fit indices matched

the theoretical Chi-squared distribution and if the count of outliers outside of 95% and 99% matched the expected counts.

Computing PASS scores with the Rasch Model

After obtaining student abilities (θs) using the BILOG-MG for both Rasch and 3PL models and then investigating model fit, the θ's were rescaled to 'Rasch' PASS scores and '3PL' PASS scores using the SCDE supplied formula. Also, using cut scores supplied by the SCDE, Rasch and 3PL PASS scores were placed into the appropriate performance category for each student ID ("Not Met," "Met," or "Exemplary"). The procedure was used for Math and ELA for both grade levels.

Methodology Research Question 1

Research Question 1

> If a different IRT model were used to score (i.e., calibrate and scale) student responses on PASS, how would state school reports cards be affected? Note that school report cards are based on the percentage of students scoring in the 'Not Met,' 'Met,' and 'Exemplary' category in each subject.

A major component of South Carolina state report cards includes the percentage of students falling into each of the performance categories ('Not Met,' 'Met,' and 'Exemplary') for each subject area. A sample of a district report card can be found in Appendix E. A sample of a school report card can be found in Appendix F.

In order to address Research Question 1, the following analysis was made for both the Rasch results and the 3PL results: the percentage of students falling into each of the

performance categories ('Not Met,' 'Met,' and 'Exemplary') was calculated for each grade and subject and broken down by school and also by district. The results of the two models were compared and reviewed for substantial differences for each grade and subject area for each school and district.

While the actual report cards include percentage in category for all grades combined, this study focuses on percentage in category for individual grades. Note that a shift in percentage one grade level would affect the percentage for combined grade levels.

Methodology Research Question 2

Research Question 2

> If a different IRT model were used to score (i.e., calibrate and scale) student response on PASS, how would federal school reports cards be affected? Note that school report cards are based on the mean score for each subject.

The Elementary and Secondary Education Act (ESEA) federal accountability portion of South Carolina school and district report card contains the mean Math PASS score and the mean ELA PASS score for all students in grades 3-5 and all students in grades 6-8 in the district or school. A sample of a district report card can be found in Appendix E. A sample of a school report card can be found in Appendix F.

In order to address Research Question 2, the following calculation was made for both the Rasch results and the 3PL results: the mean PASS score was calculated for each grade and subject and broken down by school and also by district. The results of the two

models were compared and reviewed for substantial differences for both grades and both subject areas for each school and district.

While the actual report cards include mean scores for combined grades, this study focuses on mean scores for individual grades. Note that a change in the mean for one grade level would affect the mean for combined grades.

Methodology Research Question 3

Research Question 3

Is the impact of the IRT model different among age groups?

One of the main differences in the 3PL versus the Rasch model is the inclusion of the guessing parameter. It is of interest to investigate whether the IRT model has a greater impact for younger versus older students on PASS testing. It could be that younger students are more likely to guess or vice versa, as older students may be exposed to more difficult content or higher order thinking questions.

In order to address this question, the differences in mean PASS scores based on IRT model for 3rd and 8th graders were compared to determine if either grade level is more sensitive to the change in IRT model. Similarly, the differences for percentage in performance category for the two models were compared for 3rd and 8th grade.

Methodology Research Question 4

Research Question 4

Is the impact of the IRT model different among subgroups (including a subgroup of students who received modifications or accommodations)?

The Elementary and Secondary Education Act (ESEA) federal accountability portion of South Carolina school and district report card contains a composite index score which is largely based on the performance of subgroups. For elementary and middle school grades, mean PASS scores for each subject area are used to determine if an annual measurable objective was met.

For example, if the mean Math PASS score for all students in the school meets the annual measurable objective proficiency requirement, the school or district is awarded one point on the compenent system. Also, the mean Math PASS score for *individual subgroups each* may contribute up to one point in the system as well. Subgroups with at least 30 students are included. The 'weight' of the subgroup is the same (up to one point) regardless of the size of the subgroup. That is, a subgroup with 30 students will be weighed as heavily as a sugroup with 500 students, for example.

Points are awarded in this manner for each subject area and subgroup with the potential to earn up to a total of 100 points. Points for Math and ELA subject areas are weighted at 40% each and can contribute up to 80 points on the 100 point system.

A sample of the The ESEA Federal Accountability System Components depicting the relevant subgroups and weights obtained from the ESEA Federal Accountability Brief Tehnical Document (2014) can be found in Appendix G.

In order to test the impact of the IRT model on the composite index score, Rasch mean PASS scores were compared to 3PL mean PASS scores for each subgroup and for each grade and subject area for each school and district. Because subgroups have such a large impact on the composite index score, the sensitivity of subgroups to the IRT model could have a large impact on the composite index score.

Simulation Study

Based on the results of the research questions, a simulation study was conducted for a group of examinees that appear particularly sensitive to the change in IRT model. Here, the group of examinees may be a school or district, a grade level, a subject area, an ethnic group or examinees requiring standard or non-standard accommodations. A limitation of the analysis with the actual student response matrix is that we do not know the real IRT model. For the simulation, Rasch model item parameter, ability estimates, and ability estimate standard errors obtained from the student response data were used to generate 'true' Rasch abilities and Rasch model responses. Then, 3PL model item parameter, ability estimates, and ability standard errors obtained from the student response data were used to generate 'true' 3PL abilities and 3PL responses. The 3PL responses were scaled and calibrated with the both the 3PL and the Rasch model. Also, the Rasch model responses were scaled and calibrated with both the Rasch and the 3PL model. The performance of the matched and mismatched model case scenarios was examined by comparing the estimated student abilities to the true abilities. The results of the simulation study may help to guide model selection when the true model is in question. Figure 3.1 summarizes the simulation study.

		Model used to calibrate and scale responses. (Selected model)	
		Rasch	**3PL**
Model used to simuulate responses. (True model)	**Rasch**	**Fit Model**	**How do estimated abilities compare to true abilities?**
	3PL	**How do estimated abilities compare to true abilities?**	**Fit Model**

Figure 3.1. Organization of simulation study.

CHAPTER 4

RESULTS

The purpose of this study was to determine the impact of the IRT model used to estimate student ability on statewide assessments. The study focused on South Carolina's 2014 PASS assessment results where the Rasch model was used to estimate student ability. The Rasch model is a one parameter logistic model with an item difficulty parameter for each item and also an item discrimination parameter that is constant for each item. This study compares the PASS scores obtained with the Rasch model to PASS scores obtained with the 3PL model. The 3PL model, in addition to the item difficulty parameter, also includes a guessing parameter and item discrimination parameter for each item. Furthermore, this study aims to determine that impact of the IRT model at the school and district levels where decisions are made from statewide assessment data.

Chapter Organization

This chapter begins with the results of Rasch and 3PL model fit checks on the PASS response data. Then, results are presented to address the following research questions:

Research Question 1

If a different IRT model were used to score (i.e., calibrate and scale) student responses on PASS, how would state school reports cards be affected? Note that school report cards are based on the percentage of students scoring in the 'Not Met,' 'Met,' and 'Exemplary' category in each subject.

Research Question 2

If a different IRT model were used to score (i.e., calibrate and scale) student response on PASS, how would federal school reports cards be affected? Note that school report cards are based on the mean score for each subject.

Research Question 3

Is the impact of the IRT model different among age groups?

Research Question 4

Is the impact of the IRT model different among subgroups (including a subgroup of students who received modifications or accommodations)?

Recall that all of the school and district IDs are de-identified (false) and are used solely for reference in this study. Chapter 4 concludes with the results of a simulation study designed to further investigate a subgroup that appeared especially sensitive to the change in model.

Model Fit Checks

Before comparing the results of the Rasch and 3PL models, it was of interest to check which model appeared to be a better fit the PASS response data. A goodness of fit index, z_h (Drasgow et al., 1985) was used to measure the degree to which the observed response pattern for each examinee agreed with the response pattern predicted by the item response theory model employed.

For each of the PASS data sets, the z_h index was computed for both Rasch and 3PL using the Multidimensional Item Response Theory (MIRT) package in R (Chalmers et al., 2016). Quantile plots were constructed to compare the Rasch and 3PL results. In order to obtain more readable plots, the approximately standard normal z_h indices were squared and transformed into Chi-squared distributions with 1 degree of freedom. High z_h indices indicate poor fit. Figure 4.1 shows the quantile plot constructed to compare the Rasch and 3PL results for 3rd grade ELA. Values above the diagonal theoretical Chi-squared distribution reference line indicate lack of fit. As discussed in Chapter 2, guessing could have more of an impact with low ability examinees (Lord, 1980) and therefore the 3PL model, which accounts for guessing, could be a better fit for low examinees. Therefore, the model-fit check was examined for various ability groups. Figure 4.2 shows quantile plot for low ability examinees while Figures 4.3 and 4.4 show the quantile plots for middle and high ability examinees. Note that the lower horizontal line of reference marks the 5% cut-off; 95% of the distribution is expected to be below this line. Also, the higher horizontal line of reference marks the 1% cut-off; 99% of the distribution is expected to be below this line. The black line (top line) represents the

Rasch model and the red line (lower line) represents the 3PL model. Table 4.1 shows counts of extreme values.

Figures 4.1, 4.2 and Figure 4.3 show that the 3PL z_h indices are generally below the Rasch indices and also the Rasch model has indices that are larger that expected when compared to the reference theoretical Chi-squared distribution. (Recall that high values indicate a lack of fit.) Figure 4.2 indicates that the person fit for the Rasch model appears to be worse for lower ability examinees. Figure 4.4 shows that the indices for both models are low for high ability examinees, indicating either possible over-fitting or that the fit statistics are conservative for extreme probabilities. Third grade ELA is presented for illustration but quantile plots for other grades and subject areas are similar. The quantile plots for other grades and subject areas can be found in Appendix M.

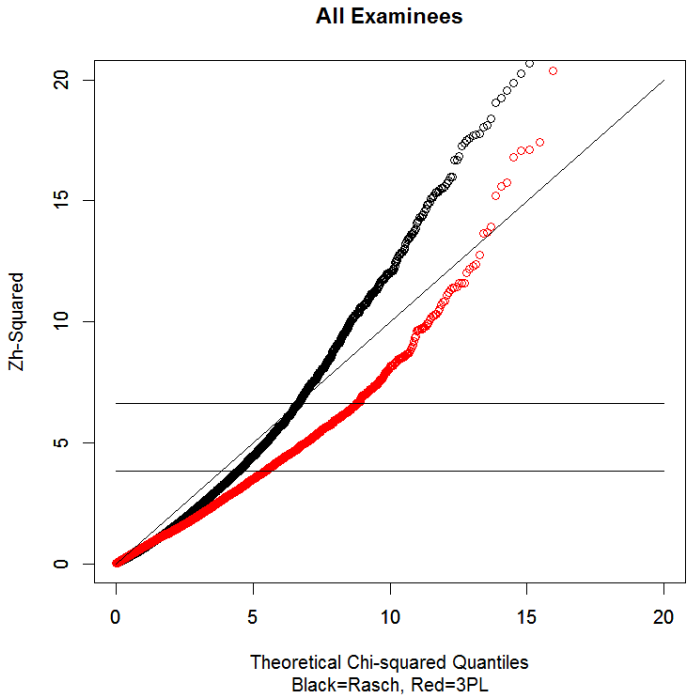


Figure 4.1. Quantile plot for person goodness of fit indices for 3rd Grade ELA all examinees.

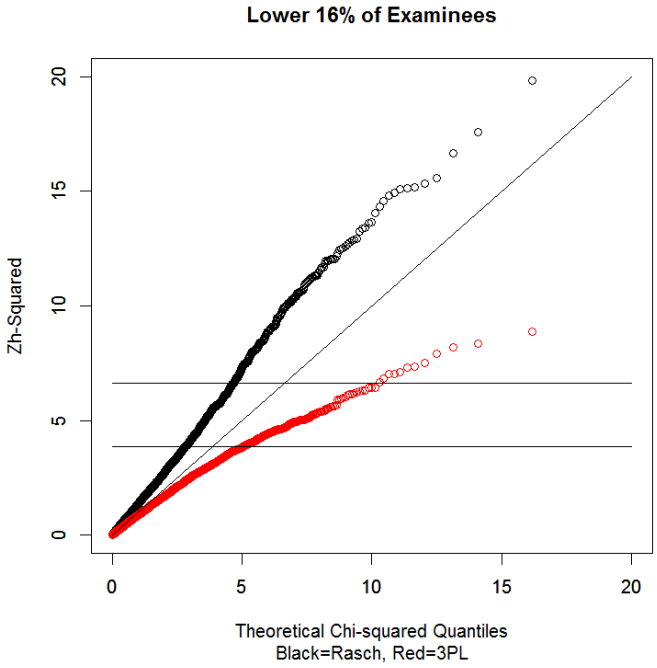


Figure 4.2. Quantile plot for person goodness of fit indices for 3rd Grade ELA low ability examinees.

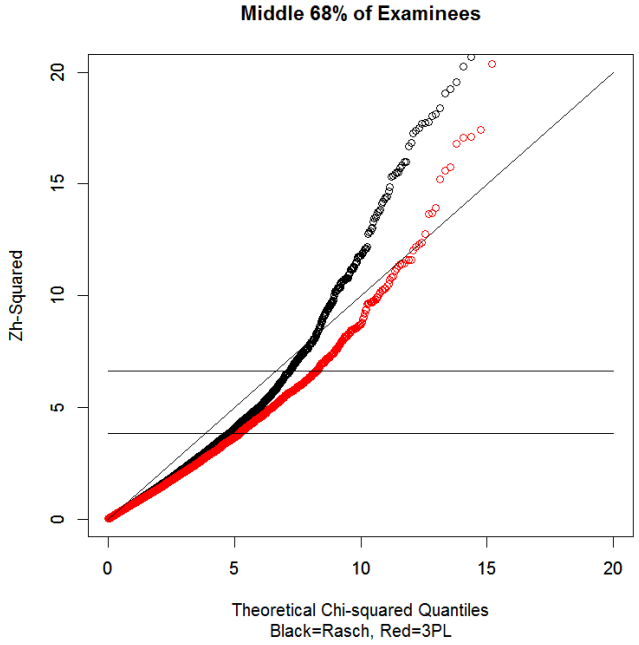


Figure 4.3. Quantile plot for person goodness of fit indices for 3rd Grade ELA middle ability examinees

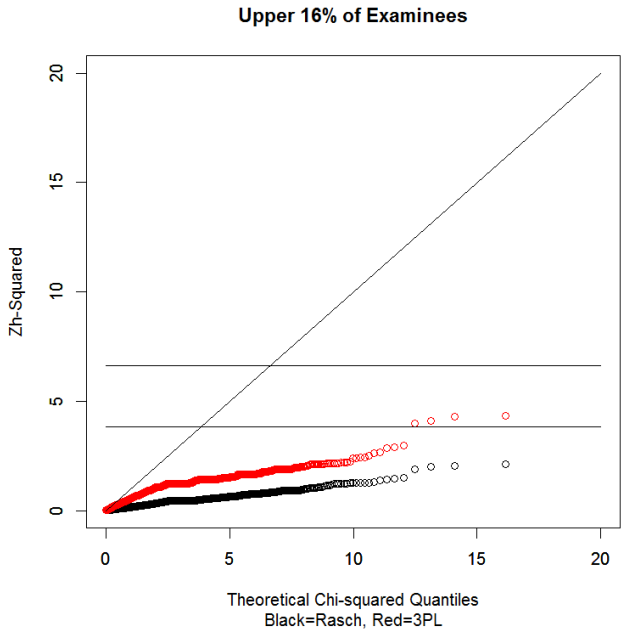


Figure 4.4. Quantile plot for person goodness of fit indices for 3rd Grade ELA high ability examinees

Table 4.1

Count of extreme z_h values 3rd Grade ELA high ability examinees

Examinee ability level		Expected Count	Rasch	3PL
All Examinees	Extreme 5%	2,687	1,844	1,043
	Extreme 1%	537	528	161
Lowest 16%	Extreme 5%	430	812	216
	Extreme 1%	86	263	12
Middle 68%	Extreme 5%	1,826	1,032	823
	Extreme 1%	365	265	149
Highest 16%	Extreme 5%	430	0	4
	Extreme 1%	86	0	0

Table 4.1 shows that both 3PL and Rasch models generally have extreme values that are within the expected count of a theoretical Chi-squared distribution. However, counts of extreme values are higher than expected for low ability students with the Rasch model and lower than expected for the 3PL. Results were similar for other grades and subject areas.

Research Question 1

This section contains results to address Research Question 1:

> If a different IRT model were used to score (i.e., calibrate and scale) student responses on PASS, how would state school reports cards be affected? Note that school report cards are based on the percentage of students scoring in the 'Not Met,' 'Met,' and 'Exemplary' category in each subject.

The analysis for Research Question 1 begins by looking at the percentage of students in PASS performance category overall for all students in the state. Next, we examine the proportion of students who changed performance levels. Then, the change in performance level for schools and districts is presented. Finally, to show how state report cards could be impacted, selected schools or districts with extreme changes in the proportion of students in performance categories for the Rasch versus 3PL model are displayed. The results are presented for each grade and subject.

3rd Grade ELA

First, Figure 4.5 shows the the percentage of students in each performance category for all 3rd grade ELA students. Figure 4.5 mimics the layout of the percentage in performance category presented on state report cards. The state report cards were reported for each school and district though, not for the overall state. Here, the percentages are shown for the overall state as a starting point. The percentage of students in the "Not Met" category is about the same for both the 3PL and Rasch model while the the 3PL has a slightly lower percentage of students in the "Met" category. Because the 3PL results were rescaled to match the Rasch scale by mean and standard deviation, it was expected that the percentage in performance level would be about the same for both models for all students combined.

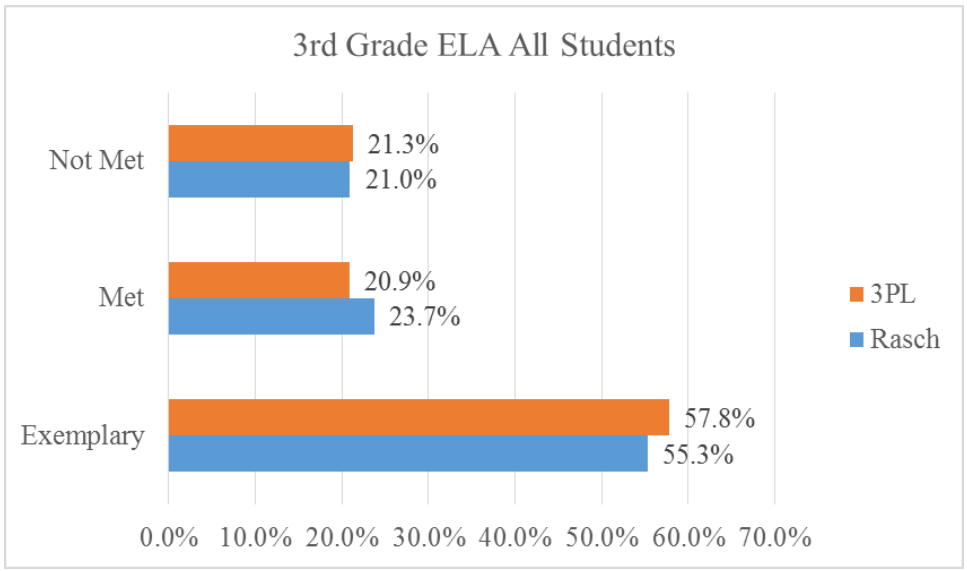


Figure 4.5. Percentage of 3rd grade ELA students in PASS performance categories for the Rasch versus 3PL model, N= 53,731 students

Table 4.2 provides more information regarding the change in performance by indicating the percentage of students in each performance level for Rasch that moved into a different performance level with the 3PL model. Figure 4.5 does not capture students who may have "swapped" performance categories. For example, if 100 students moved from "Exemplary" to "Met" with the change from Rasch to 3PL, and another 100 students moved from "Met" to "Exemplary", then the overall percentage in each category would stay the same. Table 4.2, on the other hand, shows the percentage of students who changed position in the performance category. The most noticeable change is in the "Met" category. Table 4.2 shows that of the 10,488 students who fell in the "Met" category for the Rasch model, 11.9% of those students moved into the "Exemplary" category for 3PL while 5.8% of them moved into the "Not Met" category.

Table 4.2

Change in PASS performance levels for the Rasch versus 3PL model for 3rd grade ELA students

	3PL Level			
Rasch Level	Exemplary	Met	Not Met	All
Exemplary				
Count	29,555	173	0	29,728
Row %	99.4	0.6	0.0	100.0
Met				
Count	1,513	10,488	743	12,744
Row %	11.9	82.3	5.8	100.0
Not Met				
Count	0	579	10,680	11,259
Row %	0.0	5.1	94.9	100.0
All				
Count	31,068	11,240	11,423	53,731
Row %	57.8	20.9	21.3	100.0

Note. For each of the performance categories for Rasch shown on the first column, the corresponding counts and percentages of students is shown for 3PL. For example, for students scoring in the 'Exemplary' category for Rasch, 99.4% of those students also fell into the 'Exemplary' category for 3PL but .6 moved into the 'Met' category.

Figure 4.6 addresses the change in performance level by district for the Rasch versus 3PL model. As shown by the median on the boxplots, the changes for each of the districts tends to follow the pattern shown in Figure 4.5: the change in the "Not Met" category is near zero, Rasch is slightly higher for the "Met" category and lower for the "Not Met" category. Outliers on the graph indicate that some districts had substantial shifts in performance categories. The pattern for schools, shown in Figure 4.7, is similar.

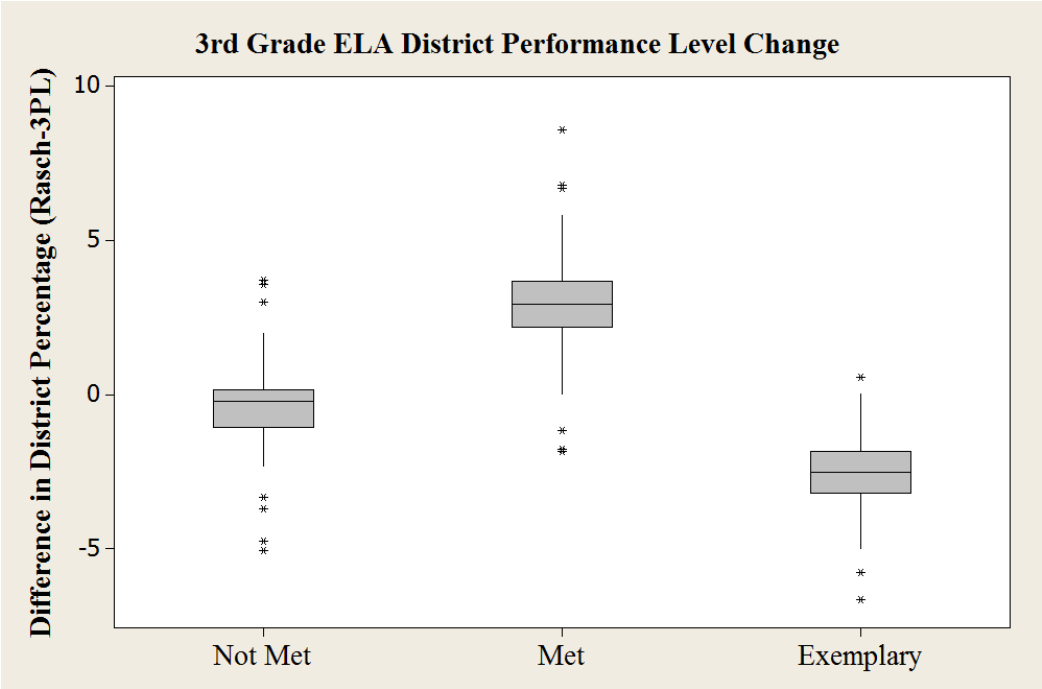


Figure 4.6. Change in percentage of 3rd grade ELA students in PASS performance categories by school district for the Rasch versus 3PL model, N= 83 districts.

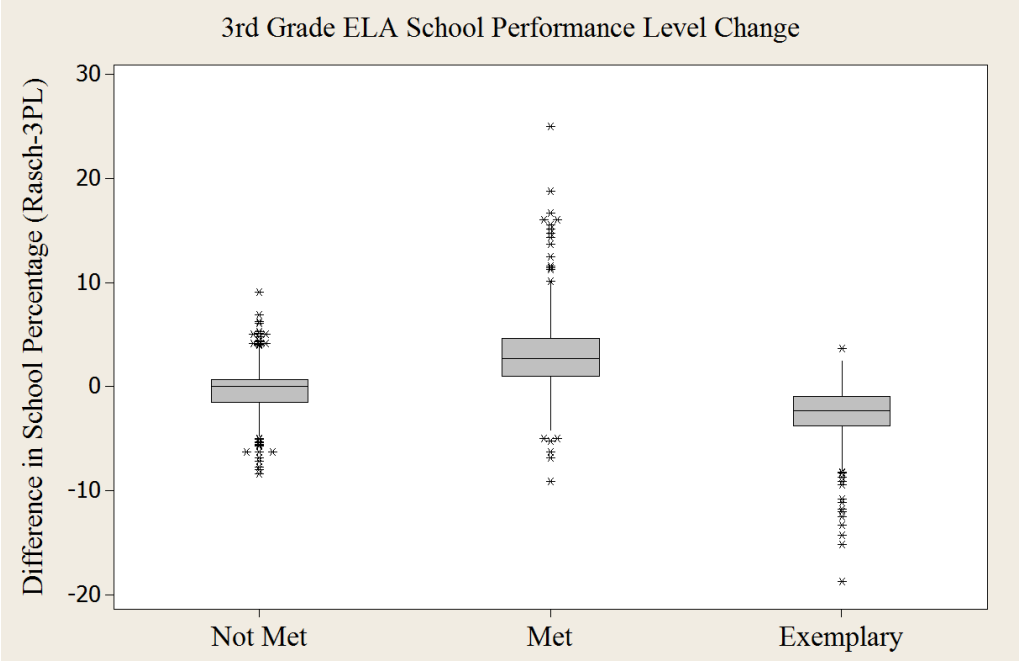


Figure 4.7. Change in percentage of 3rd grade ELA students in PASS performance categories by school for the Rasch versus 3PL model, N= 634 schools.

Figures 4.8 and 4.9 provide examples of schools with extreme shifts in performance levels based on the change from Rasch to 3PL. Recall that all of the school and district IDs are de-identified (false) and are used solely for reference in this study. In Figure 4.8, School ID 32727020, with only 16 third grade students shifted 19% in the "Exemplary" category. While this amounts to only 3 students moving from "Met" to "Exemplary", percentage in category is featured on state report cards. These 3 students each had a PASS score that was at least 8 points higher with the 3PL model than with the Rasch model.

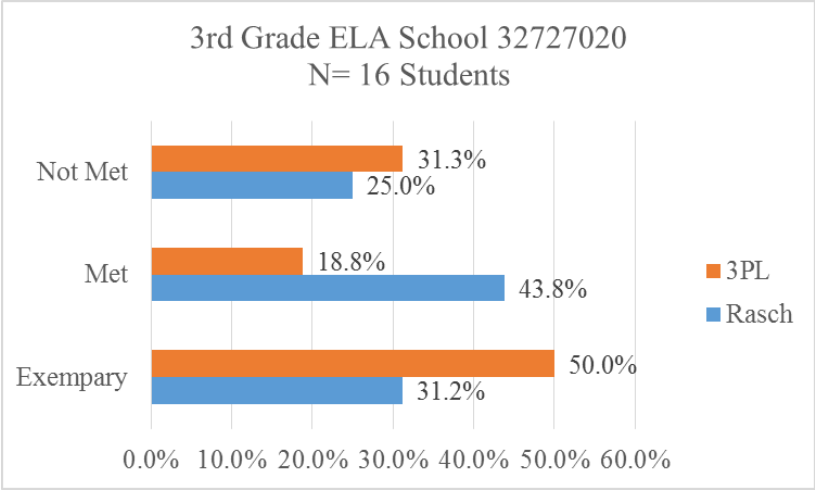


Figure 4.8. Selected sample school, School ID 32727020, with extreme changes for the percentage of students in PASS performance categories for the Rasch versus 3PL model.

In Figure 4.9, School ID 38827012, with 44 3rd grade students shifted down 9% in the "Not Met" category. These 4 students each had a PASS score that was at least 6 points higher with the 3PL model than with the Rasch model.

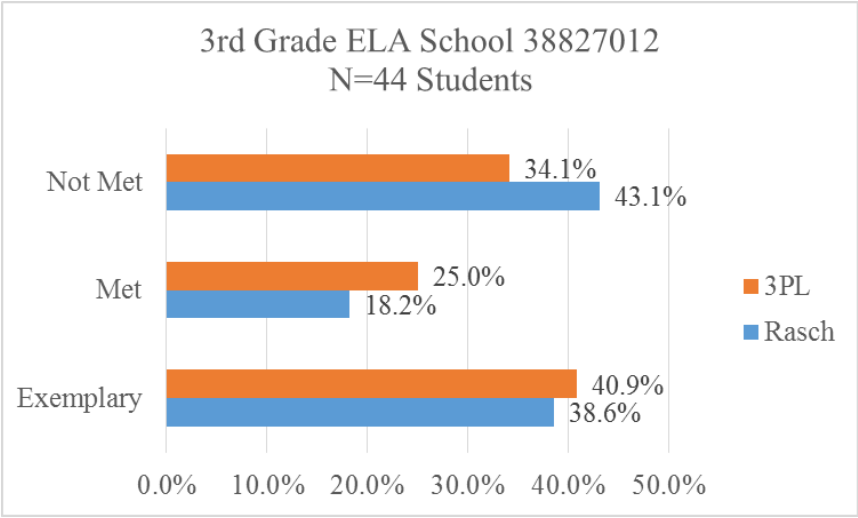


Figure 4.9. Selected sample school, School ID 38827012, with extreme changes for the percentage of students in PASS performance categories for the Rasch versus 3PL model.

3rd Grade Math

Figure 4.10 shows the percentage of students in each performance category for all 3rd grade Math students. The percentage of students in the "Met" category is about the same for both Rasch and 3PL while the the 3PL has a slightly lower percentage of students in the "Not Met" category and more students in the "Exemplary" category.

Table 4.3 shows that of the 16,273 students in the "Not Met" category for Rasch, about 10% of those students change to the "Met" category for 3PL.

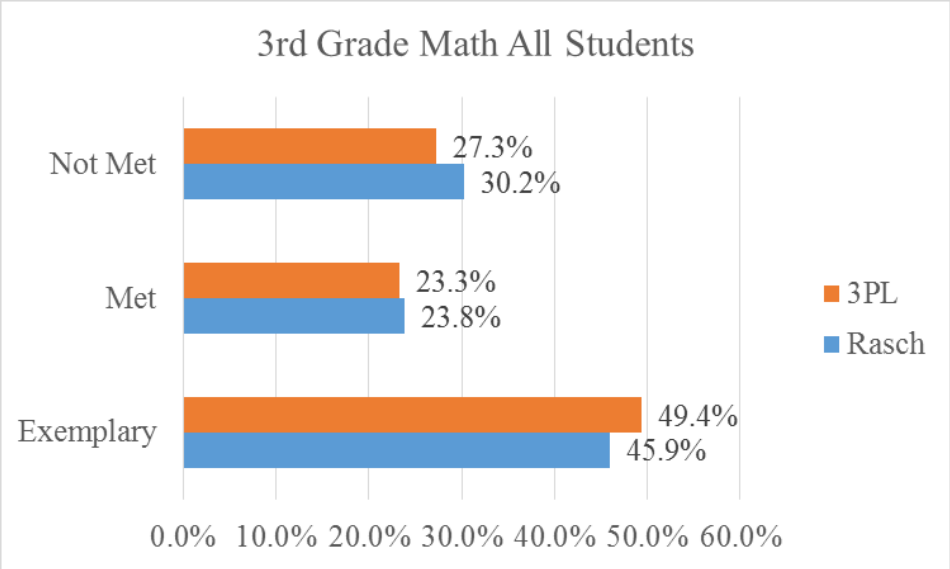


Figure 4.10. Percentage of 3rd grade Math students in PASS performance categories for the Rasch versus 3PL model, N= 53,829 students

Figure 4.11 addresses the change in performance level by district for the Rasch versus 3PL model. As shown by the median on the boxplots, the changes for each of the districts tends to follow the pattern shown in Figure 4.10: the change in the "Met" category is near zero, Rasch is slightly higher for the "Not Met" category and lower for the "Exemplary" category. Outliers on the graph indicate that some districts had

substantial shifts in performance categories. The pattern for schools, shown in Figure 4.12, is similar.

Table 4.3

Change in PASS performance levels for the Rasch versus 3PL model for 3rd grade Math students

	3PL Level			
Rasch Level	Exemplary	Met	Not Met	All
Exemplary				
Count	24,640	91	0	24,731
Row %	99.6	0.4	0.0	100.0
Met				
Count	1,945	10801	79	12,825
Row %	15.2	84.2	0.6	100.0
Not Met				
Count	0	1,654	14,619	16,273
Row %	0.0	10.2	89.8	100.0
All				
Count	26,585	12,546	14,698	53,829
Row %	49.4	23.3	27.3	100.0

Note. For each of the performance categories for Rasch shown on the first column, the corresponding counts and percentages of students is shown for 3PL. For example, for students scoring in the 'Exemplary' category for Rasch, 99.6% of those students also fell into the 'Exemplary' category for 3PL but .4% moved into the 'Met' category.

Figure 4.13 provides an example of a school with extreme shifts in performance levels based on the change from Rasch to 3PL. In Figure 4.13, School ID 33927011, with only 20 third grade students, shifted down 25% in the "Not Met" category.

Figures 4.14 provides an example of a district with extreme shifts in performance levels based on the change from Rasch to 3PL. In Figure 4.13, District ID 38355, with 64 third grade students, shifted down 8% in the "Not Met" category.

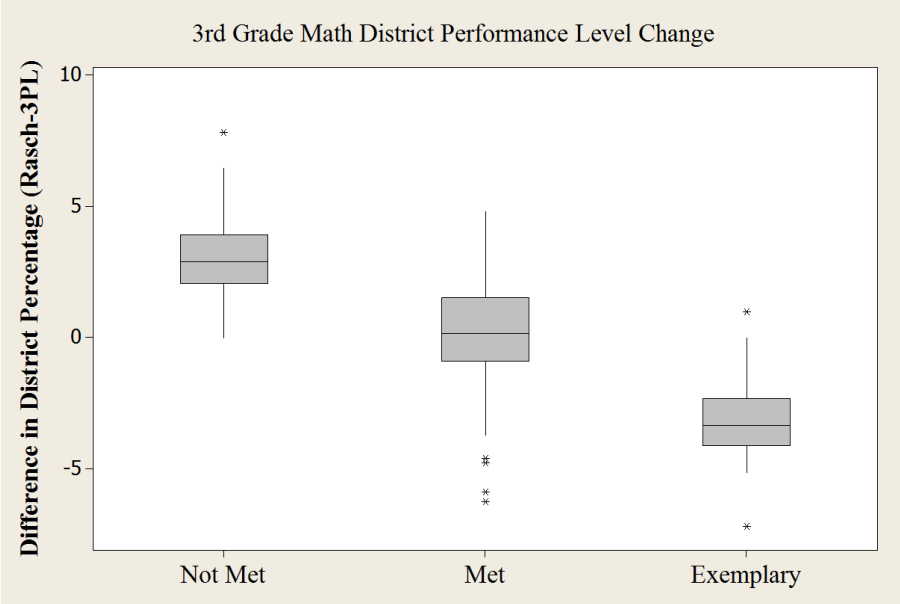


Figure 4.11. Change in percentage of 3rd grade Math students in PASS performance categories by school district for the Rasch versus 3PL model, N= 83 districts.

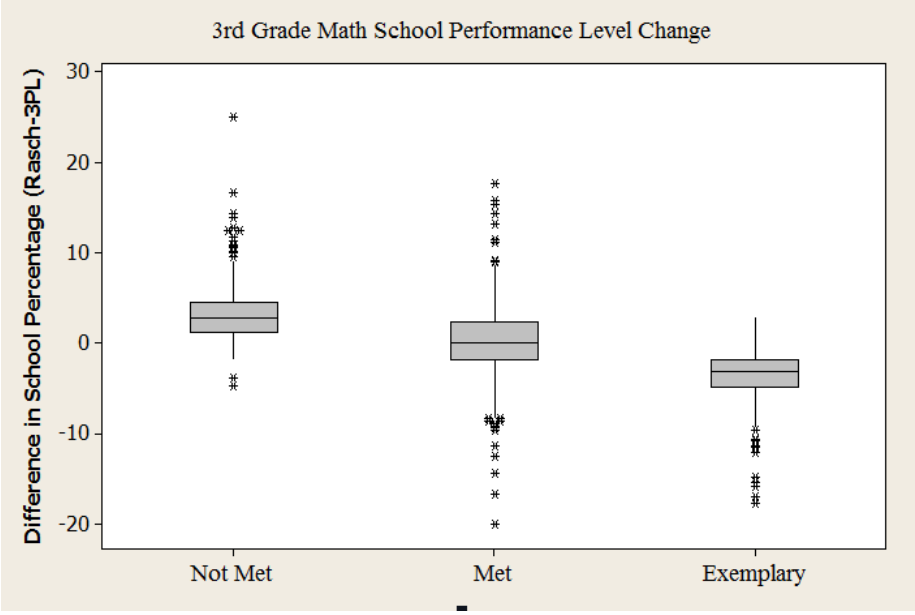


Figure 4.12. Change in percentage of 3rd grade Math students in PASS performance categories by school for the Rasch versus 3PL model, N= 634 schools.

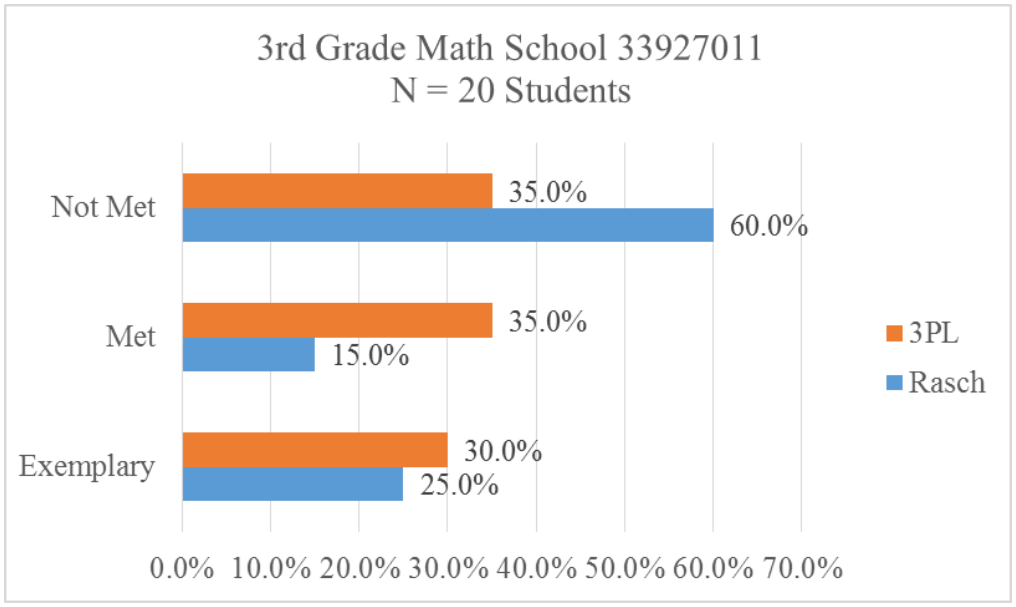


Figure 4.13. Selected sample school, School ID 33927011, with extreme changes for the percentage of students in PASS performance categories for the Rasch versus 3PL model.

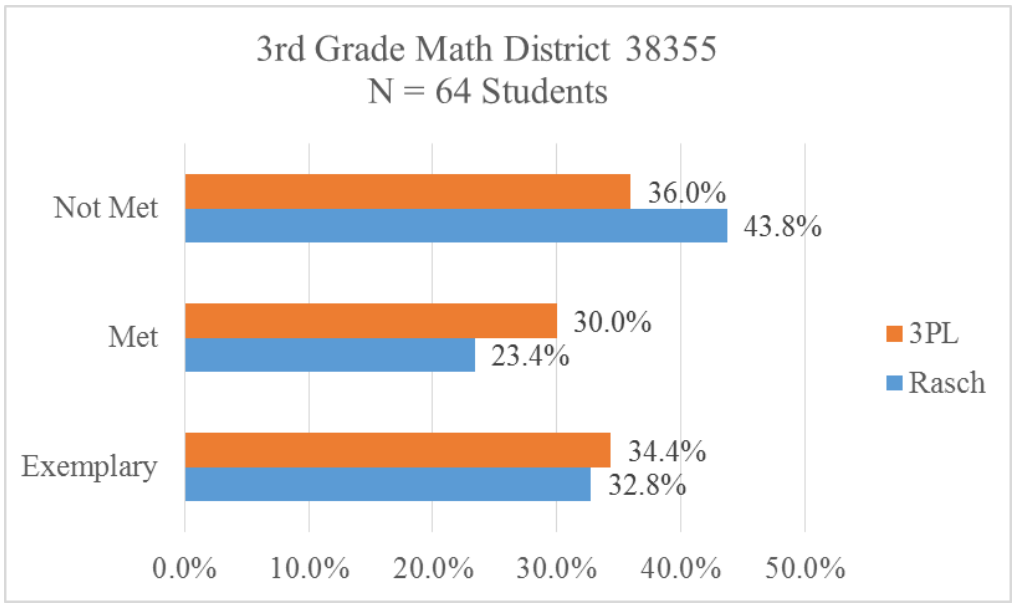


Figure 4.14. Selected sample district, District ID 38355, with extreme changes for the percentage of students in PASS performance categories for the Rasch versus 3PL model.

8th Grade ELA

Figure 4.15 shows the the percentage of students in each performance category for all 8th grade ELA students. The percentage of students in the "Met" and "Exemplary" category is about the same for both Rasch and 3PL while the the 3PL has a slightly lower percentage of students in the "Not Met" category.

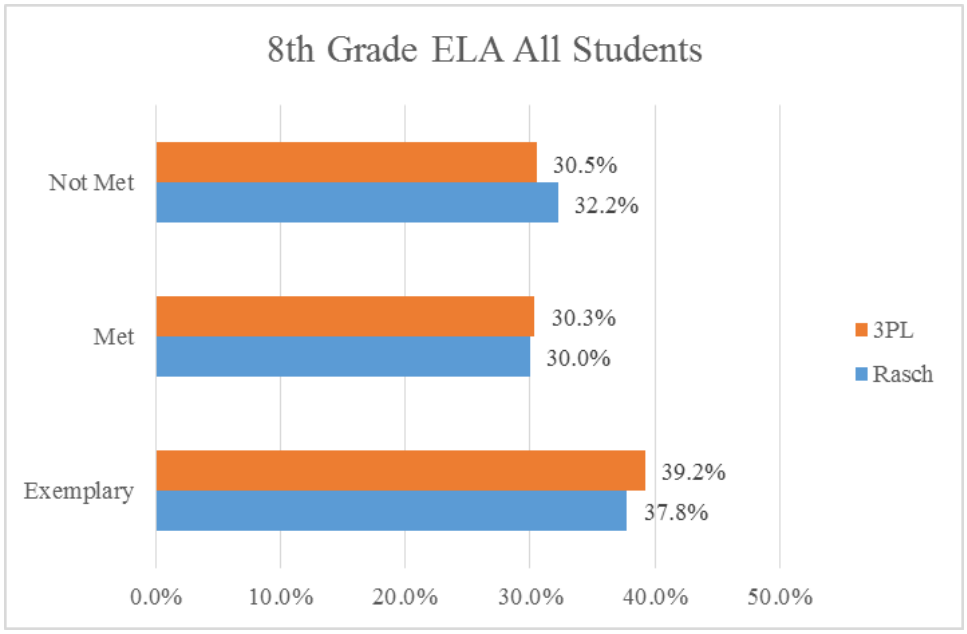


Figure 4.15. Percentage of 8th grade ELA students in PASS performance categories for the Rasch versus 3PL model, N= 54,828 students

Table 4.4 shows that of the 17,679 students in the "Not Met" category for Rasch, about 7% of those students change to the "Met" category for 3PL. Again, this analysis shows that even though overall percentage may not show differences in results between

the two models (i.e., Figure 4.15), Table 4.4 indicates that students are "swapping" positions in performance categories.

Table 4.4

Change in PASS performance levels for the Rasch versus 3PL model for 8th grade ELA students

	3PL Level			
Rasch Level	Exemplary	Met	Not Met	All
Exemplary				
Count	20,084	618	0.0	20,702
Row %	97.0	3.0	0.0	100.0
Met				
Count	1,393	14,824	230	16,447
Row %	8.5	90.1	1.4	100.0
Not Met				
Count	0.0	1,191	16,488	17,679
Row %	0.0	6.7	93.3	100.0
All				
Count	21,477	16,633	16,718	54,828
Row %	39.2	30.3	30.5	100.0

Note. For each of the performance categories for Rasch shown on the first column, the corresponding counts and percentages of students is shown for 3PL. For example, for students scoring in the 'Exemplary' category for Rasch, 97.0% of those students also fell into the 'Exemplary' category for 3PL but 3.0% moved into the 'Met' category.

Figure 4.16 displays the change in performance level by district for the Rasch versus 3PL model. As shown by the median on the boxplots, the changes for each of the districts tends to follow the pattern shown in Figure 4.15: the change in the "Met" and "Exemplary" category is near zero while Rasch is slightly higher for the "Not Met". Outliers on the graph indicate that some districts had substantial shifts in performance categories, though the shifts are slighter than what was observed in the 3rd grade subjects. The pattern for schools, shown in Figure 4.17, is similar.

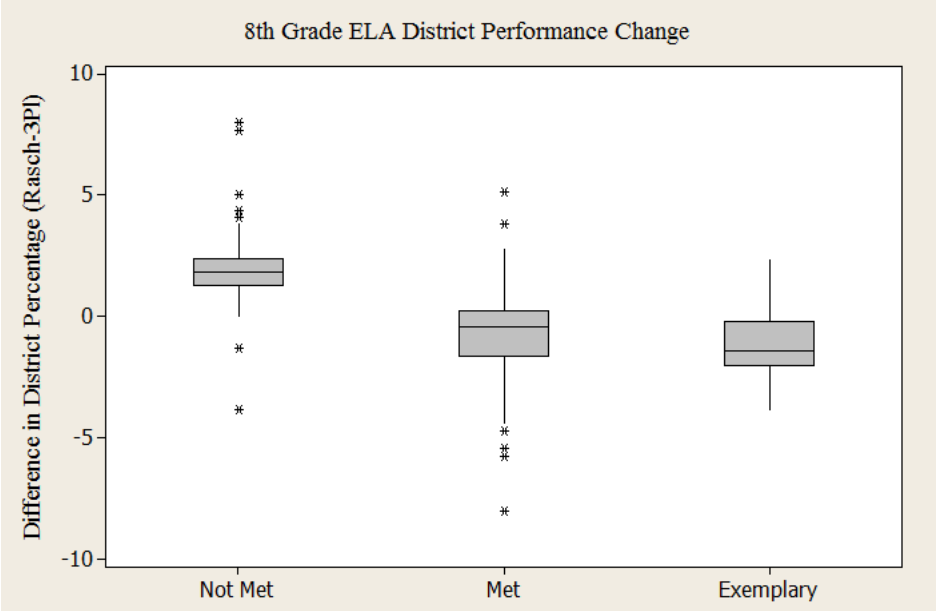


Figure 4.16. Change in percentage of 8th grade ELA students in PASS performance categories by school district for the Rasch versus 3PL model, N= 83 districts.

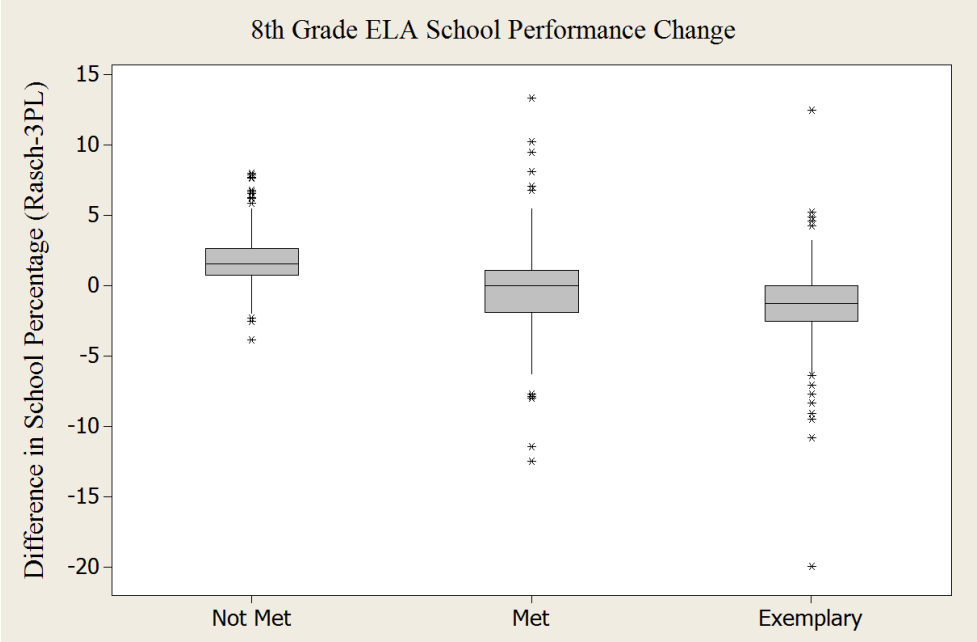


Figure 4.17. Change in percentage of 8th grade ELA students in PASS performance categories by school for the Rasch versus 3PL model, N= 301 schools.

Figures 4.18 provides an example of a district with a substantial shift in the "Not Met" category. District ID 38345 shifted down 8% in the "Not Met" category.

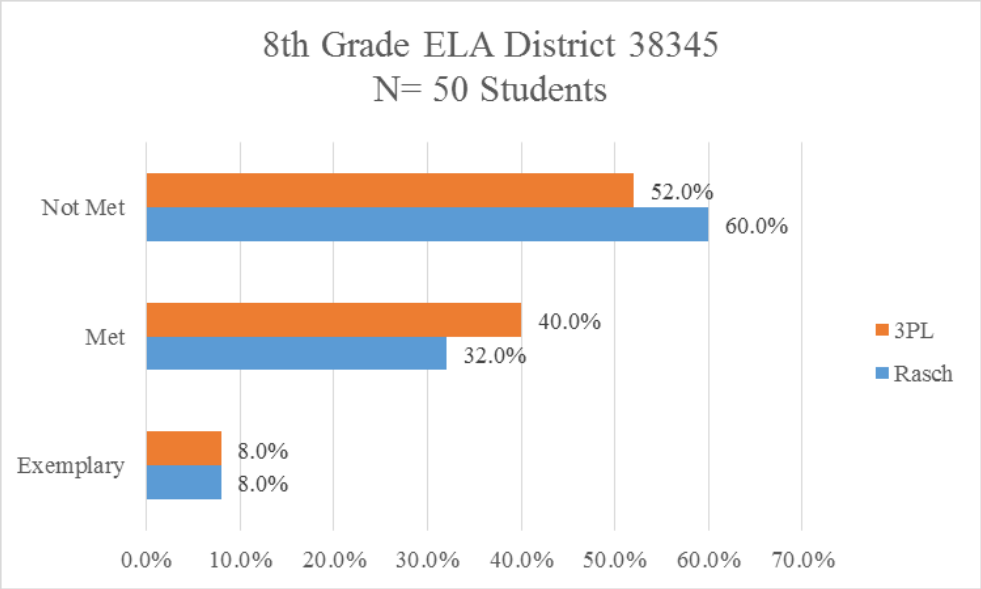


Figure 4.18. Selected sample district, District ID 38345, with extreme changes for the percentage of students in PASS performance categories for the Rasch versus 3PL model.

Figures 4.19 provides an example of a large school with a substantial shift in the "Not Met" category. There were smaller schools with more extreme shifts but in order to provide variety in the school sizes, a larger school was selected to display.

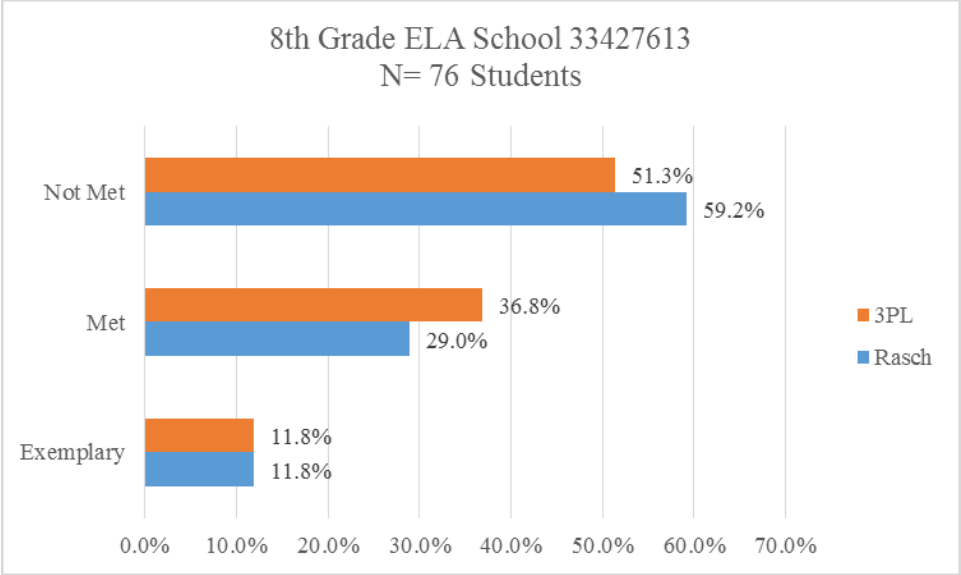


Figure 4.19. Selected sample school, School ID 33427613, with extreme changes for the percentage of students in PASS performance categories for the Rasch versus 3PL model.

8th Grade Math

A concern that developed during the analysis of 8th grade Math was that the distribution of Rasch abilities did not have a normal distribution. Figure 4.20 is a normal quantile plot of the Rasch abilities (θs) that shows a right skewed distribution. (Quantile plots for other grades and subjects were normally distributed with light tails. The quantile plots for the other data sets are in Appendix L.) Recall that the method for putting the 3PL θs on the same scale as the Rasch θs was to match the mean and standard deviation. This method is reasonable for normal distributions and seemed logical due to the SCDE's method of scaling θ's to PASS scores based on the θ mean and the θ standard deviation. However, for 8th grade Math, this method resulted in a range of PASS scores for Rasch that was too far off from the 3PL PASS scores to be reasonably comparable.

Therefore, a more stringent rescaling method was also employed for 8th grade Math. For 8th grade Math, in addition to matching the Rasch theta scale on mean and standard deviation, an equi-percentile rescaling method was also applied. Here the 3PL PASS scores are rank-ordered and matched to the Rasch PASS scale based on rank. For example, the 10th highest scoring examinee for 3PL will have a PASS score equal to the 10th highest Rasch PASS score.

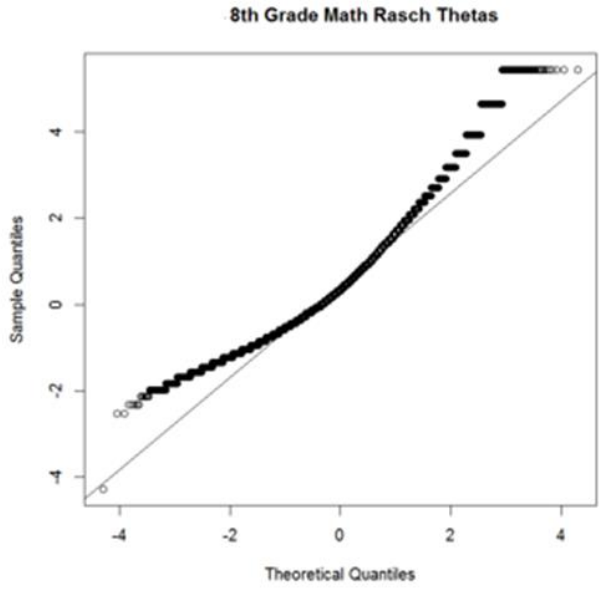


Figure 4.20. Quantile plot of 8th grade Math Rasch abilities (θs) shows a right skewed distribution. N = 54,885 students.

Figure 4.21 shows the the percentage of students in each performance category for all 8th grade Math students. As expected, the percentage of students in all categories is roughly the same for the Rasch model and the 3PL model with the equi-percentile rescaling method (3PL EQ%). The 3PL model with the original rescaling method (3PL) has a larger percentage of students in the "Exemplary" category and fewer students in the "Not Met" and "Met" categories.

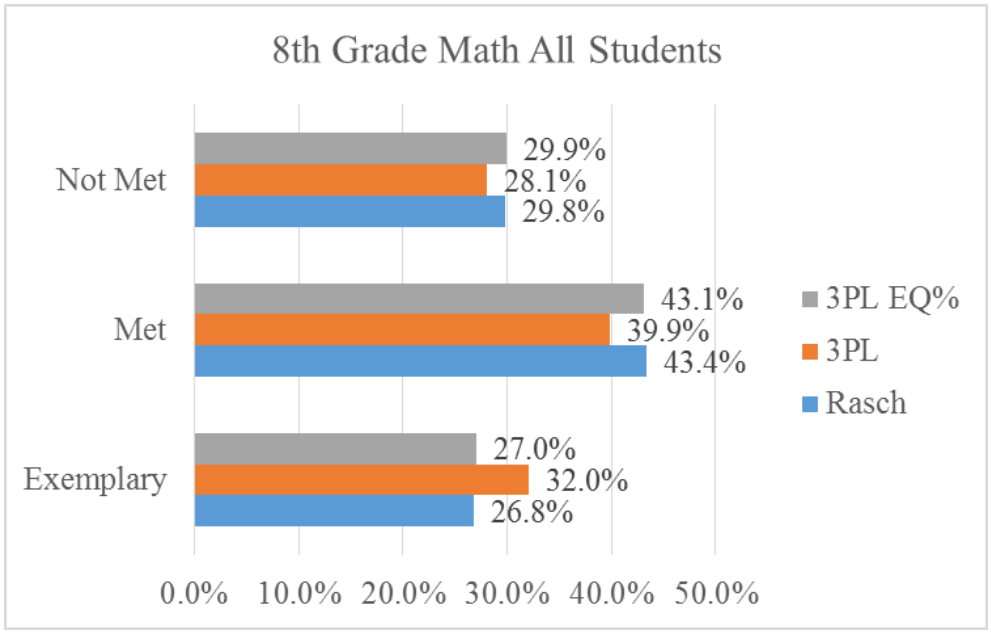


Figure 4.21. Percentage of 8th grade Math students in PASS performance categories for the Rasch versus 3PL model, and also the 3PL model with equi-percentile rescaling, N= 54,885 students.

Table 4.5 shows that of the 16,375 students in the "Not Met" category for Rasch, about 10% of those students change to the "Met" category for 3PL. In Table 4.6, we see that 6.4% of the 16,375 students move to "Not Met" for 3PL EQ%.

Figure 4.22 compares the Rasch and 3PL model for districts. Figure 4.23 compares the Rasch and 3PL EQ% for districts. Figure 4.24 compares the Rasch and 3PL for schools and Figure 4.25 compares Rasch and 3PL EQ% for schools. For each of these figures, as shown by the median on the boxplots, the changes for schools and districts tends to follow the pattern shown in Figure 4.21: there is little difference between Rasch and 3PL EQ% but 3PL is higher in the "Exemplary" category. Outliers

on the figures indicate that some schools and districts had substantial shifts in performance categories.

Table 4.5

Change in PASS performance levels for the Rasch versus 3PL model for 8th grade Math students

	3PL Level			
Rasch Level	Exemplary	Met	Not Met	All
Exemplary				
Count	14,655	38	0.0	14,693
Row %	99.7	0.3	0.0	100.0
Met				
Count	2,928	20,209	680	23,817
Row %	12.3	84.9	2.9	100.0
Not Met				
Count	0.0	1,655	14,720	16,375
Row %	0.0	10.1	89.9	100.0
All				
Count	17,583	21,902	15,400	54,885
Row %	32.0	39.9	28.1	100.0

Note. For each of the performance categories for Rasch shown on the first column, the corresponding counts and percentages of students is shown for 3PL. For example, for students scoring in the 'Exemplary' category for Rasch, 99.7% of those students also fell into the 'Exemplary' category for 3PL but 0.3% moved into the 'Met' category.

Figure 4.26 provides an example of a school with a substantial shift in the "Met" category for both the 3PL and 3PL EQ% as compared to the Rasch model. Rasch had a substantially larger proportion of students in the "Not Met" category then both 3PL and 3PL EQ%.

Figure 4.27 provides an example of a district with the reverse effect of model change than School 38527015. District 38345 has more students in the "Met" category for Rasch than for 3PL and 3PL EQ% and less students in the "Not Met" category

Table 4.6

Change in PASS performance levels for the Rasch versus 3PL model for 8th grade Math students where the equi-percentile rescaling method was used with the 3PL model.

	3PL Level Equipercentile			
Rasch Level	Exemplary	Met	Not Met	All
Exemplary				
Count	14,182	511	0.0	14,693
Row %	96.5	3.5	0.0	100.0
Met				
Count	633	22,093	1,091	23,817
Row %	2.7	92.8	4.6	100.0
Not Met				
Count	0.0	1,046	15,329	16,375
Row %	0.0	6.4	93.6	100.0
All				
Count	14,815	23,650	16,420	54,885
Row %	27.0	43.1	29.9	100.0

Note. For each of the performance categories for Rasch shown on the first column, the corresponding counts and percentages of students is shown for 3PL. For example, for students scoring in the 'Exemplary' category for Rasch, 96.5% of those students also fell into the 'Exemplary' category for 3PL but 3.5% moved into the 'Met' category.

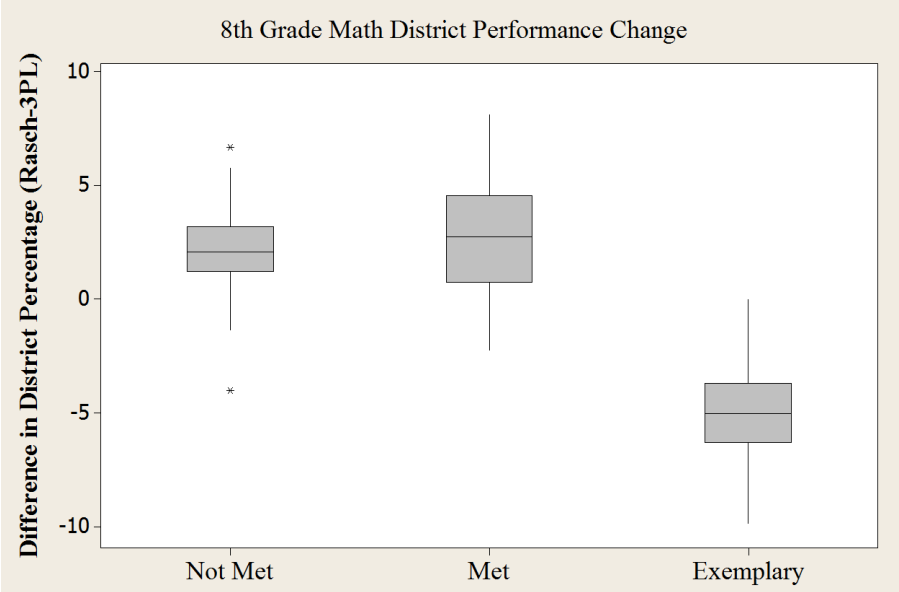


Figure 4.22. Change in percentage of 8th grade Math students in PASS performance categories by district for the Rasch versus 3PL model, (without equi-percentile rescaling). N= 301 schools.

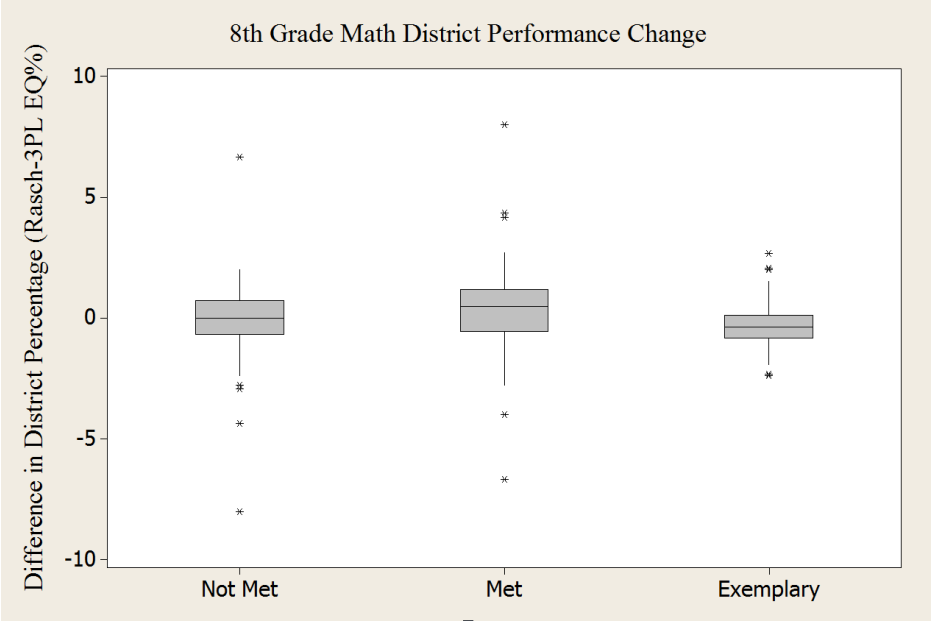


Figure 4.23. Change in percentage of 8th grade Math students in PASS performance categories by district for the Rasch versus 3PL model with equi-percentile rescaling. N= 83 districts.

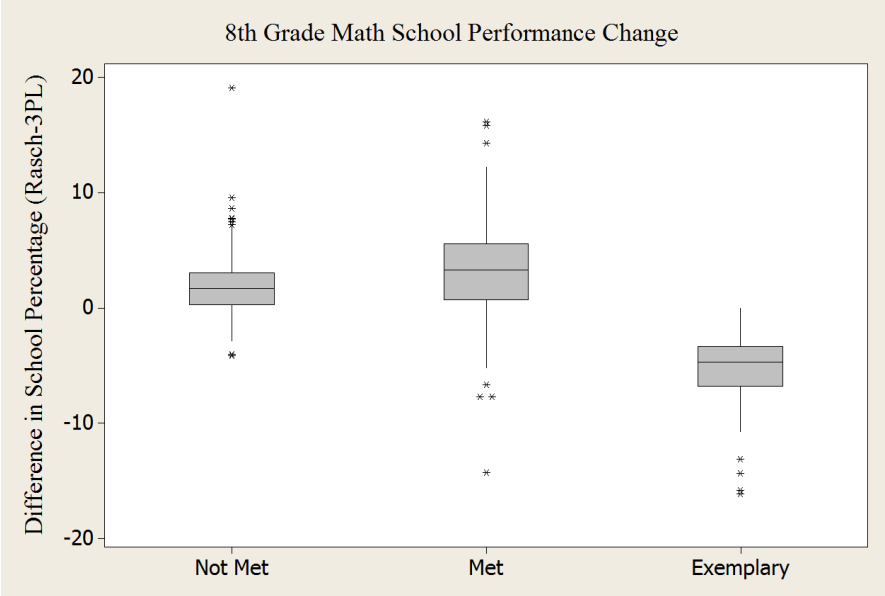


Figure 4.24. Change in percentage of 8th grade ELA students in PASS performance categories by school for the Rasch versus 3PL model (without equi-percentile rescaling). N= 301 schools.

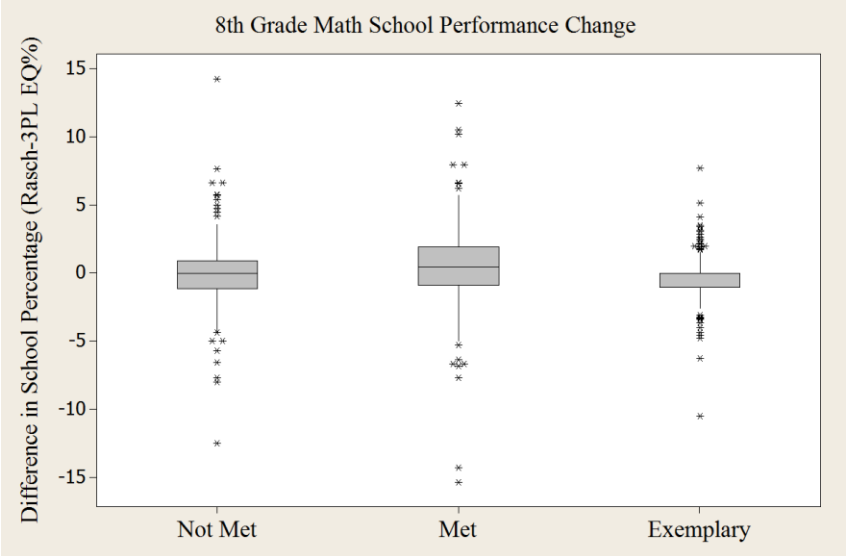


Figure 4.25. Change in percentage of 8th grade Math students in PASS performance categories by school for the Rasch versus 3PL model with equi-percentile rescaling. N= 301 schools.

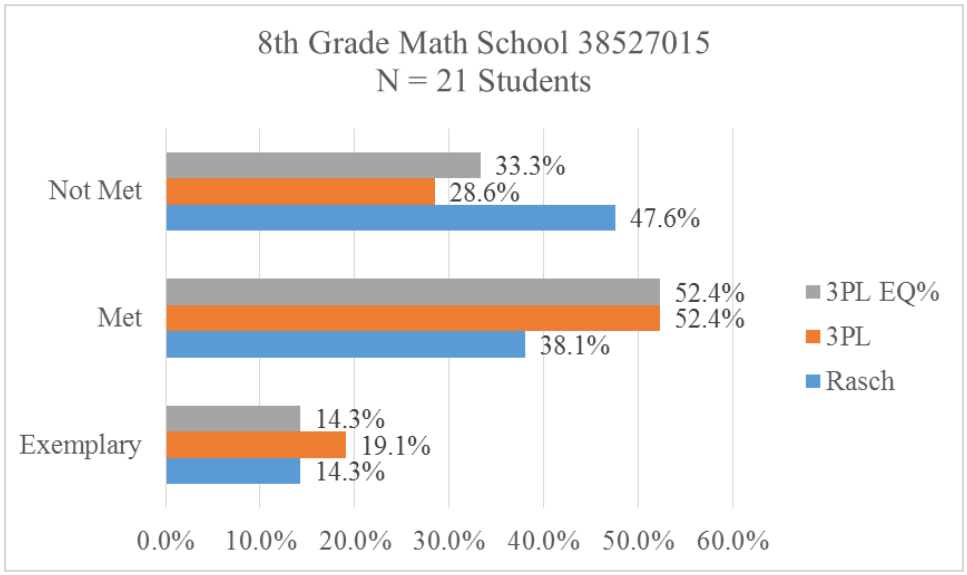


Figure 4.26. Selected sample school, School ID 38527015, with extreme changes for the percentage of students in PASS performance categories for the Rasch versus 3PL model.

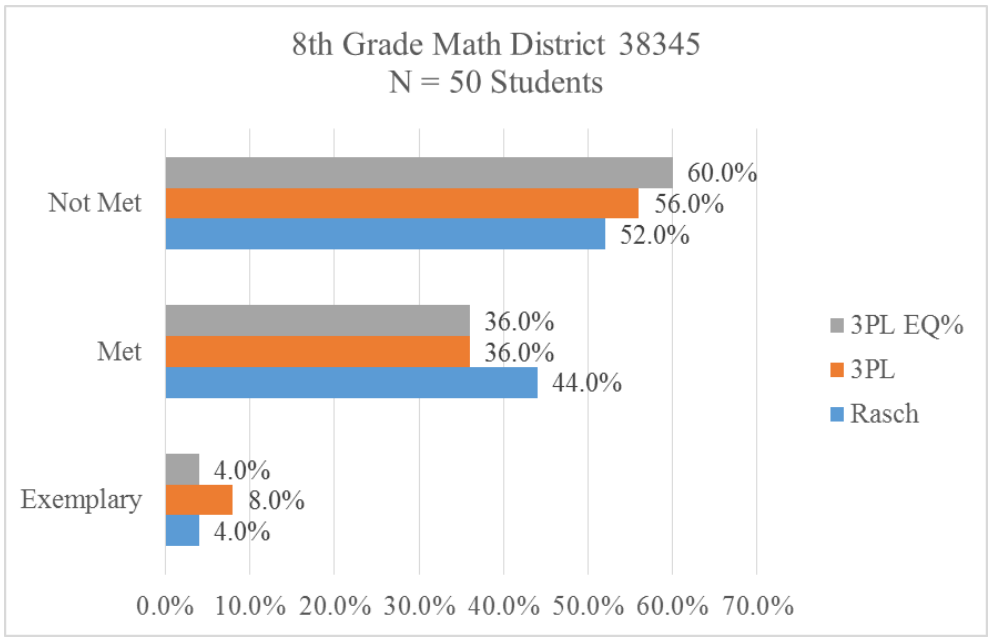


Figure 4.27. Selected sample school district, District ID 38345, with extreme changes for the percentage of students in PASS performance categories for the Rasch versus 3PL model.

Summary

Some schools, especially smaller schools, would find significant shifts for percentage in category for the Rasch versus 3PL model. As shown in the boxplots for each grade and subject, the shift could go in either direction but the most extreme shifts show schools with a higher proportion of students in the "Not Met" category for Rasch than for 3PL. With equi-percentile equating for 8th grade Math, the shifts are more symmetric, meaning the percentage in performances category shift equally; some schools will have more students in the "Not Met" category for Rasch and vice versa. This is evidenced by the symmetric pattern of the boxplots in Figure 4.25. However, for most schools, there would be little or no change for percentage in category. Students within a school or district "swap" categories for Rasch versus 3PL but this would not be reflected on state report cards.

Research Question 2

The following section addresses Research Question 2:

> If a different IRT model was used to score (i.e., calibrate and scale) student response on PASS, how would federal school reports cards be affected? Note that school report cards are based on the mean score for each subject.

The federal report cards focus on the mean PASS score for subject areas in schools and districts. The mean PASS score is considered for all students as well as for subgroups. This section will focus on PASS means for all students in schools and districts. Subgroups will be explored more thoroughly with Research Question 4.

We will begin by looking at scatterplots comparing the PASS scores for all students in the state to examine differences in PASS scores for the 3PL and Rasch model at the student level. Scatterplots comparing Rasch and 3PL for school means and district means will also be examined to help determine the impact of the 3PL versus Rasch model at the school and district level. Frequency tables are provided for further investigation of the differences in means for schools and districts. Finally, selected schools and districts are presented that have substantial differences in mean PASS scores for the 2 models to further explore the impact of the IRT model on particular schools or districts. These analyses are provided using mean and standard deviation rescaling for each grade and subject. For 8th grade Math only, equi-percentile rescaling results are presented as well.

3rd Grade ELA

For 3rd grade ELA, Figure 4.28 shows that on the student level, the PASS scores differed the most for the two models between scores 550 and 600 at the student level. However, Figures 4.29 and 4.30 show that mean scores at the school and district level differ only slightly.

Table 4.7 provides a closer look at how schools differ in PASS mean scores. There are a couple of extreme cases, but for the most part, differences in PASS means scores for schools are minimal.

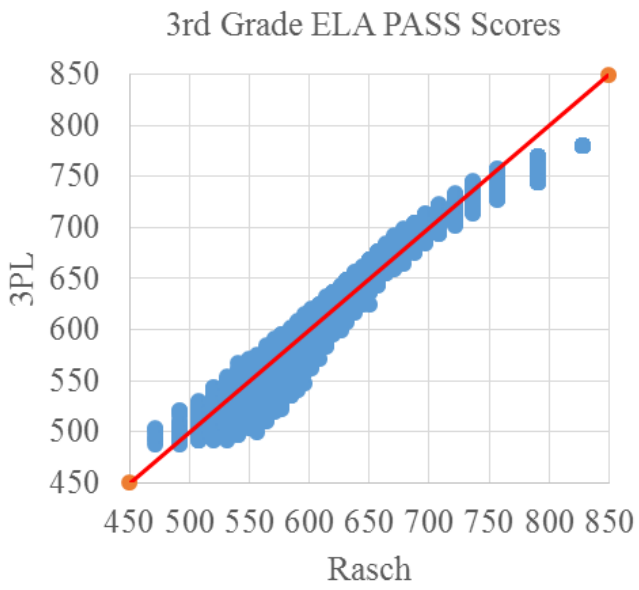


Figure 4.28. Scatterplot of PASS scores, 3PL versus Rasch for 3rd grade ELA with an "x = y" line of reference showing where the scores are equal. N = 53,731 students.

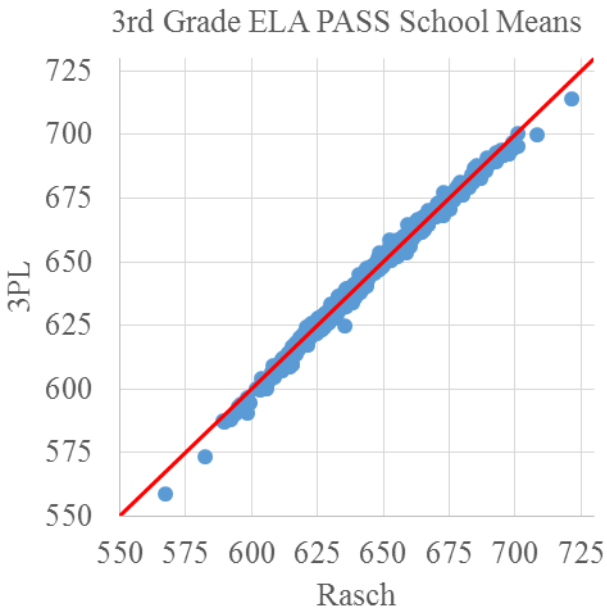


Figure 4.29. Scatterplot of school PASS mean scores, 3PL versus Rasch for 3rd grade ELA with an "x = y" line of reference showing where the scores are equal. N = 634 schools.

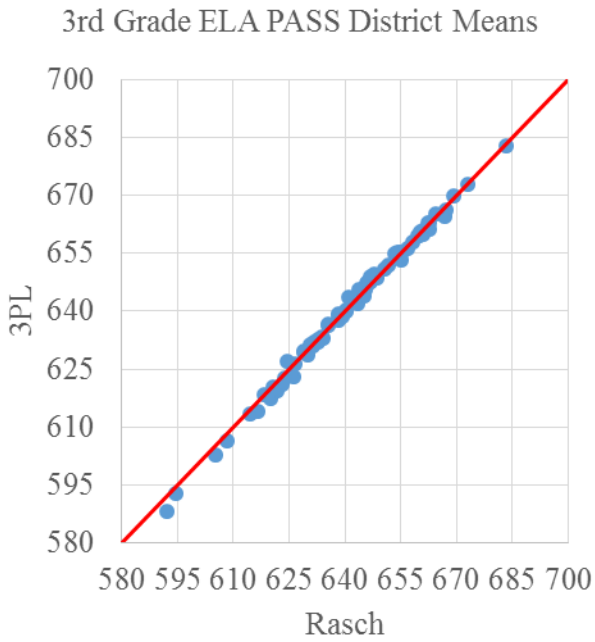


Figure 4.30. Scatterplot of district PASS mean scores, 3PL versus Rasch for 3rd grade ELA with an "x = y" line of reference showing where the scores are equal. N = 83 districts.

Table 4.7

Frequency table comparing Rasch and 3PL School PASS means for 3rd grade ELA

Diff	Count	%
-6	1	0.16
-5	3	0.47
-4	10	1.58
-3	38	5.99
-2	90	14.20
-1	121	19.09
0	125	19.72
1	108	17.03
2	64	10.09
3	40	6.31
4	18	2.84
5	9	1.42
6	1	0.16
8	3	0.47
9	2	0.32
11	1	0.16

Note. Diff = Rasch mean – 3PL mean

Table 4.8 shows a selection of schools with differences in mean scores. The schools with extreme scores tend to be schools with mean scores near 650 or above or schools with means between 550 and 600. This information agrees with the pattern shown in Figure 4.27.

Table 4.8

Selected schools with extreme differences in school 3rd grade ELA PASS means

School ID	Rasch Mean	3PL Mean	Diff	N
33327601	652.6	658.7	-6.1	33
34727015	659.3	664.7	-5.4	23
37927602	648.5	653.5	-5.0	22
34827012	654.0	658.1	-4.1	50
34027015	598.3	590.4	7.9	49
35827016	567.2	558.8	8.4	24
33427116	708.5	699.9	8.6	135
39000001	582.3	573.5	8.8	21
34931035	635.3	624.8	10.5	56

Note. Diff = Rasch mean – 3PL mean. N= number of students

Table 4.9 shows how districts differ in PASS mean scores. There are a couple of extreme cases, but for the most part, differences in PASS means scores for districts is minimal. There is only one district with PASS means that differ by more than 4 points.Table 4.10 shows that like the schools, districts that differ most in PASS means are districts with means between 550 and 600.

Table 4.9

Frequency table comparing difference in district PASS Means for 3rd grade ELA for Rasch and 3PL for 83 districts

Diff	Count	%
-3	1	1.20
-2	5	6.02
-1	18	21.69
0	31	37.35
1	14	16.87
2	9	10.84
3	3	3.61
4	1	1.20
9	1	1.20

Note. Diff = Rasch mean – 3PL mean

Table 4.10

Selected districts with extreme differences in district PASS means, 3rd grade ELA

District ID	Rasch Mean	3PL Mean	Diff	N
33230	592.1	588.1	4.0	56
39000	582.3	573.5	8.8	21

Note. Diff = Rasch mean – 3PL mean. N= number of students.

3rd Grade Math

Similar to 3rd grade ELA, Figure 4.30 shows that PASS scores for the Rasch versus 3PL model have the greatest differences near a score of 550. School and district means, shown in Figures 4.31 and 4.32 differ only slightly for 3PL versus Rasch. Figure 4.32 shows differences in district means near score 550.

Table 4.11 displays the frequency of differences in PASS means for the 3PL versus Rasch models for schools which is fairly small in most cases. Table 4.12 shows selected schools that the more extreme differences in PASS means. It can be seen that schools of various sizes are affected. School ID 34027015 has a large number of students and also a large change in PASS mean. The change appears to be likely due to the mean score being near 550 which is where the change in model is the most noticeable.

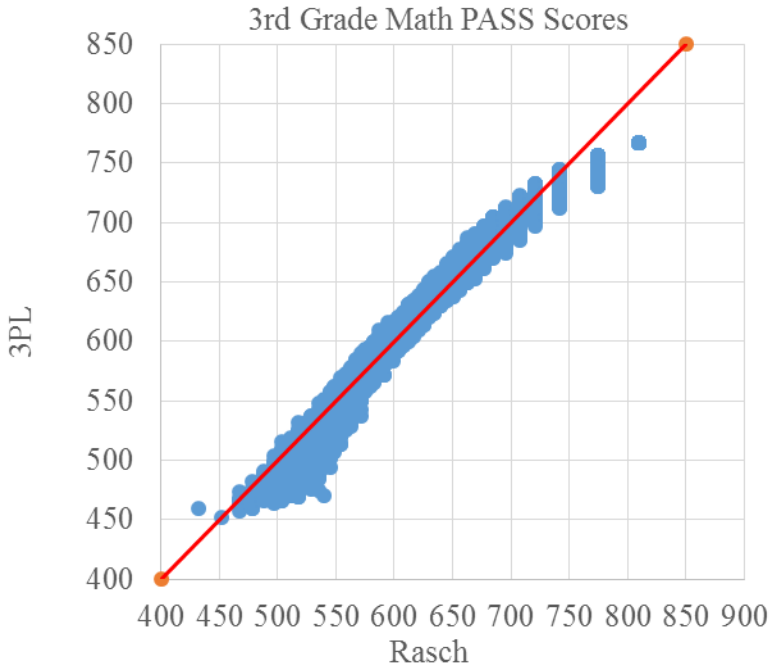


Figure 4.31. Scatterplot of school PASS scores, 3PL versus Rasch for 3rd grade Math with an "x = y" line of reference showing where the scores are equal. N = 53,829 students.

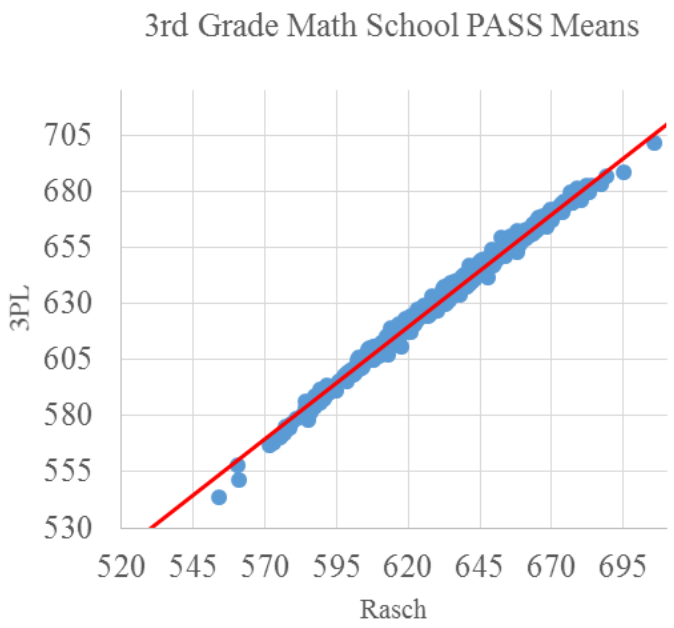


Figure 4.32. Scatterplot of school PASS mean scores, 3PL versus Rasch for 3rd grade Math with an "x = y" line of reference showing where the scores are equal. N= 634 schools

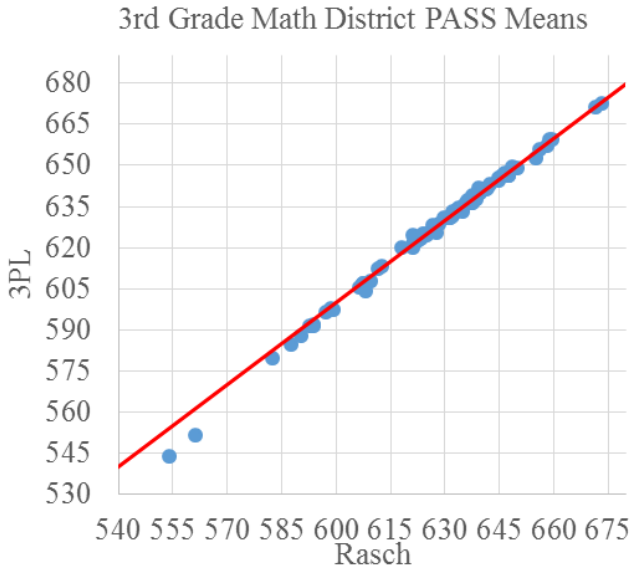


Figure 4.33. Scatterplot of district PASS mean scores, 3PL versus Rasch for 3rd grade Math with an "x = y" line of reference showing where the scores are equal. N = 83 districts.

Table 4.11

Frequency table comparing Rasch and 3PL school PASS means for 3rd grade Math

Diff	Count	%
-7	1	0.16
-6	1	0.16
-5	7	1.10
-4	15	2.37
-3	39	6.15
-2	78	12.3
-1	119	18.77
0	137	21.61
1	101	15.93
2	80	12.62
3	31	4.89
4	12	1.89
5	6	0.95
6	3	0.47
7	2	0.32
10	2	0.32

Note. Diff = Rasch mean – 3PL mean

Table 4.12

Selected schools with extreme differences in school 3rd grade Math PASS means

School ID	Rasch Mean	3PL Mean	Diff	N
37927602	652.6	659.5	-6.9	22
38727113	641.4	646.9	-5.5	19
35127094	649.1	654.2	-5.1	50
33427116	695.0	688.7	6.3	56
34931034	617.6	610.6	7.0	49
34027015	585.2	578.1	7.1	131
39000001	561.0	551.3	9.7	56
33230043	554.1	543.7	10.4	21

Note. Diff = Rasch mean – 3PL mean. N= number of students.

Table 4.13 shows that at the district level, the change in PASS means is very small for the most part for the 3PL versus Rasch model. A couple of exceptions are provided in Table 4.14, again with means near 550.

Table 4.13

Frequency table comparing difference in district PASS Means for 3rd grade Math for Rasch and 3PL for 83 districts

Diff	Count	%
-4	1	1.20
-2	4	4.82
-1	19	22.89
0	30	36.14
1	16	19.28
2	7	8.43
3	3	3.61
4	1	1.20
10	2	2.41

Note. Diff = Rasch mean – 3PL Mean

Table 4.14

Selected districts with extreme differences in district 3rd grade Math PASS means

District ID	Rasch Mean	3PL Mean	Diff	N
39000	561.0	551.3	9.7	21
33230	554.1	543.7	10.4	56

Note. Diff = Rasch mean – 3PL Mean. N= number of students

8th Grade ELA

Figure 4.34 displays a scatterplot of 3PL versus Rasch for 8th grade ELA students. Based on this scatterplot, 8th grade ELA appears to be less affected by the change in IRT model than either of the 3rd grade subjects. There is still a noticeable difference near

score 525 but it is more modest than the difference seen in the 3rd grade subjects. Figure 4.35 and 4.36 show that the difference in PASS means is almost negligible at the school and district levels for the 3PL versus Rasch model.

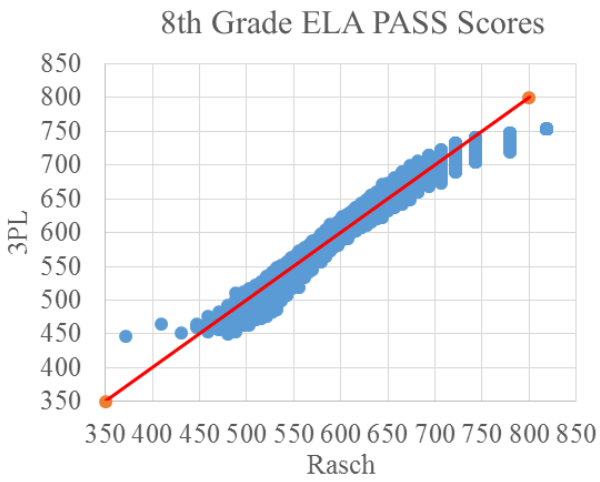


Figure 4.34. Scatterplot of PASS scores, 3PL versus Rasch for 8th grade ELA with an "x = y" line of reference showing where the scores are equal. N = 54,828 students.

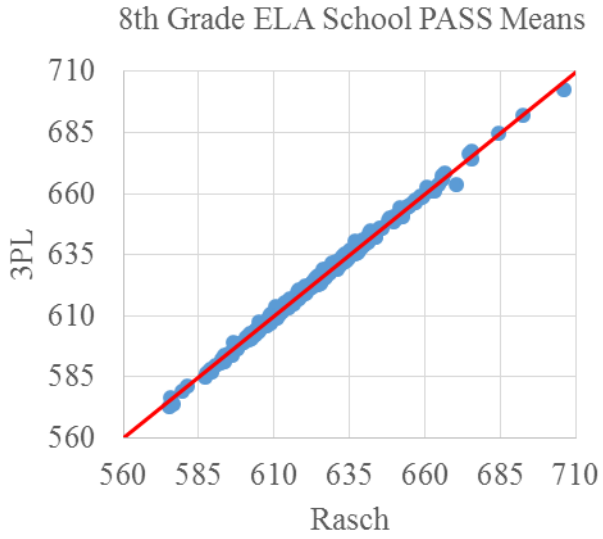


Figure 4.35. Scatterplot of district PASS mean scores, 3PL versus Rasch for 8th grade ELA with an "x = y" line of reference showing where the scores are equal. N = 83 districts.

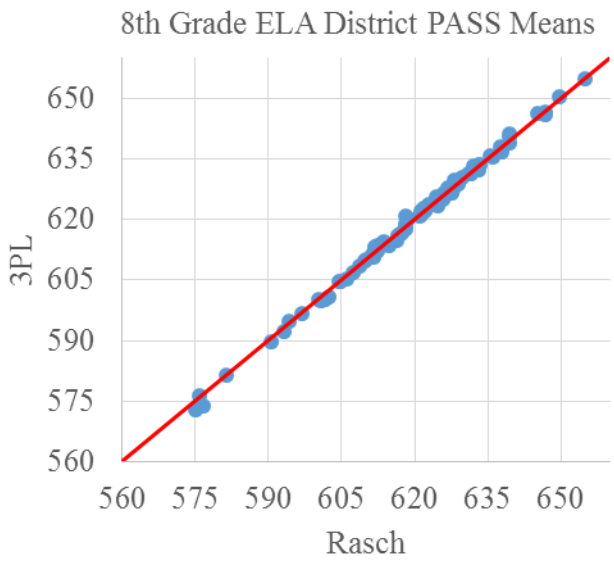


Figure 4.36. Scatterplot of district PASS mean scores, 3PL versus Rasch for 8th grade ELA with an "x = y" line of reference showing where the scores are equal. N = 301 schools.

Table 4.15 displays the frequency of differences in PASS means for the 3PL versus Rasch models for schools which shows slim differences in PASS means. Only 11 schools have more than a 3 point difference. Schools with the highest differences, presented in Table 4.16, occurred for schools with very high PASS means. Figure 4.34 indicates some differences at the high end for 8th grade ELA. Table 4.17 indicates that at the district level, the difference in PASS means is practically imperceptible for 8th grade ELA.

Table 4.15

Frequency table comparing Rasch and 3PL school PASS means for 8th grade English

Diff	Count	%
-3	9	2.99
-2	25	8.31
-1	67	22.26
0	103	34.22
1	69	22.92
2	23	7.64
3	3	1.00
4	1	0.33
7	1	0.33

Note. Diff = Rasch mean – 3PL Mean

Table 4.16

Selected schools with extreme differences in school 8th grade ELA PASS means

School ID	Rasch Mean	3PL Mean	Diff	N
33427116	705.9	702.4	3.5	71
35827006	670.3	663.7	6.6	23

Note. Diff = Rasch mean – 3PL Mean. N= number of students

Table 4.17

Frequency table comparing Rasch and 3PL district PASS means for 8th grade English

Diff	Count	%
-3	1	1.20
-2	3	3.61
-1	19	22.89
0	34	40.96
1	19	22.89
2	6	7.23
3	1	1.20

Note. Diff = Rasch mean – 3PL Mean

8th Grade Math

In Figure 4.37, the scatterplot for 3PL versus Rasch for PASS means shows large differences at the score 570 and also at the upper end, above 750. Recall from the results in Research Question 1, the distribution of Rasch θs was right skewed and therefore, so were the Rasch PASS scores. For 8th grade Math, scores at the upper end of the scale were quite different for Rasch versus 3PL even though the distribution of PASS scores for both models had the same mean (630.6) and same standard deviation (53.2). Perfect scores for Rasch resulted in a PASS score of 861 while perfect scores for 3PL resulted in PASS scores of 779. Table 4.18 shows the top 7 PASS scores for Rasch and for 3PL to give a better understanding of the difference in resulting scales. There are more extreme jumps for top scores for Rasch which has 60 unique PASS scores for 8th grade Math (because there are 60 questions) while 3PL has 285 unique PASS scores and a less "discrete" scale (recall that 3PL incorporates pattern scoring and therefore different scores can be awarded for the same number of total correct answers.) This pattern results in schools with higher PASS scores have more extreme differences in means for 3PL versus Rasch.

Table 4.18

Highest 7 PASS scores for the Rasch and 3PL model for 8th grade Math.

Rasch	3PL
861	779
824	773
790	770
770	769
755	768
743	766
733	765

Figure 4.38 shows schools with high PASS mean scores have lower means for 3PL than they do for Rasch.

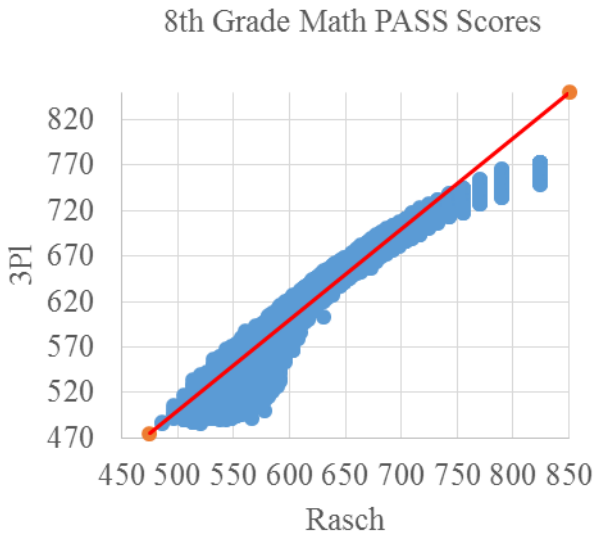


Figure 4.37. Scatterplot of PASS scores, 3PL versus Rasch for 8th grade Math with an "x = y" line of reference showing where the scores are equal. N = 54,885 students.

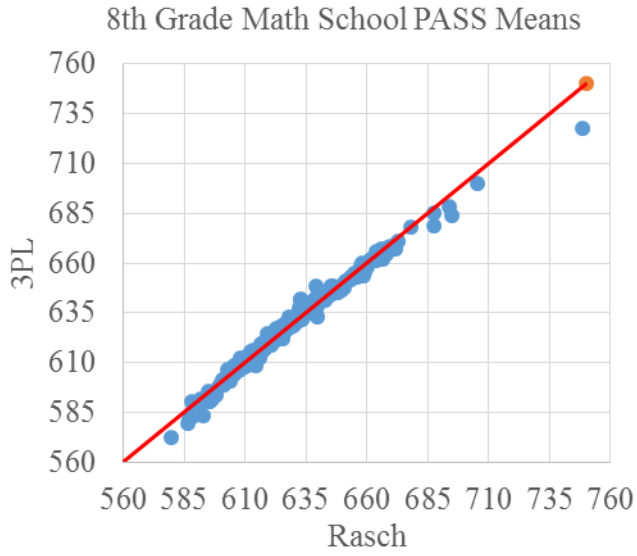


Figure 4.38. Scatterplot of school PASS mean scores, 3PL versus Rasch for 8th grade Math with an "x = y" line of reference showing where the scores are equal. N = 301 schools.

Figure 4.39 shows districts with low PASS mean scores have lower means for 3PL than they do for Rasch.

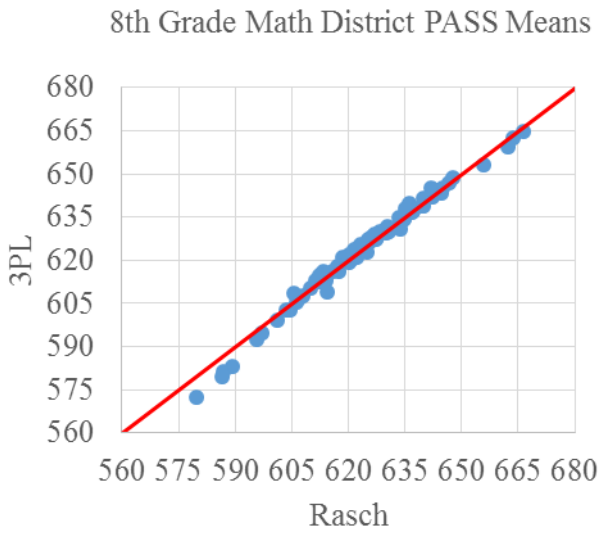


Figure 4.39. Scatterplot of district PASS mean scores, 3PL versus Rasch for 8th grade Math with an "x = y" line of reference showing where the scores are equal. N = 83 districts.

Table 4.19 shows some school means are affected by as much as 20 points by the change in IRT model. Table 4.20 shows small and large schools that have noticeable changes in PASS means.

Table 4.21 shows district means are less sensitive to the change in model than school means. Table 4.22 shows some districts with both low PASS means and high PASS means are impacted.

Table 4.19

Frequency table comparing Rasch and 3PL school PASS Means for 8th grade Math

Diff	Count	%
-10	2	0.66
-8	1	0.33
-6	1	0.33
-5	2	0.66
-4	7	2.33
-3	25	8.31
-2	48	15.95
-1	45	14.95
0	51	16.94
1	59	19.6
2	21	6.98
3	15	4.98
4	7	2.33
5	8	2.66
6	3	1.00
7	2	0.66
9	1	0.33
10	1	0.33
11	1	0.33
21	1	0.33

Note. Diff = Rasch mean – 3PL mean.

Table 4.20

Selected schools with extreme differences in school 8th grade Math PASS means

School ID	Rasch Mean	3PL Mean	Diff	N
33427617	632.6	642.1	-9.5	37
38727612	638.9	648.4	-9.5	13
35827002	632.9	641.3	-8.4	19
38527015	619.0	624.6	-5.6	21
34627025	632.1	637.3	-5.2	36
38127012	628.2	632.8	-4.6	258
39000001	579.7	572.6	7.1	23
38727113	687.5	678.7	8.8	11
35827007	593.0	583.3	9.7	40
35827006	694.9	683.8	11.1	23
33427116	748.5	727.5	21	71

Note. Diff = Rasch mean – 3PL mean. N= number of students.

Table 4.21

Frequency table comparing Rasch and 3PL district PASS Means for 8th grade Math

Diff	Count	%
-4	1	1.2
-3	4	4.82
-2	15	18.07
-1	15	18.07
0	19	22.89
1	15	18.07
2	5	6.02
3	4	4.82
5	1	1.2
6	2	2.41
7	2	2.41

Note. Diff = Rasch mean – 3PL Mean.

Table 4.22

Selected districts with extreme differences in district 8th grade Math PASS means

District ID	Rasch Mean	3PL Mean	Diff	N
32829	636.2	639.7	-3.5	138
38371	614.3	608.8	5.5	205
33827	589.1	583.3	5.8	157
33628	586.7	579.8	6.9	67
39000	579.7	572.6	7.1	23

Note. Diff = Rasch mean – 3PL Mean. N= number of students.

8th Grade Math with Equi-percentile rescaling

As discussed in Research Question 1 and again in the 8th grade Math section of Research Question 2, an equi-percentile rescaling method was also employed for 8th grade Math, noted as 3PL EQ%. This section shows comparisons of the Rasch model and 3PL EQ %. Figure 4.40 shows with 3PL EQ% instead of 3PL, the scatterplot of 3PL

EQ% versus Rasch at the student level is more evenly distributed around the reference "x = y" line but more spread out at the lower and higher PASS score levels. Figures 4.41 and 4.42 show that the school and district PASS means are essentially the same for Rasch and 3PL EQ%.

Table 4.23 provides a frequency table of differences showing that the differences range from only -4 to 4 for school PASS means. Table 4.24 shows impacted schools. Similarly, Table 4.25 shows district means are barely impacted by the change in model with 3PL EQ%. Table 4.26 shows the districts that are most affected but the difference is minimal.

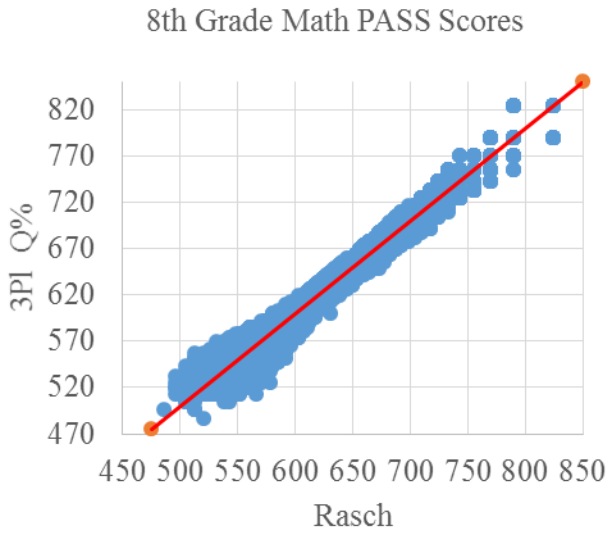


Figure 4.40. Scatterplot of PASS scores, 3PL EQ% versus Rasch for 8th grade Math with an "x = y" line of reference showing where the scores are equal. N = 54,885 students.

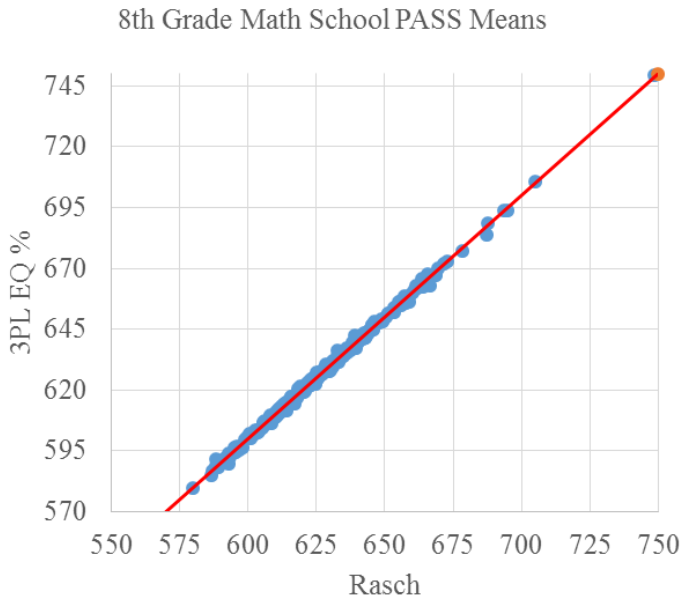


Figure 4.41. Scatterplot of school PASS mean scores, 3PL EQ% versus Rasch for 8th grade Math with an "x = y" line of reference showing where the scores are equal. N = 301 schools.

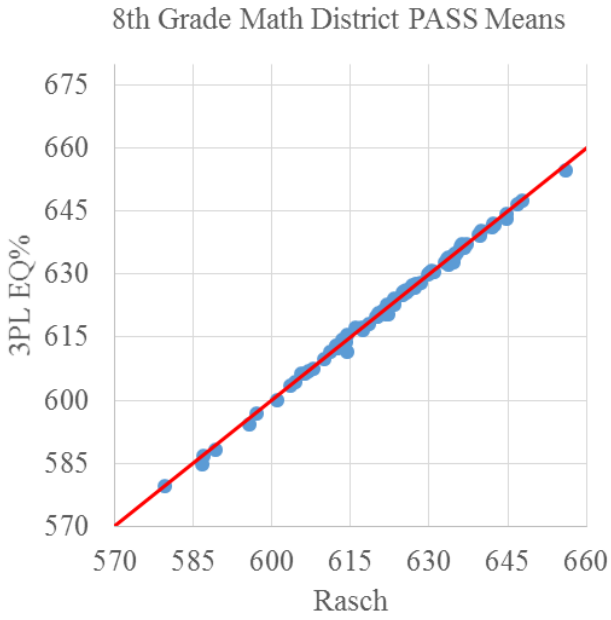


Figure 4.42. Scatterplot of district PASS mean scores, 3PL EQ% versus Rasch for 8th grade Math with an "x = y" line of reference showing where the scores are equal. N = 83 districts.

Table 4.23

Frequency table comparing Rasch and 3PL school PASS Means for 8th grade Math with equi-percentile rescaling

Diff	Count	%
-4	3	1.00
-3	1	0.33
-2	13	4.32
-1	79	26.25
0	116	38.54
1	70	23.26
2	14	4.65
3	3	1.00
4	2	0.66

Note. Diff = Rasch mean – 3PL Mean. N= number of students

Table 4.24

Selected schools with extreme differences in school 8th grade Math PASS means with equi-percentile rescaling

School ID	Rasch Mean	3PL Mean	Diff	N
37927602	588.1	591.8	-3.7	21
38727612	638.9	642.5	-3.6	13
33427617	632.6	636.2	-3.6	37
38727035	666.5	663.0	3.5	174
38727113	687.5	683.7	3.8	11

Note. Diff = Rasch mean – 3PL Mean. N= number of students

Table 4.25

Frequency table comparing Rasch and 3PL district PASS Means for 8th grade Math with equi-percentile rescaling

Diff	Count	%
-2	1	1.2
-1	13	15.7
0	49	59.0
1	14	16.9
2	5	6.0
3	1	1.2

Note. Diff = Rasch mean – 3PL Mean. N= number of students

Table 4.26

Districts with most extreme differences in district 8th grade Math PASS means with equi-percentile rescaling

District ID	Rasch Mean	3PL Mean	Diff	N
33628	586.7	584.7	2.0	67
38371	614.3	611.6	2.7	205

Note. Diff = Rasch mean – 3PL Mean. N= number of students**.**

Summary

Using a 3PL model instead of the Rasch model, the change in PASS means is minimal for most schools and districts. Schools with PASS means near the lower end of the score distribution, school means less than 600, appear to be the most sensitive to the change in model. Using an equi-percentile rescaling method for 8th grade Math (due to 8th grade Math's right skewed distribution) may remove much of the effect of model change. That is, using equi-percentile ranking forces the 3PL scores to have the discrete like distribution of Rash. Recall that with Rasch, there is one PASS score for every total score. For 8th Grade Math, this means there are 61 unique PASS scores. (There were 63 questions on the 8th grade Math exam but none of the examinees had a raw score of 0, 2, 3, or 4.) Before the equi-percentile rescaling was imposed in the 3PL PASS scores, there were 285 unique 3PL PASS scores for 3PL. Recall that 3PL utilizes pattern scoring and therefore examinees with the same total score can receive varying ability estimates and therefore different PASS scores.

Research Question 3

Research Question 3: Is the impact of the IRT model different among age groups?

The results presented for 3rd and 8th grade Research Questions 1 and 2 will be compared to address Research Question 3. There are no additional data results to present for Research Question 3. The comparison of 3rd and 8th grade results will be discussed in Chapter 5.

Research Question 4

The following section addresses Research Question 4:

> Is the impact of the IRT model different among subgroups (including a subgroup of students who received modifications or accommodations)?

The analysis for Research Question 4 includes comparing the means for different student demographic subgroups. The mean for all students is the same because the rescaling method matches the PASS scores on mean and standard deviation. However, means for subgroups could be different. Because the objective this study is to determine the impact of the change in IRT model at the decision-making level (i.e., the school and district level for PASS), subgroups were selected to reflect the subgroups represented on school and district federal report cards as closely as possible. Recall that on the federal report cards, schools may receive points or partial points based on performance of subgroups. Also, recall that a sample report card can be found in Appendix F. First means for 3PL versus Rasch are compared for subgroups for all students in each grade

and subject. Then, selected school and districts with extreme differences in selected subgroups are presented.

3rd Grade ELA

Table 4.27 shows that the means for most student subgroups are approximately the same for 3rd grade ELA. The subgroups that show a clear change in mean are the students with an individualized education plan (IEP) accommodations and the students who are English as a second language (ESL) beginners and ESL pre-functional.

Table 4.27

Rasch and 3PL PASS means for 3rd Grade ELA

3rd Grade ELA Subgroup	Rasch ELA mean	3PL ELA mean	Diff	N
All Students	649.4	649.4	0.0	53,731
Male	644.4	644.1	0.3	27,283
Female	654.7	654.8	-0.2	26,448
White	666.9	667.0	-0.1	27,988
African American	625.1	624.8	0.3	18,155
Asian	678.9	677.9	1.0	807
Hispanic	633.6	633.6	0.0	4,587
American Indian/Alaskan Native	639.8	638.7	1.1	189
Multi-ethnic	650.9	651.1	-0.1	1,931
IEP Flag	603.2	598.5	4.7	7,555
IEP Accommodation	584.0	576.9	7.1	5,131
ESL Accommodation	598.9	595.5	3.4	1,133
ESL Pre-functional	573.5	563.7	9.8	216
ESL Beginner	593.3	588.9	4.4	796
Subsidized Meals (Free)	629.6	629.5	0.1	29,898
Subsidized Meals (Reduced)	652.3	653.1	-0.8	3,287
Annual Measurable Objective (AMO)	640.0	640.0		

Note. ESL denotes English as a second language. IEP Flag indicates the students has been flagged as having an individualized education plan. IEP accommodations indicates that the student received accommodations on the PASS test due to the IEP.

3rd Grade Math

Table 4.28 shows that the means for most student subgroups are approximately the same for 3rd grade Math as well. Again, the subgroup that shows a substantial change in mean is the subgroup for the students with an IEP accommodation.

Table 4.28

Rasch and 3PL PASS means for 3rd Grade Math

3rd Grade Math Subgroup	Rasch Math mean	3PL Math mean	Diff	N
All Students	636.7	636.7	0.0	53,829
Male	636.2	635.9	0.3	27,333
Female	637.2	637.5	-0.4	26,496
White	654.7	655.0	-0.3	27,997
African American	609.9	609.5	0.4	18,176
Asian	678.5	676.2	2.2	832
Hispanic	625.4	625.8	-0.4	4,626
American Indian/Alaskan Native	625.3	626.6	-1.3	192
Multi-ethnic	636.6	637.1	-0.5	1,931
IEP Flag	593.3	589.3	4.0	7,550
IEP Accommodation	574.0	568.2	5.7	5,335
ESL Accommodation	599.6	598.9	0.8	1,419
ESL Pre-functional	578.1	573.8	4.3	283
ESL Beginner	592.9	591.4	1.5	808
Subsidized Meals (Free)	616.7	616.8	0.0	29,953
Subsidized Meals (Reduced)	639.8	640.5	-0.7	3,289
Annual Measurable Objective (AMO)	640.0	640.0		

Note. ESL denotes English as a second language. IEP Flag indicates the students has been flagged as having an individualized education plan. IEP accommodations indicates that the student received accommodations on the PASS test due to the IEP.

8th Grade ELA

Table 4.29 shows that the means for all subgroups are approximately the same for 8th grade ELA. The subgroup with IEP accommodations has a slight change in mean but it is not as substantial as the change for the 3rd grade subjects.

Table 4.29

Rasch and 3PL PASS means for 8th Grade ELA

8th Grade ELA Subgroup	Rasch ELA mean	3PL ELA mean	Diff	N
All Students	626.4	626.4	0.0	54,828
Male	617.9	617.6	0.3	27,830
Female	635.2	635.5	-0.3	26,998
White	642.0	642.1	-0.1	29,700
African American	602.5	602.3	0.2	19,085
Asian	656.3	655.8	0.5	781
Hispanic	616.4	616.8	-0.4	3,443
American Indian/Alaskan Native	619.1	619.3	-0.2	156
Multi-ethnic	630.5	630.5	-0.1	1,598
IEP Flag	568.2	565.1	3.1	6,688
IEP Accommodation	562.8	559.0	3.8	5,089
ESL Accommodation	578.7	577.1	1.6	552
ESL Pre-functional	536.2	529.3	6.9	125
ESL Beginner	554.9	551.2	3.7	307
Subsidized Meals (Free)	605.9	605.8	0.1	26,935
Subsidized Meals (Reduced)	625.8	626.4	-0.6	3,815
Annual Measurable Objective (AMO)	632.0	632.0		

Note. ESL denotes English as a second language. IEP Flag indicates the students has been flagged as having an individualized education plan. IEP accommodations indicates that the student received accommodations on the PASS test due to the IEP.

8th Grade Math

Table 4.30 shows that the means for most subgroups are approximately the same for 8th grade Math. The subgroup with IEP accommodations had the greatest change in mean as compared to the other grades and subjects. Again, we see a shift for ESL pre-functional and ESL beginners. In 8th grade Math, we also see a change in mean for the Asian subgroup. Note that this subgroup has the highest mean and is likely affected by the difference in the Rasch versus 3PL PASS scores for top scores addressed in Research Question 2.

Table 4.30

Rasch and 3PL PASS means for 8th Grade Math

8th Grade Math Subgroup	Rasch Math mean	3PL Math mean	Diff	N
All Students	630.6	630.6	0.0	54,885
Male	627.9	626.6	1.3	27,863
Female	633.4	634.8	-1.4	27,022
White	643.7	643.7	-0.1	29,699
African American	609.4	609.4	0.0	19,088
Asian	681.4	676.6	4.8	794
Hispanic	623.3	623.9	-0.6	3,484
American Indian/Alaskan Native	617.9	616.6	1.3	157
Multi-ethnic	633.5	633.8	-0.3	1,598
IEP Flag	582.8	574.1	8.7	6,682
IEP Accommodation	579.9	570.1	9.8	5,537
ESL Accommodation	598.1	595.0	3.1	686
ESL Pre-functional	576.8	564.4	12.4	176
ESL Beginner	586.2	580.1	6.0	310
Subsidized Meals (Free)	612.1	611.9	0.2	26,974
Subsidized Meals (Reduced)	628.3	629.3	-1.1	3,817
Annual Measurable Objective (AMO)	632.0	632.0		

Note. ESL denotes English as a second language. IEP Flag indicates the students has been flagged as having an individualized education plan. IEP accommodations indicates that the student received accommodations on the PASS test due to the IEP.

8th Grade Math with 3PL EQ%

Table 4.31 compares subgroups for Rasch and 3PL EQ%. With equi-percentile rescaling, none of the subgroups show substantial differences in PASS means for the whole state.

Table 4.31

Rasch and 3PL EQ% PASS means for 8th Grade Math with equi-percentile rescaling

8th Grade Math Subgroup	Rasch Math mean	3PL EQ% Math mean	Diff	N
All Students	630.6	630.6	0.0	54,885
Male	627.9	626.6	1.3	27,863
Female	633.4	633.6	-0.2	27,022
White	643.7	643.6	0.1	29,699
African American	609.4	609.4	0.0	19,088
Asian	681.4	682.2	-0.8	794
Hispanic	623.3	623.2	0.1	3,484
American Indian/Alaskan Native	617.9	617.2	0.7	157
Multi-ethnic	633.5	633.4	0.1	1,598
IEP Flag	582.8	581.5	1.3	6,682
IEP Accommodation	580.0	578.4	1.6	5,537
ESL Accommodation	598.1	597.7	0.4	686
ESL Pre-functional	576.8	574.8	2.0	176
ESL Beginner	586.2	585.4	0.8	310
Subsidized Meals (Free)	612.1	612.0	0.1	26,974
Subsidized Meals (Reduced)	628.3	628.1	0.2	3,817
Annual Measurable Objective (AMO)	632.0	632.0		

Note. ESL denotes English as a second language. IEP Flag indicates the students has been flagged as having an individualized education plan. IEP accommodations indicates that the student received accommodations on the PASS test due to the IEP.

Students with IEP Accommodations

Overall, for each grade and subject, students with IEP accommodations on the PASS exam appeared to be most sensitive to the change in IRT model. In order to see the impact of the change in IRT model at the school and district level for this subgroup, selected schools and districts are presented in Table 4.32. Note that the PASS mean for School ID 35127010 dropped by 30 points due to the change in IRT model for the IEP subgroup.

Table 4.32

Rasch and 3PL PASS means for subgroups of students with IEP accommodations on PASS for selected schools and districts with large differences.

Group	ID	Rasch Mean	3PL Mean		Diff	N
3rd Grade ELA						
School	35627024	578.1	557.2		20.9	12
District	34627	572.0	555.1		16.9	23
3rd Grade Math						
School	34027015	531.2	509.2		22.0	19
	34931034	560.8	541.6		19.2	13
District	34930	542.4	528.5		13.9	33
8th Grade ELA						
School	34227015	534.6	520.6		14.0	15
	33230042	528.1	512.8		15.3	8
	34930049	550.5	534.8		15.7	4
District	33230	528.1	512.8		15.3	8
	34227	534.6	520.6		14.0	15
8th Grade Math						
School	35127010	563.3	531.9		31.4	12
District	34930	572.2	551.5		20.7	20
	34227	566.3	545.3		21.0	17
	38371	572.0	549.9		22.1	12
8th Grade Math						
School	38527029	578.1	570.8	*	7.3	12
	34930049	573.7	564.2	*	9.5	6
	37927042	580.7	588.8	*	-8.1	6
District	32527	623.3	625.3	*	-2.0	44
	33628	589.5	587.0	*	2.5	61

Note. The "*" indicates 3PL EQ% rescaling. Diff = Rasch -3PL.

Summary

Students with IEP accommodations appear to be the most sensitive to the change in IRT model. Therefore, this subgroup was selected for a more in depth simulation study.

Simulation Study

Because the IEP accommodation group appeared to be the most sensitive to the change in IRT model, and because the most extreme difference occurred with 8th grade Math, the 8th grade Math IEP accommodation subgroup was selected for a simulation study.

Reason for Simulation Study

While working with the actual PASS student response matrix is beneficial because we are working with results that occurred in practice, a limitation is that we do not know definitively if the data resulted from a true Rasch or 3PL model. The advantage of a simulation study, is that a known model can be used to simulate response data and then we can fit the response data with different models and compare their results to see how well they match the true model. This analysis may help to select an IRT model when the true model is unknown.

Rasch as true model

In this study, the student abilities (θs) were estimated by fitting a Rasch model to the real response matrix. These θs, along with their associated standard error, were then used to generate a set of "true" Rasch the student abilities (θs). The "true" θs remained

linked to the student, school and district of the original data set. Also, the item parameters estimated by the Rasch model were treated as "true" item parameters. Using the "true" θs and "true" item parameters, a new response matrix was simulated. The simulated data was then fit with both a Rasch and a 3PL model to find new θ estimates. The estimated θs transformed to PASS scores were then compared to the PASS scores transformed from the "true" θs.

Figures 4.43 and 4.44 show that when the true Rasch data was fit with either the Rasch model or the 3PL model, the resulting estimated Rasch and 3PL PASS scores were about the same. The shapes of the scatterplots for Fitted Rasch versus True Rasch and Fitted 3PL versus true Rasch are very similar. This suggest that the Rasch model and the 3PL model fit the the true Rasch data similarly.

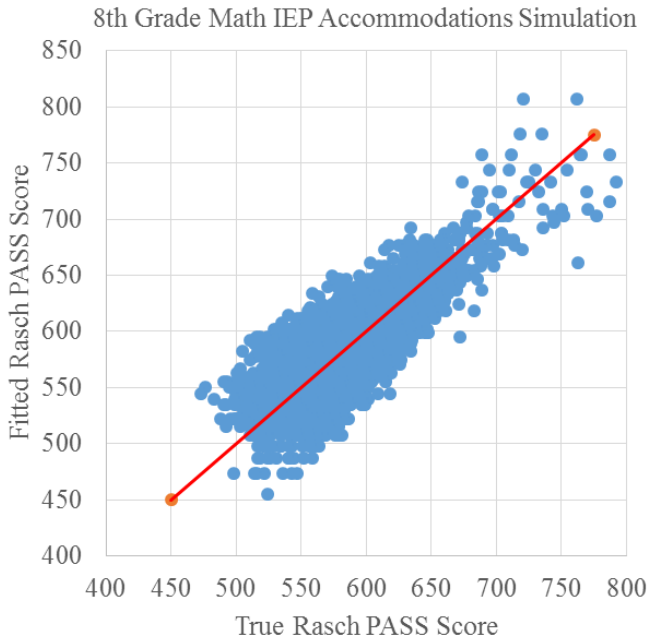


Figure 4.43. Comparison of true Rasch θs *transformed to PASS scores and the the resulting estimated* θs *transformed to PASS scores from the fitted Rasch model. An "x = y" line is provided for reference to show where the scores are equal.* N = 5,537 students

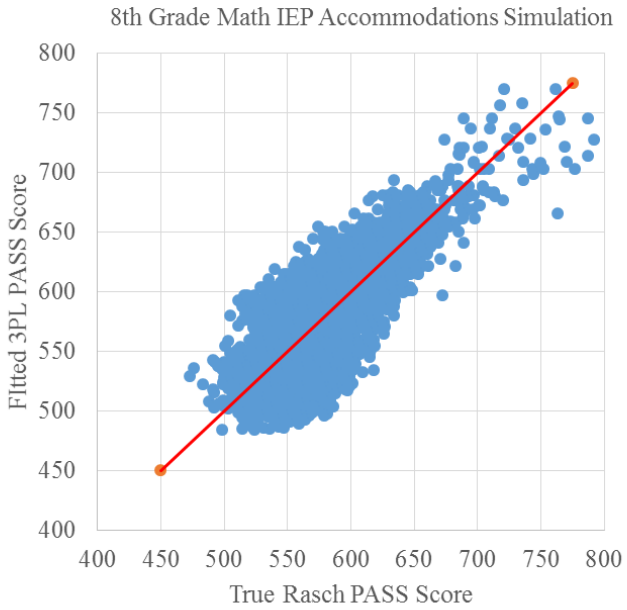


Figure 4.44. Comparison of true Rasch θs *transformed to PASS scores and the the resulting estimated* θs *transformed to PASS scores from the fitted 3PL model. An "x = y" line is provided for reference to show where the scores are equal.* N = 5,537 students

3PL as true model

Next, the student abilities (θs) were estimated by fitting a 3PL model to the real response matrix. These θs, along with their associated standard error, were used to generate a set of "true" 3PL θs. The "true" 3PL θs remained linked to the student, school and district of the original data set. Also, the item parameters estimated by the 3PL model were treated as "true" item parameters. Using the "true" 3PL θs and "true" item parameters, a new response matrix was simulated. The simulated data was then fit with both a Rasch and a 3PL model to find new θ estimates. Estimated θs were then compared to the "true" 3PL θs after transforming the θs to PASS scores.

Figure 4.45 shows the 3PL model estimates to be higher than the true 3PL PASS scores at the low end. The Rasch estimates in Figure 4.46 are higher at the low end than the true 3PL scores. Rasch estimates appear to be further away from the true 3PL scores at the low end than the 3PL estimates are from the true 3PL values.

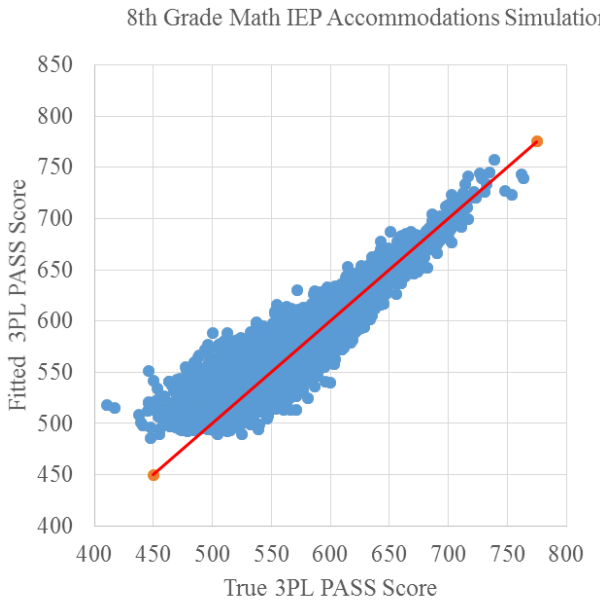


Figure 4.45. Comparison of true 3PL θ*s transformed to PASS scores and the the resulting estimated* θ*s transformed to PASS scores from the fitted 3PL model. An "x = y" line is provided for reference to show where the scores are equal.* N = 5,537 students

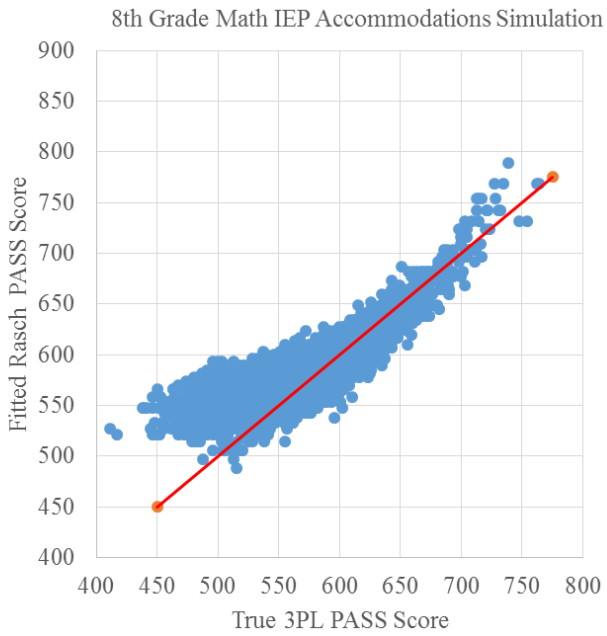


Figure 4.46. Comparison of true Rasch θs *transformed to PASS scores and the the resulting estimated* θs *transformed to PASS scores from the fitted 3PL model. An "x = y" line is provided for reference to show where the scores are equal.* N = 5,537 students

Simulation with Equi-percentile Rescaling

It was noted that many of the "true" scores that resulted from the simulation student resulted in unusually high or low PASS scores. This is because the true score simulation incorporated the standard error of the θs that were originally estimated and some of the standard errors were quite high. Therefore, the equi-percentile rescaling method was used again to put all of the PASS scores from the true and estimated θs on the original PASS scale. This is a fairly stringent rescaling method that may, in effect, remove the impact of the 3PL versus Rasch model by forcing 3PL PASS scores on to a more discrete scale as discussed with Research Question 2.

The simulation analysis was repeated using the equi-percentile scaling method. Figures 4.47 – 4.48 show that with equi-percentile rescaling, the Rasch and 3PL model

give similar results when Rasch is the true model. Also, Figures 4.49-4.50 show that with equi-percentile rescaling, the Rasch and 3PL models give similar results when 3PL is the true model.

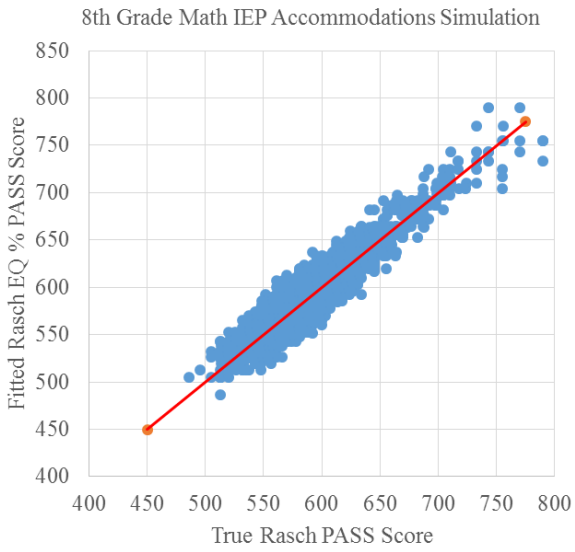


Figure 4.47. Comparison of true Rasch θs *transformed to PASS scores and the the resulting estimated* θs *transformed to PASS scores from the fitted Rasch model with equipercentil rescaling. An "x = y" line is provided for reference to show where the scores are equal.* N = 5,537 students

Table 4.33 provides summary statistics for the simulation study. Without equi-percentile rescaling, the Rasch model does not appear to estimate student ability well when 3PL is the true model. The true 3PL mean is 570 but estimated Rasch mean is 583. The 3PL estimated mean was off as well, at 577, but not as poorly fit as the Rasch model. Both the 3PL model and the Rasch model were close to matching the true Rasch mean. With equi-percentile rescaling, the Rasch and the 3PL model performed equally well when Rasch was the true model and also when 3PL was the true model.

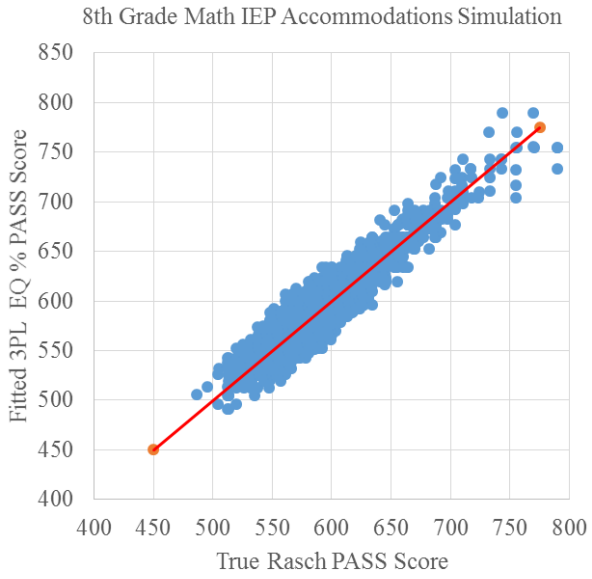


Figure 4.48. Comparison of true 3PL θs *transformed to PASS scores and the the resulting estimated* θs *transformed to PASS scores from the fitted Rasch model with equipercentil rescaling. An "x = y" line is provided for reference to show where the scores are equal.* N = 5,537 students.

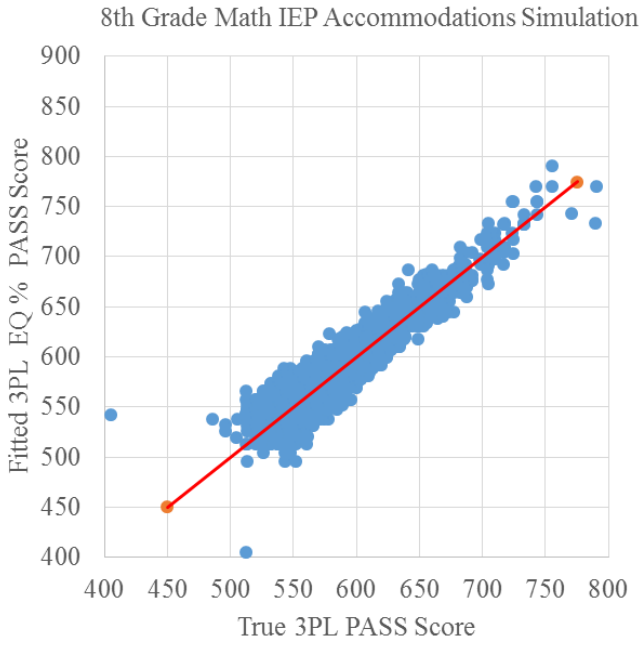


Figure 4.49. Comparison of true 3PL θs *transformed to PASS scores and the the resulting estimated* θs *transformed to PASS scores from the fitted 3PL model with equipercentil rescaling. An "x = y" line is provided for reference to show where the scores are equal.* N = 5,537 students.

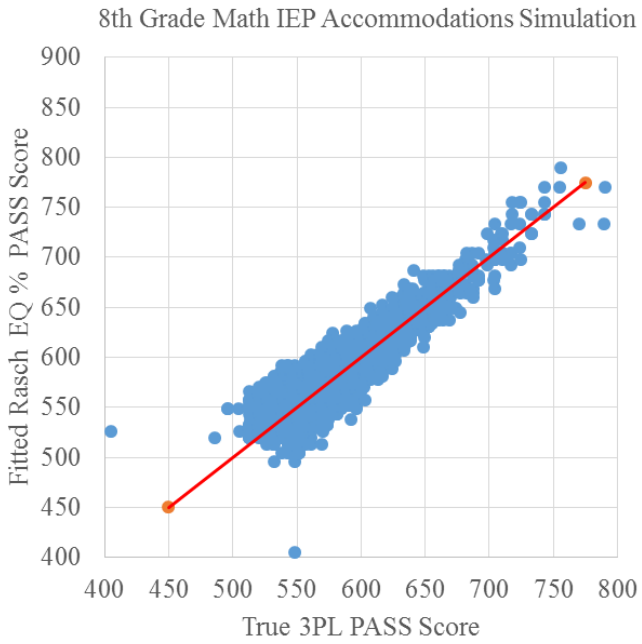


Figure 4.50. Comparison of true Rasch θs *transformed to PASS scores and the the resulting estimated* θs *transformed to PASS scores from the fitted 3PL model with equipercentil rescaling. An "x = y" line is provided for reference to show where the scores are equal.* N = 5,537 students

Table 4.34 shows the mean of the differences for districts. Again without equi-percentile rescaling, the Rasch model estimates result in large differences when 3PL is the true model.

Table 4.33

Summary statistics for the simulation study of 8th Grade Math students with IEP accommodations

Model	*M*	*SD*	Minimum	Q1	Median	Q3	Maximum
M/SD Rescale							
True Rasch	580	34	473	558	577	597	792
Fitted Rasch	584	34	455	563	582	602	807
Fitted 3PL	580	40	484	553	580	604	770
True 3PL	570	46	411	538	569	601	820
Fitted 3PL	577	40	486	548	573	603	781
Fitted Rasch	583	32	488	562	578	600	859
EQ% Rescale							
True Rasch	582	32	486	561	578	596	790
Fitted Rasch	584	32	486	566	581	600	790
Fitted 3PL	584	32	491	561	581	600	790
True 3PL	581	32	405	561	578	596	861
Fitted 3PL	583	32	405	561	578	600	861
Fitted Rasch	583	32	405	561	578	600	861

Note. N = 5,537 students.

Table 4.34

Summary of District Differences 8th Grade Math with Accommodations

Rescaling Method	True Model	Fit Model	*M*	*SD*	Minimum	Q1	Median	Q3	Maximum
M/SD									
	Rasch	Rasch	5	4	-4	2	4	5	20
	Rasch	3PL	1	4	-9	-2	0	2	18
	3PL	3PL	8	5	-1	5	7	10	29
	3PL	Rasch	15	7	1	10	14	18	41
EQ%									
	Rasch	Rasch	3	4	-5	1	2	3	17
	Rasch	3PL	3	3	-5	1	2	3	16
	3PL	3PL	3	3	-3	1	2	4	17
	3PL	Rasch	3	4	-4	0	2	4	18

Note. Differences calculated as the Fit Model – True Model. N=73 districts.

Summary

This study used estimated abilities from an 8th grade Math subgroup of students with IEP accommodations which appeared to be sensitive to the change in IRT model from Rasch to 3PL to conduct a simulation analysis. The Rasch and 3PL models performed about the same for matching true Rasch model results. However, when 3PL was the true model, 3PL estimates more closely matched 3PL true values than Rasch estimates. With equi-percentile rescaling, Rasch and 3PL estimates matched Rasch true values very closely. Also, with equi-percentile rescaling, Rasch and 3PL estimates matched 3PL true values very closely.

CHAPTER 5

DISCUSSION

The purpose of this study was to investigate the impact of the IRT model used in the analysis of statewide assessment data with the intention of acquiring knowledge to increase the likelihood that valid interpretations are drawn from assessment results. The data used for this dissertation was the scored student response matrix from the 2014 administration of South Carolina's PASS statewide assessment. The study involved analyzing the student response matrix with the Rasch model, the IRT model used in practice by SCDE and many other states for other statewide assessments. The data was also analyzed with the 3PL model, another popular IRT model used in a large number of states for educational statewide assessments. Unlike Rasch, the 3PL models accounts for varying item discrimination and guessing.

Model fit checks investigated the fit of the Rasch and 3PL models. Resulting student PASS scores for both models were then summarized and compared at the school and district level. The study was unique because it used real statewide assessment student responses (as opposed to simulated data) and because the analysis was performed at the school and district level. It was established in Chapter 2 that many decisions and interpretations made from statewide assessments occur at the school and district level. Therefore, the analysis at this level contributes greatly to the validity evidence for

statewide assessments.Reporting of PASS statewide assessment data at the school and district level centers around state and federal report cards. The analyses focused on the following research questions which all relate to state and federal report cards:

Research Question 1

If a different IRT model were used to score (i.e., calibrate and scale) student responses on PASS, how would state school reports cards be affected? Note that school report cards are based on the percentage of students scoring in the 'Not Met,' 'Met,' and 'Exemplary' category in each subject.

Research Question 2

If a different IRT model were used to score (i.e., calibrate and scale) student response on PASS, how would federal school reports cards be affected? Note that school report cards are based on the mean score for each subject.

Research Question 3

Is the impact of the IRT model different among age groups?

Research Question 4

Is the impact of the IRT model different among subgroups (including a subgroup of students who received modifications or accommodations)? Note that federal report cards report PASS means for subgroups.

Additionally, a simulation study was conducted on a subgroup of students with IEP accommodations for 8th grade Math. This group was found to be particularly

sensitive to the change in IRT model. With the simulation study, student responses were simulated from a known model and then fit with the Rasch and 3PL. Student ability estimates from the fit models were compared to the known abilities used for the simulation. The known abilities used for the simulation characterized the subgroup with IEP accommodations because they were generated using the student abilities and standard errors estimated from the real data.

Findings

Model Fit Checks

As discussed in Chapter 2, the basis for the Rasch model is that the PASS assessment was constructed for Rasch measurement. That is, the assessment was designed such that item discrimination would be consistent among all items and that guessing on items would not be a factor. Pilot studies for PASS were conducted to examine item difficulty and item discrimination (SCDE, 2012) and presumably, Rasch fit statistics. Presumably, based on the pilot studies, only items appropriate for the Rasch model were included on the PASS exam. The Rasch model is attractive due to its simplistic interpretation; the total score is a sufficient statistic for Rasch. That is, students with the same total score have the same estimated student ability (θ) and ultimately the same PASS score. This straight forward measurement is easy to explain to stakeholders. Additionally with Rasch, item difficulty and student ability parameters are on the same scale.

However, there were concerns over the reality of "no guessing" on a multiple choice test as well as the expectation that all items would discriminate equally. A model

fit check was used in this study to check the fit of the Rasch model to the student response matrix resulting from the actual administration of of PASS in 2014. A person goodness of fit index, z_h (Drasgow et al., 1985) was used to examine the fit of the Rasch and 3PL IRT models on the PASS student response matrix. Findings indicated that both the Rasch and the 3PL model fit well for high ability examinees and for most middle ability level examinees. However, for low ability level examinees, the 3PL model fit well but the Rasch model had a poor fit. The finding was not unexpected; it is reasonable to assume that low ability examinees would be more likely to guess on items and the 3PL model is structured to adjust for guessing while the Rasch model is not.

Implications of findings for practice

Because the 3PL model was found to fit as well as or better than the Rasch model at all ability levels, state officials should consider the 3PL for analysis as opposed to Rasch. Alternatively, the PASS assessment items may need additional testing to ensure Rasch measurement is an appropriate IRT model to analyze student responses on PASS, especially at low ability levels. Because the Rasch model is a version of the 3PL model (with the guessing parameter equal to zero and the same item discrimination parameter for all items), it seems reasonable to fit the PASS response data with the 3PL model.

Research Question 1

For Research Question 1, it was determined that most schools and districts would have only minor shifts for the percentage of students in performance categories for the Rasch versus 3PL model. Generally, these shifts tended to be within a 5% range and thus not a striking change on state report cards. The greatest differences for 3PL versus PASS

scores were for students scoring below 600. Since 600 was the cutoff for the "Met" category, the differences did not occur as much near key PASS scale values for performance categories.

However, for small schools in particular, a small change in the number of students in a particular category had a big effect on the percentage in category report on state report cards. Communication with various school officials confirmed that the percentage of students falling in the "Not Met" category has the greatest implications for schools and districts. Therefore, this discussion focuses on the impact in the "Not Met" category. Table 5.1 shows the percentage in the "Not Met" category can be as much as 25% higher for 3rd grade Math at the school level for Rasch. Also, the percentage in the "Not Met" category can be as much as 8% lower for Rasch in 3rd grade ELA for a school. With the equi-percentile rescaling method discussed in Chapter 4, percentage in the "Not Met" category could be as much as 13% lower for Rasch or as much as 14% higher for Rasch versus 3PL EQ%. The importance of these findings is that for some schools, state report cards could appear very different due to the IRT model. Consequently, decistions for those schools based on test scores such as curriculum changes or eligibility for grant funding could be impacted by the IRT model.

Another finding with Research Question 1, was that students who "change places" due to the IRT model in the performance categories were not reflected on the percentage in category state card reporting method. For example. Table 4.6 shows that for 8th grade Math, 1,046 student who fell in the "Not Met" category for Rasch moved to the "Met" category for 3PL. Meanwhile, a different 1,096 students who were in the "Met" category

for Rasch, fell into the "Not Met" category for 3PL. Some other less extreme "swaps" were found for the other grades and subjects.

Table 5.1

Range for differences in school and district percentage in the "Not Met" category with a change to the 3PL model.

Grade	Subject	School level lowest difference	School level highest difference	District level lowest difference	District level highest difference
3	ELA	-8%	9%	-5%	4%
3	Math	-5%	25%	0%	8%
8	ELA	-4%	8%	-4%	8%
8	Math	-4%	19%	-4%	7%
8	Math*	-13%	14%	-8%	7%

Note. Order of subtraction is Rasch – 3PL. The "*" indicates 3PL EQ%.

Implications of findings for practice

For most schools and districts the impact of the IRT model did not have striking effects on state report findings and therefore may not affect decisions made from state school and district summaries. However, some small schools, where just a couple of students changing performance category greatly affects percentages, could be largely impacted by the choice if IRT model. Also, at the student level, thousands of students "swap" between the "Not Met" and "Met" category. Students selected for certain programs based on performance level, such as Multi-tier System of Support (MTSS),

could be largely impacted by the change in IRT model. Similarly, 3rd grade ELA student mandated to be retained by the Read to Succeed Law (2014) based in 3rd grade ELA scores would be impacted by a change IRT model. Based on this study, applying the 3PL model provides a better person-fit and therefore would increase validity evidence for decisions made regarding individual students and small schools.

Research Question 2

For Research Question 2, it was determined that the vast majority of schools and districts overall have similar PASS means for each school and district for Rasch versus 3PL. Chapter 4 results showed that in most cases school Rasch and 3PL PASS means were within 5 points of each other. Third grade Math only had 9 schools and 2 districts with Rasch and PASS means that differed more than 5 points while 3rd grade ELA had 8 schools and 1 district that differed more than 5 points. For 8th grade ELA, all district Rasch and 3PL PASS means were with 5 points of each other and only 2 schools differed by more than 5 points. Eighth grade Math had 13 schools and 4 districts that differed by more than 5 points but with 3PL EQ%, no schools or districts differed by more than 5 points.

A key value on federal report cards for school and district PASS means was the Annual Measurable Objective (AMO). Schools received points on their composite index score on the federal report card for meeting the AMO. Table 5.2 provides the AMO values for each grade and subject. Table 5.2 shows that the 3PL versus Rasch model did not appear to affect whether schools or districts met or did not meet the AMO. The schools and districts matched for 3PL versus Rasch on meeting the AMO for the mean of

all of their students almost 100% of the time for each grade and subject. It did not appear that Rasch versus 3PL affected whether a school or district met the AMO in almost all cases for the mean of all students at the school and district level.

Table 5.2

Annual Measureable Objective for PASS 2014

Group	2014 AMO	% Schools met AMO (Rasch)	% Schools met AMO (3PL)	% Matches (Schools)	% Districts met AMO (Rasch)	% Districts met AMO (3PL)	% Matches (Districts)
3rd Grade ELA	640	58%	58%	**97%**	59%	59%	**100%**
3rd Grade Math	640	40%	40%	**97%**	23%	25%	**98%**
8th Grade ELA	632	32%	32%	**99%**	19%	20%	**99%**
8th Grade Math	632	36%	36% 38%*	**99%** **98%***	31%	29% 31%*	**98%** **100%***

Note. * Indicates results with 3PL EQ%.

Furthermore, for schools that fell below the AMO, partial points are awarded using a composite index system based on quartiles between 600 and the AMO (SCDE, 2014). For example, in 8th Grade Math, a mean PASS score (below the AMO of 632) of 624.2-631.4 would receive .9 points. The quartiles were either 8 point (for middle school) or 10 point (for elementary school) ranges. Since we have found that almost all schools and districts were within 5 points of each other for Rasch versus 3PL, the IRT model was not expected to heavily impact partial point awards.

Partial points were also awarded for PASS means below 600 based on improvement from the previous year as opposed to the PASS mean value. The schools and districts with the greatest differences for PASS presented in Chapter 4 tended to have PASS means either over the AMO or below 600. Thus, schools and districts with different PASS means based on the IRT method either received the full 1 point credit or received partial points based on improvement from the previous year instead of the current PASS mean.

Implications of findings for practice

It does not appear the the IRT model would affect decisions made at the school or district level that are based on the PASS mean for all students. Schools with low PASS scores, below 600, seemed to be more sensitive to the change in IRT model. On federal report cards these schools and districts received partial points based on the improvement from the previous year. A future study might include the impact of the IRT model with linking and equating from previous years on school and district PASS means.

Also schools with high PASS means for all students were sensitive to the change in model. However, these school tended to be above the AMO for both Rasch and 3PL and therefore would have received the full point for meeting the AMO goal regardless of method used.

Research Question 3

In order to determine if one age group was more sensitive to the change in IRT model, the results for Research Question 1 and Research Question 2 were compared for 3rd and 8th grade.

Third Grade

In general, 3rd Grade ELA and 3rd Grade Math responded similarly to the change in IRT model. In both cases, districts and schools tended to have a slightly greater percentage of students in the "Exemplary" category with the 3PL model. With 3rd Grade ELA, as shown in Figures 4.6 and 4.7, schools and districts tended to have more students in the "Met" category with Rasch. With 3rd Grade Math, as shown in Figures 4.11 and 4.12, schools and districts tended to have more students in the "Not Met" category with Rasch. As shown in Tables 4.2 and 4.3, at the student level, both subject areas had between 3,000 and 4,000 students overall who changed performance levels. 3rd grade Math and 3rd Grade ELA both had approximately 9 schools and 2 districts with Rasch and 3PL PASS means that differed more than 5 points.

8th Grade

The most striking difference between 3rd grade and 8th grade is that while 3rd grade subject areas responded similarly to the change in IRT model, the 8th grade subject areas were impacted differently by the change in IRT model. Of all grades and subject areas, 8th grade ELA appeared to be the least sensitive to the change while 8th grade Math was the most sensitive.

For 8th grade ELA, at the school and district level, there was little change, on average, for the percentage of students in performance categories. All PASS district means were within 3 points for 3PL versus Rasch. All PASS school means were within 4 points for 3PL versus Rasch with one exception; a small school with a couple of very

high Rasch PASS scores compared to 3PL. At the student level, between 3,000 and 4,000 students changed performance levels, just as we saw for 3rd grade. The scatterplot for 3PL versus Rasch, Figure 4.34, shows less discrepancy between Rasch and 3PL than the scatterplots for the other grades and subjects.

Meanwhile, 8th grade Math (partially due to the right skewed distribution of Rasch PASS scores for 8th grade Math discussed in Chapter 4, along with the extreme large jumps in top Rasch scores presented in Chapter 4) showed the most contrasting results for 3PL versus Rasch. This discrepancy is visible in Figure 4.37, the scatterplot of 3PL versus Rasch scores which shows differences for 3PL versus Rasch near score 575 and also above 750. Also, cases of schools with different means for 3PL versus Rasch were more extreme for 8th grade Math than for the other grades and subjects. One school's PASS mean differed by 21 points. School ID 33427116 with 71 students had a Rasch PASS mean of 748.5 and 3PL PASS mean of 727.5. Both scores were above the AMO and therefore the IRT model would not affect the score on the federal report card. On the other end, School ID 35827007 with 40 students in 8th grade Math, had a Rasch PASS mean of 593.0 and a 3PL PASS mean of 583.3. In this case, both PASS scores were below 600 and the impact on the report card would be based on improvement from last year.

However, with equi-percentile rescaling, the impact of the IRT model for 8th grade Math is greatly reduced. As discussed in Chapter 4, the equi-percentile rescaling method is rather stringent and forces the 3PL PASS scores onto the discrete Rasch scale.

Summary

The initial idea for Research Question 3 was to compare 3rd grade to 8th grade with the theory that one age group might guess more than another age group. However, the findings indicated that there is more of a discrepancy between Math and ELA for 8th grade than there is for 3rd Grade. Eighth grade ELA generally showed about the same or even less sensitivity to the change in IRT model as both 3rd grade subject areas. Eighth grade Math shows the most sensitivity to the change in IRT model.

However, 8th grade Math had characteristics such as a right skewed distribution for Rasch PASS scores and extreme jumps in Rasch PASS scores for top scores that were not present for the 3PL PASS scores. These factors may have contributed to the sensitivity of the IRT model change more so than guessing. Although, looking back at the item parameters for each of the grades and subjects, the mean guessing parameter (c parameter) for 8th grade Math items was higher than any of the other areas. Eighth grade Math had an average guessing parameter of .20 while 3rd Grade ELA, 3rd Grade Math, and 8th Grade ELA all had average guessing parameters of .14 or .15. This suggests that guessing could be more prevalent for 8th grade Math than for the other subjects. However, using equi-percentile re-scaling for 8th grade Math removed the impact of the IRT model.

Implications of findings for practice

By subject and and grade level, 8th grade Math appeared to have the greatest impact for the 3PL versus Rasch model. Therefore, it appears that 8th grade Math should be given priority for a review in the utilization of the Rasch model. Because the 3PL

model estimated an overall guessing parameter that was higher than for the other grades and subjects, test items should be re-examined for Rasch fit for 8th grade Math especially. The finding that students guess more on 8th grade Math as estimated by the 3PL model could indicate that students do not know the material as well and may inform curriculum development. The result is potentially related to the variety of course placement for students in 8th grade Math: General Math, Pre-Algebra, or Algebra. Because person-fit was better for the 3PL model, state contractors should consider using the 3PL model instead of Rasch.

Additional equating methods might be employed for future studies to compare Rasch to 3PL. Smoothing (Livingston, 2004) or kernel equating (Davier, Holland, & Thayer, 2004) may provide better solutions for equating the discrete-like Rasch scale with the more continuous 3PL scale. Future studies could investigate more sophisticated rescaling methods and also explore the implications of the nature of the discrete-like versus more continuous scale.

The distribution of Rasch scale PASS scores at the high end is concerning. On the Rasch PASS scale used in practice, a perfect score of 63 on 8th grade Math receives a PASS score of 881 while a total score of 62 receives a PASS score of 824. This is a 57 point jump for a difference of one question! Considering that the "Met" category for 8th grade Math only has a 56 point range (from 600 (total score 27) to 656 (total score 43)), a 57 point jump for one question seems extreme! For 3PL, the difference between a perfect score and the next highest score is only 9 points.

For Rasch, perfect scores could easily inflate the mean for a small school or especially a small class. If the mean PASS score was being used to evaluate teachers, an inflated mean due to a perfect score could be misleading. The standard deviation should be reported in addition to the mean so the variability among student scores can be taken into account when comparing small schools or classes. It is also probably worthwhile to compare students with perfect scores to students who missed one question on other assessment measures for evidence of concurrent validity for extreme differences in PASS score. For example, studies might compare the performance of students with perfect PASS scores on routine school assessments to the performance of students who missed one question on PASS to find evidence (or lack there of) for perfect scores on state assessments warranting a 57 point difference from those who missed on question.

Research Question 4

With regard to subgroups, it is clear that students who are ESL beginners, ESL pre-functional, or students with IEP accommodations on the PASS test are the most sensitive to the change in IRT model. Students in these subgroups tended to score in the range of PASS scores (roughly between 550 and 600) where there were the largest differences between Rasch and 3PL. It is reasonable to infer that students at the lower ability level would be more likely to guess and therefore it is not surprising that subgroups with score ranges on the lower end would be the most affected by Rasch versus 3PL since 3PL accounts for guessing.

Because the ESL subgroups are fairly small, the analysis for this dissertation focused on the subgroup with IEP accommodations. As shown in Chapter 4, for each grade and subject area, the mean for the entire state was approximately 10 points lower for 3PL PASS than for Rasch PASS for students with IEP accommodations.

Recall that the weight for subgroups on federal report cards as has much bearing on the composite index score as the mean for all students in the school. This finding could have a large impact on federal report cards which contains a 'With Disability' subgroup determined by instructional codes on PowerSchool (SCDE, 2014). Presumably, many of the students with IEP accommodations coded on PASS would fall into the "With Disability" subgroup on the federal report card.

As an example of how this could affect federal report cards, consider 3rd grade ELA. For 3rd Grade ELA, 417 schools had 9 or more students with an IEP accommodation. On federal report cards, 30 students are needed in the subgroup to count on the federal report card composite index system for the school or district. Since there is only 3rd grade data in this study (without 4th and 5th grade), let us assume that about 10 students in 3rd grade alone is enough to count as a subgroup. A total of 87 of the 417 schools with IEP subgroups had Rasch PASS means that were at least 10 points higher than 3PL PASS means. Most of these schools had PASS means that were below 600 and therefore their partial point on the composite index system would be determined by improvement from last year. About 10 of these schools had PASS means over 600 and the partial point would therefore be determined by the quartile system discussed with

Research Question 2. Again, with equi-percentile equating used for 3PL on 8th grade Math, the effect on IEP accommodations was removed.

Implications of findings for practice

Pilot testing for PASS and PASS-like assessments should carefully consider item functioning with the Rasch model for students with disabilities or consider the 3PL model. This would be particularly important for students with IEPs who scored, on average, closer to 550 than other subgroups. A PASS score of 550 was the cut off for 3rd grade ELA for "Not Met 1" (the lower end of the "Not Met" performance level) which is the marker for students who need to be retained based on the Read to Succeed Act (2014). For these students, the IRT model selected could result in whether the student is retained or not. Since the 3PL model is a better for for examinees, especially at this ability level, using the 3PL model would strengthen the validity evidence for the decisions students who should be retained. Alternatively, PASS-like assessments should be re-evaluated for Rasch fit and pilot testing should ensure students with IEPs are included in the pilot.

Simulation Study

Students with an IEP accommodations for 8th grade Math (IEP subgroup) were chosen for a simulation study. The Chapter 4 results of the simulation study show that when Rasch was the true model, both the 3PL model and the Rasch model were fairly accurate in estimating PASS scores that resulted from student ability estimates. However, when 3PL was the true model, the fit 3PL model estimated student abilities and

resulting PASS scores were closer to the "true" abilities and resulting PASS scores than those of the fit Rasch model.

An issue with the simulation study was that the "true" θs were generated as a normal distribution from the estimated θs and standard errors from the actual PASS responses matrix. The standard errors were rather large. For example, the most extreme estimated θ from the real PASS data was -4.26 with and standard error of 1. Therefore the resulting simulated true theta was expected to range within 3 standard errors of -4.26 (between -7.26 and -1.26). This translates to PASS scores ranging from 263 to 546. While in practice PASS scores can range from 300 to 900, the Rasch PASS scores for 8th grade Math in this study ranged from 405 to 861. Equi-percentile rescaling was used to place all simulated true θs and model fit θs and resulting PASS scores on the Rasch PASS scale. In this case, when 3PL was the true model, both the Rasch model and the 3PL fit model estimated the true student abilities accurately.

Implications of findings for practice

In every case scenario from the simulation study, the 3PL model appears equal to or better than the Rasch model for estimating student abilities that most closely match the true abilities. From this standpoint, it seems logical to use the 3PL model when the true model is unknown. A disadvantage of this would be loss of simplicity offered by the Rasch model.

In order to strengthen validity evidence for interpretations of the IEP subgroup, additional studies are needed to determine if Rasch is an appropriate model. The rescaling methods made big differences in the findings. The additional studies could

examine other options for rescaling methods to compare the 3PL versus Rasch models in simulation studies.

Suggestions for future studies

Additional studies could contribute to the comparison of Rasch versus 3PL in statewide assessments. One area that needs further investigation is the impact of the IRT model on linking from year to year that was not addressed in this study. For PASS, year to year linking might affect partial points awarded for improvement for groups falling below 600 on PASS on the federal report card composite index system.

Another area for further research includes determining rescaling methods to compare 3PL scores (which are more continuous) to Rasch scores (which are more discrete). This study mainly used a common mean and standard deviation rescaling method because the transformation of θ to PASS scores is mean and standard deviation based and also because the data sets were normally distributed. The exception was 8th grade Math which was right skewed. Here, an equi-percentile rescaling method was also employed. However, with this method, the 3PL scores were transformed to the Rasch PASS scores based on rank. That is, examinees were rank-ordered by their 3PL PASS scores and then given the same Rasch PASS score as the examinee in their corresponding rank for Rasch PASS scores. This method, in effect, transformed 3PL to a more discrete distribution and removed much of the effect of the change in IRT model. Additional rescaling methods should be explored to compare the 2 models.

A third area of interest is the effect of the change in model for the subgroup of students who are gifted. This study focused on subgroups represented on state an federal

report cards but additional studies could include gifted students, especially because the Rasch versus 3PL scores showed substantial differences at the high end of the score distribution.

Summary

The goal of this study was to examine the impact of the 3PL versus Rasch IRT model in scoring and scaling statewide assessment at the school and district level. The analysis was motivated by the many decisions that are made based on school and district summaries of statewide assessment data. Because decisions from statewide assessments are often made from school and district summaries, analysis of the impact of IRT model at this level contributes to the validity evidence for the assessment.

In general, results of this study indicate that the IRT model, 3PL versus Rasch, does not have a large impact on at school and district summary PASS results. There are some exceptions:

- Small schools or districts where percentage in category was greatly affected by the shift of just a couple of students.
- Schools or districts with low PASS mean scores near 550 which were be below the "Met" cut score of 600 for Rasch or 3PL either way.
- Schools or districts with PASS means over 640 which met the AMO objective for Rasch or 3PL either way.
- The IEP subgroup and ESL subgroups consistently had lower PASS mean scores for 3PL than for Rasch. Because subgroups were weighed as heavily on federal report cards as the entire group of students for a school or district, this was the

area where the IRT model has the greatest impact on school and district summaries.

The IRT model appear does have an impact at the student level where about 3,000-4,000 out of roughly 60,000 students change performance levels for Rasch versus 3PL. This could impact students selected for certain programs such as MTSS or students who are retained by laws such as Read to Succeed (2014).

Person fit statistics and also a simulation study indicate that the 3PL model is a better choice than the Rasch model in the areas where the IRT model has impact: the IEP subgroup and the examinees scoring in the lower range for PASS. The simulation results for this study agree with the findings the simulation study of Jiao and Lau (2003) presented in Chapter 2: if the true model is in question, the 3PL performs better at estimating ability on true 1PL data than the other way around. This indicates that further studies are needed for PASS or PASS-like statewide assessments for students in the IEP subgroup where the Rasch model is employed. Alternatively, the 3PL model should be considered in order to obtain the best estimate of student ability for these examinees.

While this study focused in South Carolina's PASS assessment, the findings inform future PASS-like assessments in South Carolina or other states' educational assessments. For PASS-like assessments where the Rasch model is used, it is informative to note that for the most part, school and district summary scores are not largely impacted by the Rasch versus 3PL model. In this regard, it could be argued that it is time and cost effective to continue the status quo, using the Rasch model for assessments where the Rasch model is already employed. Additionally, state contractors

may be motivated to continue with the Rasch model due to its ease of interpretation in cases where stakeholders will view the raw test scores; the total score is a sufficient statistic for examinee ability, and this property appeals to stakeholders as discussed in Chapter 2. Also, better screening of Rasch items could prevent some of the problems identified with Rasch in this study. This could be investigated using the methods in this study such as checking model fit or simulation studies with the pilot data.

On the other hand, there are many reasons state contractors should strongly consider employing the 3PL model instead of the Rasch model. This study showed that even though PASS items went through pilot testing to ensure Rasch fit, in practice, the 3PL had better person fit, especially at the low ability examinee level. At the individual examinee level, many examinees changed performance level for 3PL versus Rasch. With this in mind, the 3PL model seems more appropriate for making decisions at the examinee level.

The 3PL model has the disadvantage of additional guessing and item discrimination parameter estimation requirements. In the past, this drawback would have been a significant hindrance. However, modern software is readily available to handle this type of estimation. Also, Rasch proponents may argue that with 3PL, the assessment would not be constructed as carefully for sound measurement. But, assessment items could still be screened for proper item functioning through pilot testing before the assessment is administered. State contractors could still choose to discard items that are prone to guessing or show undesirable discrimination properties. Then, after the assessment is administered full-scale, the 3PL will have the advantage of fitting the student response

and perhaps correcting for guessing or varying item discrimination that was not prevented through pilot testing.

Weighing the pros and cons of the 2 models, utilizing the 3PL model appears to be the best choice. At the school or district summary level, the models give similar results. However, at the examinee level, the 3PL model fit better even though the assessment was constructed for the Rasch model. Especially in cases where individual student scores will be used, it makes sense then, to use the model that gives the best estimate at the examinee level. The transition to the 3PL model for Rasch assessments may take time to implement but is clearly worthwhile to ensure the best decisions are made at the examinee level.

Note that this study does not show that the 3PL model is the optimal IRT model for PASS-like assessments. Only the Rasch and 3PL model were compared in this study because they are the two most popular IRT models used in statewide assessments. There are other guessing models that are not as popular that are also worth investigating. San Martin and del Pino (2006) proposed a 1PL with ability-based guessing model, for example. In cases such as adaptive testing where items are targeted to examinee ability levels, perhaps the 2PL model (which excludes the guessing parameter) would work because examinees would not be as likely to guess.

Due to the many decisions made at the student, school and district level, state contractors should continuously investigate IRT models models used for scoring statewide assessment with regard to the practical significance of model misfit. Sinharay et al. noted that "several layers of analysis" are necessary to investigate the practical

significance of model misfit. This study provides one layer but recognizes the need for continued research in this area.

REFERENCES

Abedalaziz, N., & Leng, C. H. (2013). The Relationship between CTT and IRT Approaches in Analyzing Item Characteristics. *Malaysian Online Journal of Educational Sciences, 1*(1), 64-70.

AdvancedED. (2015). AdvancED Performance Accreditation Step by Step. Retrieved from http://www.advanc-ed.org

Alabama State Department of Education. Process used to determine cut scores for the Alabama Science Assessments (Grade 5 and Grade 7). Retrieved from https://www.alsde.edu

Alabama State Department of Education. (2014). Alabama college and career ready assessment system timeline. Retrieved from http://www.alsde.edu/sec/sa/Testing/Student%20Assessment%20Timeline%20Revised%20September%202014.pdf.

American Educational Research Association, A. P. A., National Council on Measurment in Education, (2014). *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Foundation.

American Institutes for Research. (2010). Ohio achievement assessments Reading grades 3-8 Mathematics grades 3-8 Science Grades 5 and 8 May 2010 Administration Early Return Technical Report. Retrieved from http://www.ohiodocs.org/Technical%20Docs/May_2010_OAA_EarlyReturnTechnicalReport_07142010sa_combined.pdf.

American Institutes for Research. (2012). Standard setting technical report Setting performance standards for the Computer-Adaptive Delaware Comprehensive Assessment System (DCAS). Retrieved from http://www.doe.k12.de.us/cms/lib09/DE01922744/Centricity/Domain/111/Vol3_DCAS_StanSetting_TechRep.pdf

American Institutes for Research. (2014). Technical Manual for Minnesota's Title I and Title III Assessments for the academic year 2013–2014. Retrieved from http://www.education.state.mn.us/mde/schsup/testadmin/mntests/techrep/index.html.

American Psychological Association (2016) Retrieved from http://www.apa.org/science/programs/testing/standards.aspx

Arizona Department of Education. (2014). Detailed assessment testing calendar for 2014-2015. Retrieved from http://www.azed.gov/assessment/files/2014/12/detail-calendar-2014-2015-v3.pdf.

Arizona State Department of Education. (2015). Assessment. Retrieved from http://www.azed.gov/assessment/.

Arkansas Department of Education. (2015). Assessment. Retrieved from http://www.arkansased.org/divisions/learning-services/assessment.

Baker, F. B. (2001). *The Basics of Item Response Theory. Second Edition.* College Park, MD: Eric Clearinghouse on Assessment Evaluation.

Bergan, J. R. (2010). Assessing the Relative Fit of Alternative Item Response Theory Models to the Data. Assessment Technology, Incorporated**.** Tuscon, Arizona. Retrieved from http://www.ati-online.com/pdfs/researchK12/AlternativeIRTModels.pdf

Bergan, J. R. (2013). Rasch versus Birnbaum: New arguments in an old debate. Assessment Technology, Incorporated. Tuscon, AZ. Retrieved from http://www.ati-online.com/pdfs/researchK12/RaschVsBirnbaum.pdf

Birnbaum, A. (1968). Some latent trait models. In F. Lord & M. Novic (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing Company, Inc.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues & Practice, 28*(4), 3-14. doi: 10.1111/j.1745-3992.2009.00158.x

California Department of Education. (2015). California Assessment of Student Performance and Progress (CAASPP) System. Retrieved from http://www.cde.ca.gov/ta/tg/ca/.

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C., & Meade, A. (2016). mirt: Multidimensional Item Response Theory. Retrieved September 1, 2016, from https://cran.r-project.org/web/packages/mirt/index.html

Chiu, T.W., & Camilli, G. (2013). Comment on 3PL IRT Adjustment for Guessing. *Applied Psychological Measurement, 37*(1), 76-86.

Colorado Department of Education. (2014). Colorado Measures of Academic Success: Science and Social Studies Assessments Technical Report Spring 2014. Retrieved from http://www.cde.state.co.us/search/node/technical%20reports.

Colorado Department of Education. (2015). 2014 - 2015 School Year State Testing Windows. Retrieved from http://www.cde.state.co.us/assessment/statetestingwindows.

Connecticut State Department of Education. (2015). Student assessment. Retrieved from http://www.sde.ct.gov/sde/cwp/view.asp?a=2748&Q=334726.

Crocker, L., & Algina, J. (2006). *Introduction to Classical & Modern Test Theory*. Mason, Ohio: Thompson Wentwoth.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. W. H. I. Braun (Ed.), *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

CTB McGraw Hill Education. (2008). Accuracy of test scores: Why IRT models matter. (White paper). The McGraw-Hill Companies. Retrieved from https://www.ctb.com/ctb.com/control/researchArticleMainAction?p=ctbResearch&articleId=18746

CTB McGraw Hill Education. (2010). Spring 2009 WESTEST 2 Reading/Language Arts, Mathematics, Science, and Social Studies Assessments. CTB McGraw-Hill Retrieved from http://wvde.state.wv.us/oaa/pdf/WestVirginia_TechnicalReport_2009.pdf.

CTB McGraw Hill Education. (2012). Wisconsin knowledge and concepts examinations Fall 2011 WKCE Technical report. Monterey, CA: CTB-McGraw-Hill Retrieved from http://oea.dpi.wi.gov/sites/default/files/imce/oea/pdf/td-2011-techman.pdf.

CTB McGraw Hill Education. (2013). Oklahoma School Testing Program Oklahoma Core Curriculum Tests Grades 3 to 8 Assessments 2012–2013 Technical Report Monterey, CA: Retrieved from http://www.ok.gov/sde/sites/ok.gov.sde/files/documents/files/OCCT_G3-8_2012-13_TR_FINAL.pdf.

CTB McGraw Hill Education. (2014). Guide to Test Interpretation Grades 3-8 Indiana Spring 2014. Monterey, California: CTB/McGraw-Hill LLC Retrieved from http://www.doe.in.gov/sites/default/files/assessment/2014springinterpretiveguide-1.pdf.

CTB McGraw Hill Education. (2014). Missouri Assessment Program Grade-Level Assessments Technical Report 2014 Final. Monterey, CA: CTB/McGraw-Hill LLC Retrieved from https://dese.mo.gov/sites/default/files/asmt-gl-2014-tech-report.pdf.

CTB McGraw Hill Education. (2015). Fall 2013 Administration Final Technical Report. Monterey, CA: CTB/McGraw-Hill Retrieved from http://www.dpi.state.nd.us/testing/assess/NDSAFall2013TechReport.pdf.

CTB McGraw-Hill Education. (2011). *Colorado Student Assessment Program Technical Report 2011*. : CTB/McGraw-Hill Retrieved from http://www2.cde.state.co.us/artemis/edserials/ed210212internet/ed210212201101internet.pdf.

Data Recognition Corporation. (2012). Mathematics Operational and Field Test Science Operational and Field Test Technical Report October 2012. Retrieved from http://www.education.ne.gov/assessment/pdfs/Final%20NeSA%202012%20Technical%20Report%20for%20NDE.pdf.

Data Recognition Corporation. (2013). Alaska Comprehensive System of Student Assessment Technical Report Spring 2013 Grades 4,8, and 10 Science Standards Based Assessment (SBA). Retrieved from http://education.alaska.gov/tls/assessment/TechReports/Spring13_SBAScience/SpSci13_Tr.pdf.

Data Recognition Corporation. (2013). Idaho Standards Achievement Tests Spring 2013 Technical Report. Retrieved from http://www.sde.idaho.gov/site/assessment/isat/docs/technicalReports/EID244_ISAT_Technical%20Report_Final.pdf.

Data Recognition Corporation. (2013). Iowa Assessments Standard Setting Technical Report. Retrieved from https://www.educateiowa.gov/sites/files/ed/documents/Iowa%20Assessment%20Standard%20Setting%20Technical%20Report%20October%202013.pdf.

Data Recognition Corporation. (2014). Integrated iLEAP 2014 Operational Technical Summary. Retrieved from https://www.louisianabelieves.com/docs/default-source/assessment/ileap-technical-summary.pdf?sfvrsn=4.

Data Recognition Corporation. (2014). Technical Report for the 2014 Pennsylvania System of School Assessment. Retrieved from http://www.portal.state.pa.us/portal/server.pt/community/technical_analysis/7447.

Davier, A., Holland, P. W., & Thaywer, D. T. (2004). The Kernel Method of Test Equating. Princeton, NJ: Springer-Verlag New York, Inc.

Davis, M. D., Assessment Research and Development Assessment and Accountability, Georgia Department of Education, (Personal Communication, February 28, 2015).

Delaware Department of Education. (2014). Delaware System of Student Assessments (DeSSA): 2014–2015 Calendar. Retrieved from http://de.portal.airast.org/wp-content/uploads/2013/07/2014-2015_DeSSA-Calender_Updated.pdf.

Delaware Department of Education. (2015). Assessment. Retrieved from http://dedoe.schoolwires.net/domain/111.

Drasgow, F., Levine, M. &Williams, E. (1985). Appropriateness measurement with polychogomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67-86.

Education Accountability Act, South Carolina Code of Laws. Title 59 Chapter 18 (2008).

Educational Testing Service. (2013). California Department of Education Assessment Development and Administration Division California Standards Tests Technical Report Spring 2012 Administration. Retrieved from http://www.cde.ca.gov/ta/tg/sr/documents/cst12techrpt.pdf.

Educational Testing Service. (2014). Technical Report Proficiency Assessments for Wyoming Students (PAWS) and Student Assessment of Writing Skills (SAWS). Retrieved from http://edu.wyoming.gov/downloads/assessments/2014/PAWS2013-2014TechnicalManualFinal.pdf.

Educational Testing Service. (2015). WASHINGTON COMPREHENSIVE ASSESSMENT PROGRAM GRADES 3–8, High School Spring 2014 Technical Report. Olympia, WA: Retrieved from http://www.k12.wa.us/assessment/pubdocs/WCAP2014SpringAdministrationTechnicalReport.pdf.

Educational Testing Service in accordance with PARCC RFP 06. (2014). PARCC Technical Memorandum for Field Test Phase Final.

Edwards, C. (2015). Iowa State Board of Education Resolution on Statewide Assessment and Recommendations for the Iowa Legislature as Required by Iowa Code 253.7(21). Retrieved from https://www.educateiowa.gov/sites/files/ed/documents/AssessmentResolution_0.pdf.

Engelhard, G. J. (2013). *Invariant Measurement Using Rasch Models in the Social, Behavioral, and Health Sciences*. New York, NY: Psychology Press.

Ewell, P., Boeke, M., Ziz, S., & National Center for Higher Education Management Systems. (2010). State Uses of Accreditation: Results of a Fifty-State Inventory. Washington, DC: Council for Higher Education Accreditation.

Fischer, G. H. (2007). Rasch Models. In Rao, C.R. & Sinharay, S. (Eds.) *Handbook of Statistics* (Vol. 26): Elsevier B.V. doi: 10.1016/S10169-7161(06)26016-4

Florida Department of Education. (2015). Assessments. Retrieved from http://www.fldoe.org/accountability/assessments.

Fulmer, C. , Richland School District One of South Carolina. (Personal Communication, December 4, 2015 and August 31, 2016).

Georgia Department of Education. (2014). Georgia Milestones Assessment System. Retrieved from http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Georgia-Milestones-Assessment-System.aspx.

Haladyna, T. (2004). Developing and Validating Multiple-Choice Test Items. Mahway, NJ: Lawrence Erlbaum Associates, Publishers.

Hambleton, R, Swaminathan, H., & Rogers, H. (1991). *Fundamentals of Item Response Theory,* Sage Publications.

Hancock, G. (1994). Cognitive Complexity and the Comparability of Multiple-Choice and Constructed-Response Test Formats. *Journal of Experimental Education, 62*(2), 143-157.

Hawaii State Department of Education. State Assessment (Smarter Balanced). Retrieved from http://www.hawaiipublicschools.org/TeachingAndLearning/Testing/StateAssessment/Pages/home.aspx.

Hendrawan, I., & Wibowo, A. (2013). The Connecticut Mastery Test. Retrieved from http://www.sde.ct.gov/sde/lib/sde/pdf/student_assessment/research_and_technical/Public_2013_CMT_Tech_Report.pdf.

Human Resources Research Organization Under Subcontract to and in Cooperation with Harcourt Assessment, I. (2007). Technical Report for 2006 FCAT Test Administrations. San Antonio, TX: Retrieved from http://fcat.fldoe.org/pdf/fc06tech.pdf.

Huynh, H. (2006). A Clarification on the Response Probability Criterion RP67 for Standard Settings Based on Bookmark and Item Mapping. *Educational Measurement: Issues and Practice, 25*(2), 19-20.

Idaho Department of Education. (2014). 2014-2015 Testing Dates/Windows. Retrieved from http://www.sde.idaho.gov/site/assessment/docs/Assessment%20Calendar.pdf.

Illinois State Board of Education. (2015). Assessment PARCC - Partnership for Assessment of Readiness for College and Careers Retrieved from http://www.isbe.state.il.us/assessment/parcc.htm.

Illinois State Board of Education Division of Assessment. (2013). Illinois Standards Achievement Test2013 Technical Manual Retrieved from http://www.isbe.net/assessment/pdfs/isat_tech_2013.pdf.

Indiana Department of Education. (2014). 2014-2015 Indiana Assessment Program Manual. Retrieved from http://www.doe.in.gov/assessment.

IRT from SSI: BILOG-MG MULTILOG PARSCALE TESTFACT. (2003). (M. du Toit Ed.). Linconwood, IL: Scientific Software International, INC.

Irwin, P. M., Poggio, A. J., Yang, X., Glasnapp, D. R., Poggio, J. P., Center for Educational Testing and Evaluation, & The University of Kansas. (2006). 2006 Technical manual for the Kansas General Assessments Kansas Assessments of Multiple Measures (KAMM) Kansas Alternate Assessments (KAA). Retrieved from http://cete.ku.edu/sites/cete.drupal.ku.edu/files/docs/Technical_Reports/2007/irwin2007_Math.pdf.

Jacques, A., School District Five of Lexington and Richland Counties of South Carolina. (Personal Communication, August 14, 2015 and August 26, 2016).

Jiao, H., & Lau, A. (2003). The Effects of Model Misfit in Computerized Classification Test. Paper presented at the National Council of Education Measurement, Chicago, IL. Retrieved from http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/ji03-01.pdf

Kane, M. (2009) Validating Proposed Interpretations and Uses of Test Scores. In Lizzitz, R.W. (Ed.) *(2009). The Concept of Validity (41-64)*. Charlotte, NC: Information Age Publishing.

Kang, T., & Cohen, A. S. (2007). IRT Model Selection Methods for Dichotomous Items. *Applied Psychological Measurement, 31*(4), 331-358. doi: 10.1177/0146621606292213

Kansas Department of Education. (2015). Kansas Assessment Program. Retrieved from http://www.ksassessments.org/.

Karantonis, A., & Sireci, S. G. (2006). The Bookmark Standard-Setting Method: A Literature Review. *Educational Measurement: Issues and Practice, 25*(1), 4-12. doi: 10.1111/j.1745-3992.2006.00047.x

Kennedy, P., & Walstad, W. (1997). Combining Multiple-Choice and Constructed-Response Test Scores: An Economist's View. *Applied Measurement in Education, 10*(4), 359-375.

Kentucky Department of Education. (2015). K-PREP. Retrieved from http://education.ky.gov/AA/Assessments/Pages/K-PREP.aspx.

Kim, D.-H. (2006). *A comparison of student performance between paper-and-pencil and computer-based testing in four subject areas.* (3232519 Ph.D.), University of South Carolina, Ann Arbor. Retrieved from https://pallas2.tcl.sc.edu/login?url=http://search.proquest.com/docview/305274598?accountid=13965

Le, D.-T. (2013). *Applying Item Response Theory Modeling in Educational Research.* (Doctor of Philosophy Graduate Theses and Dissertations), Iowa State University, Ames, Iowa. (13410)

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). *Standard setting: A Bookmark approach*. Paper presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoeniz, AZ.

Linn, R. (1990). Has Item Response Theory Increased the Validity of Achievement Test Scores? *Applied Measurement in Education, 3*(2), 115-141.

Lissitz, R., Hou, X., & Slater, S. (2012). The Contribution of Constructed Response Items to Large Scale Assessment: Measuring and Understanding their Impact. *Journal of Applied Testing Technology, 13*(3), 1-52.

Livingston, S. (2004). Equating Test Scores: Educational Testing Service.

Lord, F. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdate, NJ: Lawrence Erlbaum Associates.

Lord, F. (1986). Maximum Likelihood and Bayesian Parameter Estimation in Item Response Theory. Journal of Educational Measurement, 23(2), 157-162.

Lord, F., & Novic, M. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Welsey Publishing Company, Inc.

Louisiana Department of Education. (2015). Assessment Month-by-Month Checklist. Retrieved from http://www.louisianabelieves.com/docs/default-source/assessment/assessment-month-by-month-checklist.pdf?sfvrsn=2.

Maine Department of Education. (2014). MeCAS Testing Dates. Retrieved from http://www.maine.gov/doe/assessment/dates.html.

Mantie, S. (2015). Assessment Memo 2014-2015. Retrieved from http://www.education.nh.gov/instruction/accountability/documents/statewide_assess2014-15.pdf.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of Test Validity*. New York, NY: Routledge.

Maryland Department of Education. (2011). Maryland School Assessment Mathematics Grades 3 through 8 Technical Report: 2011 Administration. Retrieved from http://www.marylandpublicschools.org/msde/divisions/planningresultstest/2011+MSA+Mathematics+Technical+Report.htm.

Maryland State Department of Education. (2014). Maryland State Department of Education Student Testing Calendar School Years 2014-2015 through 2016-2017. Retrieved from http://marylandpublicschools.org/MSDE/testing/docs/MSDE_Testing_Calendar_2014-2015-2016-2017_Rev100714.pdf.

Massachusetts Department of Education. (2015). 2014–2015 Statewide Testing Schedule and Administration Deadlines continued. Retrieved from http://www.doe.mass.edu/mcas/1415schedule.pdf.

Measured Progress. (2010). New England Common Assessment Program Science 2008–2009 Technical Report Dover, NY: Measured Progress Retrieved from http://www.ride.ri.gov/Portals/0/Uploads/Documents/Instruction-and-Assessment-World-Class-Standards/Assessment/NECAP/TechnicalReports/2008-09-NECAP-Science-Technical-Report-with-Appendices.pdf.

Measured Progress. (2011). New England Common Assessment Program Scienc2010-2011 Technical Report. Dover, NH: Retrieved from http://www.education.nh.gov/instruction/assessment/necap/documents/sciencetechreport_2011.pdf.

Measured Progress. (2012). MontCas Criterion-Referenced Test (Montana CRT) 2011–12 Technical Report. Dover, NH: Retrieved from http://opi.mt.gov/PDF/Assessment/CRT/TechRprts/CRT/11-12CRTTechRpt.pdf.

Measured Progress. (2013). 2013 MCAS and MCAS-Alt Technical Report. Dover, NH: Retrieved from http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2013/2013%20MCAS%20and%20MCAS-Alt%20Technical%20Report.pdf.

Measured Progress. (2014). Maine Educational Assessment Grades 5 and 8 Science MeCAS Part I 2013–14 Technical Report. Dover, NH: Measured Progress Retrieved from http://www.maine.gov/doe/mea/documents/2013-14%20MeCAS%20Tech%20Rep%20Pt%20I.pdf.

Measured Progress. (2014). New England Common Assessment Program 2013–14 Technical Report. Dover, NH: Measured Progress Retrieved from http://www.ride.ri.gov/Portals/0/Uploads/Documents/Instruction-and-Assessment-World-Class-Standards/Assessment/NECAP/TechnicalReports/2013-14-NECAP-Math-Reading-Writing-Technical-Report.pdf.

Measured Progress. (2014). New Mexico Standards Based Assessment 2013–14 Technical Report. Dover, NH: Retrieved from http://ped.state.nm.us/AssessmentAccountability/AssessmentEvaluation/SBA/2012/2013-14%20NM%20Technical%20Report.pdf.

Messick, S. (1980). Test Validity and the Ethics of Assessment. *American Psychologist, 35*(11), 1012-1027. doi: 10.1037/0003-066X.35.11.1012

Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance Asessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8. doi: 10.1111/j.1745-3992.1995.tb00881.x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi: 10.1037/0003-066X.50.9.741

Messick, S. (1996). Technical issues in large-scale performance assessment. Retrieved from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=96802.

Michigan Department of Education. (2015). Spring 2015 Michigan Statewide Assessment Transition. Retrieved from http://www.mi.gov/documents/mde/Assessment_Transition_FINAL_11-13-14_473989_7.pdf.

Michigan Department of Education Bureau of Assessment Accountability And Measurement Incorporated. (2011). Michigan Educational Assessment Program Technical Report 2010-2011. Retrieved from http://www.michigan.gov/documents/mde/MEAP_20102011_Technical_Report_393868_7.doc.

Minnesota Department of Education. (2015). Minnesota Statewide Testing Program. Retrieved from http://education.state.mn.us/mdeprod/groups/communications/documents/basic/057901.pdf.

Mississippi Department of Education. (2014). Spring 2015 Test Administration Update. Retrieved from http://parcconline.org/update-session-times.

Missouri Department of Education. (2015). Assessment. Retrieved from http://dese.mo.gov/college-career-readiness/assessment.

Montana Department of Education. (2015). MontCAS Montana Comprehensive Assessment System. Retrieved from http://opi.mt.gov/curriculum/MontCAS/.

Nebraska Department of Education. (2014). Update: Standards, Assessment, and Accountability (SAA) Beginning the School Year 2014-2015. Retrieved from http://www.education.ne.gov/assessment/pdfs/SAA_14_2014%20Final.pdf.

Nevada Department of Education. (2012). Smarter Balanced Assessments for Grades 3-8. Retrieved from http://www.doe.nv.gov/Assessments/Smarter_Balanced_Assessment_Consortium_(SBAC)/.

New Jersey Department of Education. (2014). New Jersey Assessment of Skills and Knowledge 2013 TECHNICAL REPORT Grades 3-8. Retrieved from http://www.state.nj.us/education/assessment/es/njask_tech_report13.pdf.

New Mexico Public Education Department. (2014). New Mexico Statewide Assessment Program 2014–2015. Retrieved from http://ped.state.nm.us/ped/AssessmentEvalDocs/2014-2015%20NMSAP%20Assessment%20Calendar%2001252015.pdf.

New York State Department of Education.(n.d.) http://www.p12.nysed.gov/assessment/timeline-historyrev.pdf.

North Carolina Department of Public Instruction. (2008). The North Carolina Mathematics Tests Edition 3 Technical Report. Retrieved from http://www.ncpublicschools.org/docs/accountability/reports/Mathtechmanualdrafted2.pdf.

North Carolina Department of Public Instruction. (2009). North Carolina Reading Comprehension Tests Technical Report Retrieved from http://www.ncpublicschools.org/docs/accountability/testing/reports/eogreadingtechman3.pdf.

North Carolina Department of Public Instruction. (2009). The North Carolina Science Tests Technical Report. Retrieved from http://www.ncpublicschools.org/docs/accountability/testing/reports/eogscitechmanual.pdf.

North Carolina Department of Public Instruction. (2014). Operational Testing Calendar. Retrieved from http://www.ncpublicschools.org/docs/accountability/testing/calendars/1415optestcal.pdf.

North Dakota Department of Public Instruction. (2014). 2014-15 Assessment Test Windows. Retrieved from http://www.dpi.state.nd.us/testing/assess/2014-15-test-windows.pdf.

O'Gorman, K. Berkeley County School District of South Carolina. (Personal Communication, October 17, 2014 and August 27, 2016).

Ohio Department of Education. (2015). Testing. Retrieved from http://education.ohio.gov/Topics/Testing.

Oklahoma State Deparment of Education. (2014). Oklahoma School Testing Program (OSTP) Frequently Asked Questions. Retrieved from http://www.measuredprogress.org/static/OK/materials/Oklahoma%20School%20Testing%20Program%20%28OSTP%29%20Frequently%20Asked%20Questions%20%28FAQ%29%20Version%2012.9.14.pdf.

Oregon Department of Education. (2010). Oregon Department of Education 2009–2010 Technical Report Oregon's Statewide Assessment System Annual Report Retrieved from http://www.ode.state.or.us/wma/teachlearn/testing/manuals/2011/asmttechmanualvol1_annualreport.pdf.

Oregon Department of Education. (2015). Statewide Assessment Reference Pages. Retrieved from http://www.ode.state.or.us/search/results/?id=169.

Orlando, M., & Thissen, D. (2003). Likelihood-based item-fit indeices for dichotomous item response theory modles. *Applied Psychological Measurement*, 24.

Pearson. (2012). South Dakota State Test of Educational Progress Dakota STEP Technical Report: 2012 Spring Administration. San Antonio, TX: Retrieved from http://www.doe.sd.gov/oats/documents/DSTEP12TR.pdf.

Pearson. (2013). Arizona's Instrument to Measure Standards 2013 Technical Report. Retrieved from http://www.azed.gov/assessment/files/2014/05/aims_tech_report_2013_final.pdf.

Pearson. (2013). New York State Examination in Grade 4 Elementary-Level Science 2012 Field Test Analysis, Equating Procedure, and Scaling of Operational Test Forms Technical Report. Pearson Retrieved from http://www.p12.nysed.gov/assessment/reports/2012/els-tr12w.pdf.

Pearson. (2013). New York State Testing Program 2013: English Language Arts Mathematics Grades 3–8 Technical Report. Iowa City, Iowa: Pearson Retrieved from http://www.p12.nysed.gov/assessment/reports/2013/ela-Math-tr13.pdf.

Pearson. (2014). Kentucky Performance Rating for Educational Progress Every Child Proficient and Prepared for Success Kentucky Department of Education 2013-2013 Technical Manual Version 1.0. Pearson Retrieved from http://education.ky.gov/AA/KTS/Documents/2013-2014%20K-PREP%20Technical%20Manual%20v1.pdf.

Pearson. (2014). Mississippi Curriculum Test, Second Edition (MCT2) Technical Manual 2013–2014. Retrieved from https://districtaccess.mde.k12.ms.us/studentassessment/Public%20Access/Statewide_Assessment_Programs/Technical%20Manuals/MCT2/MCT2%202013-2014%20Tech%20Manual.pdf.

Pelton, T. (2002). Where are the Limits to the Rasch Advantage. Paper presented to the International Objective Measurement Workshop (IOMW). Retrieved from https://scholar.google.com/scholar?q=related:TBCRu4YHoAcJ:scholar.google.com/&hl=en&as_sdt=0,41

Pennsylvania Department of Education. (2015). Pennsylvania System of School Assessment (PSSA). Retrieved from http://www.portal.state.pa.us/portal/server.pt/community/state_assessment_system/20965/pennsylvania_system_of_school_assessment_(pssa)/1190526.

Potts, Jeffrey. Richland School District Two of South Carolina. (Personal Communication, August 14, 2015).

Public Schools of North Carolina. (2014). Technical Brief October 16, 2014. Retrieved from http://www.ncpublicschools.org/docs/accountability/testing/eoceogtechbrief14.pdf.

Rao, V. (2012). *Impact of psychofmetric decisions on assessment outcomes in an alternate assessment.* (3548904 Ph.D.), University of South Carolina, Ann Arbor. Retrieved from https://pallas2.tcl.sc.edu/login?url=http://search.proquest.com/docview/1282654399?accountid=13965

Rasch, G. (1960). Probablistic Models for Some Intelligence and Attainment Tests. Copenhagen.

Rawls, A. M. W. (2009). *The importance of test validity: An examination of measurement invariance across subgroups on a reading test.* (3366531 Ph.D.), University of South Carolina, Ann Arbor. Retrieved from https://pallas2.tcl.sc.edu/login?url=http://search.proquest.com/docview/304995440?accountid=13965

Rhode Island Department of Education. (2015). Assessment Schedules and Workshops. Retrieved from http://www.ride.ri.gov/InstructionAssessment/Assessment/AssessmentSchedulesWorkshops.aspx#15701-schedules.

Royal, K. D., Gilliland, K. O., & Kernick, E. T. (2014). Using Rasch measurement to score, evaluate, and improve examinations in an anatomy course. *Anatomical Sciences Education, 7*(6), 450-460. doi: 10.1002/ase.1436

San Martin, E., & del Pino, G. (2006). IRT Models for Ability-Based Guessing. Applied Psychological Measurement, 30(3).

Saunders, J. South Carolina Department of Education. (Personal Communication, November 24, 2014).

Schneider, M. (2014, 12/8/2014). PARCC Is Down to DC Plus Ten States, and Louisiana Isn't One of Them. *Huffpost Education*. Retrieved from http://www.huffingtonpost.com/mercedes-schneider/parcc-is-down-to-dc-plus-_b_6286010.html

Sinharay, S., & Haberman, S. J. (2014). How Often Is the Misfit of Item Response Theory Models Practically Significant? *Educational Measurement: Issues & Practice, 33*(1), 23-35. doi: 10.1111/emip.12024

Sinhary, S., Johnson, M., Stern, H.(2006). Posterior Predictive Assessment of Item Response Theory Models. *Applied Psychological Measurment, 30(298). doi* 1177/0146621605285517

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the Reliability of Testlet-Based Tests. *Journal of Educational Measurement, 28*(3), 237-247. doi: 10.1111/j.1745-3984.1991.tb00356.x

South Carolina Department of Education. (2010). Technical Documentation for the 2010 Palmetto Assessment of State Standards of Writing, English Language Arts, Mathematics, Science, and Social Studies. Retrieved from http://ed.sc.gov/agency/ie/Assessment/documents/PASS-2010TechnicalReport.pdf.

South Carolina Department of Education. (2014). ESEA Federal Accountability Brief Technical Documentation.

South Carolina Department of Education. (2015). Assessment. Retrieved from http://ed.sc.gov/agency/ie/Assessment/.

South Carolina Department of Education. (2015). South Carolina Intervention Guidance Document. Retrieved from https://www.ed.sc.gov/scdoe/assets/File/instruction/read-to-succeed/Interventions/Working_Draft_12-17-15_Intervention_Guidance_Document.pdf

South Carolina Department of Education Office of Assessment Division of Accountability. (2012). Technical Documentation for the 2012 Palmetto Assessment of State Standards of Writing, English Language Arts, Mathematics, Science, and Social Studies.

South Carolina Department of Education. (2015). Principal Evaluation (PADEPP). Retrieved from http://ed.sc.gov/educators/school-and-district-administrators/evaluation/principal-evaluation-padepp/.

South Carolina Read to Succeed Act, South Carolina General Assembly § Section 59-155-110 (2014).

South Dakota Department of Education. (2015). SD DOE Testing Dates for the 2014-2015 School Year. Retrieved from http://www.doe.sd.gov/Assessment/.

State of New Jersey Department of Education. (2015). Assessment Schedule. Retrieved from http://www.state.nj.us/education/assessment/.

State of Washington Office of Superintendent of Public Instruction. State Testing Overview. Retrieved from http://www.k12.wa.us/assessment/StateTesting/default.aspx.

Swaminathan, H., Hambleton, R., & Rogers, H. (2007). Assessing the fit or item response theory models. In C.R. Rao & S. SInharay (Eds), *Handbook of statistics:* Psychometrics (26).

Tennessee Department of Education. (2014). 2014-2015 TCAP Schedule. Retrieved from http://www.state.tn.us/education/assessment/testing_dates.shtml.

Texas Department of Education. (2014). 2014-2015 Testing Calendar. Retrieved from http://tea.texas.gov/student.assessment/calendars/.

Texas Education Agency, & Pearson. (2014). Technical Digest for the Academic Year

2012-2013. Retrieved from http://tea.texas.gov/Student_Testing_and_Accountability/Testing/Student_Assessment_Overview/Technical_Digest_2012-2013/.

U.S Department of Education. (2001). *No Child Left Behind Act of 2001*. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/index.html.

Utah State Office of Education. (2013). UCAS Utah's Comprehensive Accountability System SAGE Student Assessment for Growth and Excellence. Retrieved from http://www.schools.utah.gov/assessment/Adaptive-Assessment-System/SAGEUCASBrochureWeb.aspx.

Vermont Agency of Education. NECAP Science 2015 Assessment Information. Retrieved from http://education.vermont.gov/documents/EDU-NECAP_Science_2015_Information.pdf. Vermont Agency of Education. (2015). Smarter Balanced Assessment Portal. Retrieved from http://vt.portal.airast.org/important-dates/.

Virginia Department of Education. 2014-2015 Virginia Standars of Learning (SOL) Assessments Test Administration Dates. Retrieved from http://www.doe.virginia.gov/testing/test_administration/testing_calendars/2014-2015_sol_testing_calendar.pdf.

Virginia Department of Education. Virginia Standards of Learning Assemssments Technical Report 2011-2012 Administration Cycle. Retrieved from http://www.doe.virginia.gov/testing/test_administration/technical_reports/sol_technical_report_2011-12_administration_cycle.pdf.

von Davier, M. (2009). Is There Need for the 3PL Model? Guess What?. *Measurement: Interdisciplinary Research and Perspectives, 7*(2), 110-114.

West Virginia Department of Education. West Virginia Smarter Balanced Timeline. Retrieved from http://wvde.state.wv.us/smarter-balanced/timeline.html.

Wheaton, B., Muth, xe, n, B., Alwin, D. F., & Summers, G. F. (1977). Assessing Reliability and Stability in Panel Models. Sociological Methodology, 8, 84-136. doi: 10.2307/270754

Winsteps, retrieved from http://www.winsteps.com/winman/ebiascorrection.htm.

Wisconsin Department of Public Instruction. Assessment in Wisconsin. Retrieved from http://oea.dpi.wi.gov/assessment.

Wright, B. (1992). IRT in the 1990s: Which Models Work Best: 3PL or Rasch? Ben Wright's opening remarkds in his invited debate with Ron Hambleton. Paper presented at the AERA Annual Meeting Session 11.05.

Wright, B. (1997). A History of Social Science Measurement. *Educational Measurement: Issues & Practice, 16*(4), 33-45.

Wyoming Department of Education. School Year 2014-2015 Test Dates. Retrieved from http://edu.wyoming.gov/download/2014-2015-Test-Dates-Calendar-Updated.pdf.

APPENDIX A

IRT MODELS USED IN STATEWIDE ASSESSMENTS PRIOR TO SPRING 2015

Table A.1

IRT models used in statewide assessments prior to Spring 2015

State	**Statewide Assessment Prior to Spring 2015**	**Abbreviation**	**IRT Model**
Colorado	Colorado Student Assessment Program	CSAP	3PL
Florida	Florida Comprehensive Assessment Test	FCAT	3PL
North Carolina	North Carolina Standardized Test	EOG	3PL
North Dakota	North Dakota State Assessment (	NDSA	3PL
Rhode Island	New England Common Core Assessment Plan (	NECAP	3PL
New York	New York State Testing Program	NYSTP	3PL for ELA, Math; Rasch for Science
Illinois	Illinois Standards Achievement Tests	ISAT	3PL
Indiana	Indiana Statewide Testing for Educational Progress	ISTEP	3PL
Louisiana	Louisiana Educational Assessment Program	iLeap	3PL
Maine	New England Common Core Assessment Plan	NECAP	3PL
Massachusetts	Massachusetts Comp. Assess. System	MCAS	3PL
Minnesota	Minnesota Comprehensive Assessments Series II	MCA II and MCA III	3PL
Mississippi	Mississippi Curriculum Test	MCT2	3PL
Missouri	Missouri Assessment Program	MAP	3PL
New Hampshire	New England Common Core Assessment Plan	NECAP	3PL
New Mexico	New Mexico Standards Based assessment,	NMSBA	3PL

Vermont	New England Common Core Assessment Plan	NECAP	3PL
Wisconsin	Wisconsin Knowledge and Concepts Examinations	WKCE	3PL
Oklahoma	Oklahoma Core Curriculum Tests	OCCT	3-PL
Kansas	Kansas State Assessment	KSA	2PL
Connecticut	Connecticut Mastery Test	CMT	Rasch
Georgia	Georgia Milestones Assessment System	Georgia Milestones	Rasch
Idaho	Idaho State Achievement Tests	ISAT	Rasch
Kentucky	Kentucky Core Contents tests	KPREP	Rasch
Maryland	Maryland School Assessment	MSA	Rasch
Michigan	Michigan Educational Assessment Program	MEAP	Rasch
Montana	Montana Comprehensive Assessment System	MontCAS	Rasch
Nebraska	Nebraska State Accountability Assessments	NeSA	Rasch
New Jersey	New Jersey 's Core Curriculum Content Standards	NJASK	Rasch
Ohio	Ohio Achievement Test	OAT	Rasch
Oregon	Oregon Statewide Assessment System	OAKS	Rasch
Pennsylvania	Pennsylvania System of School Assessment	PSSA	Rasch
South Carolina	South Carolina Statewide Assessment Program	PASS	Rasch
South Dakota	Dakota State Test of Educational Progress	STEP	Rasch
Virginia	Virginia Standards of Learning	SOL	Rasch
Washington	Washington Comprehensive Assessment Program	WCAP (MSP)	Rasch
West Virginia	West Virginia Educational Standards Test	WESTTEST2	Rasch
Alaska	Standards Based Assessment	SBA	Rasch
Delaware	Computer-Adaptive Delaware Comprehensive Assessment System (DCAS)	DCAS	Rasch
Arizona	Arizona's Instrument to Measure Standards	AIMS	Rasch
Texas	Texas Assessment of Knowledge and Skills	STAAR	Rasch
Wyoming	Proficiency Assessments for Wyoming Students	PAWS	Rasch

Tennessee	Tennessee Comprehensive Assessment Program	TCAP	*
Alabama	Alabama Reading and Mathematics Test	ARMT	*
Arkansas	Arkansas' Augmented Benchmark Exam	AABE	*
California	Standardized Testing and Reporting	STAR	*
Hawaii	Hawaii State Assessment	HAS	*
Iowa	Iowa Tests of Educational Development	Iowa Assessments	*
Nevada	Nevada Poficiency Examination Program	NPEP - CRT	*
Utah	Utah's Comprehensive Accountability System, Student Assessment for Growth and Excellence	UCAS, SAGE	*

*IRT model information for these states was not available on the corresponding state department of education website at the time that this information was collected at the beginning of 2015. Many states were in a state of transition as they were moving over to a new state assessment. Websites were under construction.

APPENDIX B

STATEWIDE ASSESSMENTS SPRING 2015 FOR ELA AND MATH

Table B.1

IRT models used in statewide assessments Spring 2015 for ELA and Math

STATE	**2014-2015 ELA and Math**	**IRT**
Arkansas	Partnership for Assessment of Readiness for College and Careers	*
Colorado	Partnership for Assessment of Readiness for College and Careers	*
Illinois	Partnership for Assessment of Readiness for College and Careers	*
Louisiana	Partnership for Assessment of Readiness for College and Careers	*
Maryland	Partnership for Assessment of Readiness for College and Careers	*
Massachusetts	Partnership for Assessment of Readiness for College and Careers	*
Mississippi	Partnership for Assessment of Readiness for College and Careers	*
New Jersey	Partnership for Assessment of Readiness for College and Careers	*
New Mexico	Partnership for Assessment of Readiness for College and Careers	*
New York	Partnership for Assessment of Readiness for College and Careers	*
Ohio	Partnership for Assessment of Readiness for College and Careers	*
Rhode Island	Partnership for Assessment of Readiness for College and Careers	*
California	Smarter Balanced	*
Connecticut	Smarter Balanced	*
Delaware	Smarter Balanced	*
Hawaii	Smarter Balanced	*

Idaho	Smarter Balanced	*
Iowa	Smarter Balanced	*
Maine	Smarter Balanced	*
Michigan	Smarter Balanced	*
Montana	Smarter Balanced	*
Nevada	Smarter Balanced	*
New Hampshire	Smarter Balanced	*
North Dakota	Smarter Balanced	*
Oregon	Smarter Balanced	*
South Dakota	Smarter Balanced	*
Vermont	Smarter Balanced	*
Washington	Smarter Balanced	*
West Virginia	Smarter Balanced	*
Wisconsin	Smarter Balanced	*
Alabama	ACT ASPIRE	*
Alaska	Alaska Measure of Progress	*
Arizona	AzMERIT -	*
Florida	Florida Standards Assessment	*
Kansas	Kansas Assessment Program	*
Oklahoma	Oklahoma Core Curriculum Tests	*
South Carolina	ACT ASPIRE	*
Tennessee	Tennessee Comprehensive Assessment Program	*
Utah	Utah's Comprehensive Accountability System, Student Assessment for Growth and Excellence	*
North Carolina	End of Grade Tests	3 PL
Indiana	Indiana Statewide Testing for Educational Progress	3PL
Minnesota	Minnesota Comprehensive Assessments Series III	3PL
Missouri	Missouri Assessment Program	3PL
Georgia	Georgia Milestones Assessment System	Rasch
Kentucky	Kentucky Performance Rating for Educational Progress	Rasch
Nebraska	Nebraska State Accountability Assessments	Rasch
Pennsylvania	Pennsylvania System of School Assessment	Rasch
Virginia	Virginia Standards of Learning	Rasch
Wyoming	Proficiency Assessments for Wyoming Students	Rasch
Texas	State of Texas Assessments of Academic Readiness	Rasch

*Technical reports describing IRT methods were not available for these assessments which were being administered in Spring 2015. Many states transitioned to a new statewide assessment for ELA and Math in Spring 2015 due to the implementation of the Common Core.

APPENDIX C

STATEWIDE ASSESSMENTS SPRING 2015 FOR SCIENCE

Table C.1

IRT models used in statewide assessments Spring 2015 for Science

STATE	Spring 2015 Science Assessment	Abbreviation	IRT
Colorado	Colorado Measure of Academic Progress	CMAS	3PL
Florida	Florida Comprehensive Assessment Test	FCAT 2.0	3PL
Indiana	Indiana Statewide Testing for Educational Progress	ISTEP	3PL
Louisiana	Louisiana Educational Assessment Program	iLeap	3PL
Maine	Maine Educational Assessment	MEA	3PL
Massachusetts	Massachusetts Comp. Assess. System	MCAS	3PL
Minnesota	Minnesota Comprehensive Assessments-III	MCA III	3PL
Missouri	Missouri Assessment Program	MAP	3PL
New Hampshire	New England Common Core Assessment Plan	NECAP	3PL
New Mexico	New Mexico Standards Based assessment,	NMSBA	3PL
North Carolina	North Carolina End of Grade Tests	EOG	3PL
North Dakota	North Dakota State Assessment	NDSA	3PL
Rhode Island	New England Common Core Assessment Plan	NECAP	3PL
Vermont	New England Common Core Assessment Plan	NECAP	3PL
Wisconsin	Wisconsin Knowledge and Concepts Examinations	WKCE	3PL
Alaska	Alaska Science Assessment		Rasch
Arizona	Arizona's Instrument to Measure Standards	AIMS	Rasch
California	California Standards Test	STAR	Rasch
Connecticut	Connecticut Mastery Test	CMT,	Rasch
Georgia	Georgia Milestones Assessment System	Rasch	Rasch
Idaho	Idaho State Achievement Tests	ISAT	Rasch
Kentucky	K-PREP	K-PREP	Rasch

Maryland	Maryland School Assessment	MSA	Rasch
Montana	Criterion Referenced Test	CRT	Rasch
Nebraska	Nebraska State Accountability Assessments	NeSA	Rasch
New Jersey	New Jersey Assessment of Skills and Knowdege	NJASK	Rasch
Oregon	Oregon Assessment of Knowledge and Skills	OAKS	Rasch
Pennsylvania	Pennsylvania System of School Assessment	PSSA	Rasch
South Carolina	Palmetto Assessment of Standards and Skills	PASS	Rasch
South Dakota	Dakota State Test of Educational Progress	STEP	Rasch
Texas	State of Texas Assessments of Academic Readiness	STAAR	Rasch
Virginia	Virginia Standards of Learning	SOL	Rasch
Washington	Measurement of Student Progress	MSP	Rasch
Wyoming	Proficiency Assessments for Wyoming Students	PAWS	Rasch
Alabama	Alabama Science Assessment	ARMT	*
Arkansas	Augmented Benchmark Examinations for Science	AABE	*
Delaware	Delaware Comprehensive Assessment System	DCAS	*
Hawaii	Hawaii State Assessment	HAS	*
West Virginia	General Summative Assessment		*
Kansas	Kansas Assessment Prgram (KAP)	KAP	*
Michigan	Michigan Developed Assessment	MEAP	*
Nevada	Nevada Poficiency Examination Program	NPEP - CRT	*
Ohio	New State Test - American Research Institution		*
Oklahoma	Oklahoma Core Curriculum Tests	OCCT	*
Tennessee	Tennessee Comprehensive Assessment Program	TCAP	*
Utah	Utah's Comprehensive Accountability System, Student Assessment for Growth and Excellence	UCAS, SAGE	*
Illinois	*		*
Iowa	*		*
Mississippi	*		*
New York	*		*

*IRT model information for these states was not available on the corresponding state department of education website at the time that this information was collected at the beginning of 2015. Many states were in a state of transition as they were moving over to a new state assessment. Websites were under construction.

APPENDIX D

VARIABLE DESCRIPTIONS PROVIDED BY THE SCDE

Table D.1

Variable descriptions provided by the South Carolina State Department of Education

Length	Variable Type	Variable Name	Field Name	Field Description	Notes
5	$	xDistID	De-identified District ID	5-digit number	The same disguising algorithm was used for all years/subjects (i.e., a districtl is represented by the same number in all cases). The value 39000 was used for the composite of all schools with too few records to report separately.
8	$	xSchool ID	De-identified School ID	8-digit number Columns 1–5 = district code (xDistID) Columns 5–7 = school code	The same disguising algorithm was used for all years/subjects (i.e., a school is represented by the same number in all cases). The value 39000001 was used for the composite of all schools with too few records to report separately.
				Students must respond to at least one operational item in a given subject area to be considered as having attempted that test. **ELA, Math:** Y = attempted the test N = No	Fields are populated based on all operational MC items. If subject attempt = yes, then a scale score will be assigned.
1	$	ELAAttempt	Subject Attempted - ELA		
1	$	MathAttempt	Subject Attempted - Math		

2	$	Grade	EFA Grade Level	03, 05, 08	
11	$	xStudID	De-identified StateIDState ID	Statewide student ID number (11 digits)	xStudID should be unique within the state. The same disguising algorithm was used for all years/subjects (i.e., a student is represented by the same number in all cases).
1	$	Gender	Gender	M = Male F = Female	
1	$	Ethnic	**OBSERVED** Federal Reporting Category	This field is the federal reporting category based on values in the Hispanic, RaceI, RaceA, RaceB, RaceP, and RaceW fields. Blank, H, I, A, B, P, W, M	
1	$	English	ESL/English proficiency	1 = Pre-functional 2 = Beginner 3 = Intermediate 4 = Advanced 5 = Initially English Proficient 6 = Title III First Year Exited 7 = Title III Second + Year Exited 8 = English Speaker I 9 = English Speaker II–Native English speaker A = Pre-functional—Waiver B = Beginner—Waiver C = Intermediate—Waiver D = Advanced—Waiver	
1	$	Meals	Meals	blank, P = Paid or not eligible for free/reduced meals F = Free R = Reduced	
1	$	IEP	IEP flag	Y = has IEP with at least one IEP category precoded or marked on the document N = no IEP categories were indicated	IEP flag is based on EFA disability codes (IEP_AU through IEP_TBI) AND DeafBl and MultDis
2	$	ELAGrade	ELA Grade Tested	2 digit = numeric (03 , 05, 08).	
4		ELASS	ELA Scale Score		
1		ELALev	ELA Performance Level	1 = not met 2 = met 3 = exemplary Blank = not tested or did not attempt	
2	$	MathGrade	Grade Tested – Math	2 digit = numeric (03, 05, 08).	
4		MathSS	Math Scale Score		
1		MathLev	Math Performance Level	1 = not met 2 = met 3 = exemplary Blank = not tested or did not attempt	

50	$	ELAXi1 - ELAXi50	ELA Scored Item Responses (50 max.)	1 = correct, 0 = incorrect, blank = No response	
63	$	MathXi 1- MathXi 63	Math Scored Item Responses (63 max.)	1 = correct, 0 = incorrect, blank = No response	
1	$	ELAAcc1	ELA Setting	Y = marked on the answer document	**ELA IEP/504 ACCOMMODATIONS** Refer to the page towards the bottom of this document for a listing of nonstandard accommodations.
1	$	ELAAcc2	ELA Timing		
1	$	ELAAcc3	ELA Scheduling		
1	$	ELAAcc4	ELA Presentation – Oral Administration Script		
1	$	ELAAcc5	ELA Presentation – Oral Administration CD-ROM		
1	$	ELAAcc6	ELA Presentation – Signed Administration Script		
1	$	ELAAcc7	ELA Presentation – Signed Administration DVD		
1	$	ELAAcc8	ELA Presentation – Other		
1	$	ELAAcc9	ELA Response Options		
1	$	ELAAcc10	ELA Supplemental Materials or Devices		
1	$	ELAAcc11	Filler		
1	$	ELAESLAcc1	ELA ESL Bilingual Dictionary	Y = marked on the answer document	**ELA ESL ACCOMMODATIONS**
1	$	ELAESLAcc2	ELA ESL Directions Translated		
1	$	ELAESLAcc3	ELA ESL Individual and Small Group Administration		
1	$	ELAESLAcc4	ELA ESL Scheduling		
1	$	ELAESLAcc5	ELA ESL Timing		

1	$	ELASpecReq	ELA IEP Special Request Code	SCDE-approved special request code for an accommodation. Values are 1, 2, or B (both marked). 1 = standard accommodation 2 = non-standard accommodation	
1	$	MathAcc1	Math Setting	Y = marked on the answer document	**MATH IEP/504 ACCOMMODATIONS** Refer to the page towards the bottom of this document for a listing of nonstandard accommodations.
1	$	MathAcc2	Math Timing		
1	$	MathAcc3	Math Scheduling		
1	$	MathAcc4	Math Presentation – Oral Administration Script		
1	$	MathAcc5	Math Presentation – Oral Administration CD-ROM		
1	$	MathAcc6	Math Presentation – Signed Administration Script		
1	$	MathAcc7	Math Presentation – Signed Administration DVD		
1	$	MathAcc8	Math Presentation – Other		
1	$	MathAcc9	Math Response Options		
1	$	MathAcc10	Math Supplemental Materials or Devices		
1	$	MathAcc11	Calculator		
1	$	MathAcc12	Filler		
1	$	MathESLAcc1	Math ESL Bilingual Dictionary	Y = marked on the answer document	**MATH ESL ACCOMMODATIONS**
1	$	MathESLAcc2	Math ESL Directions Translated		
1	$	MathESLAcc3	Math ESL Individual and Small Group Administration		
1	$	MathESLAcc4	Math ESL Oral Administration		

1	$	MathESLAcc5	Math ESL Scheduling		
1	$	MathESLAcc6	Math ESL Timing		
1	$	MathSpecReq	Math Special Request Code	SCDE-approved special request code for an accommodation. Values are 1, 2, or B (both marked). 1 = standard accommodation 2 = non-standard accommodation	

Table D.2

PASS Accommodations

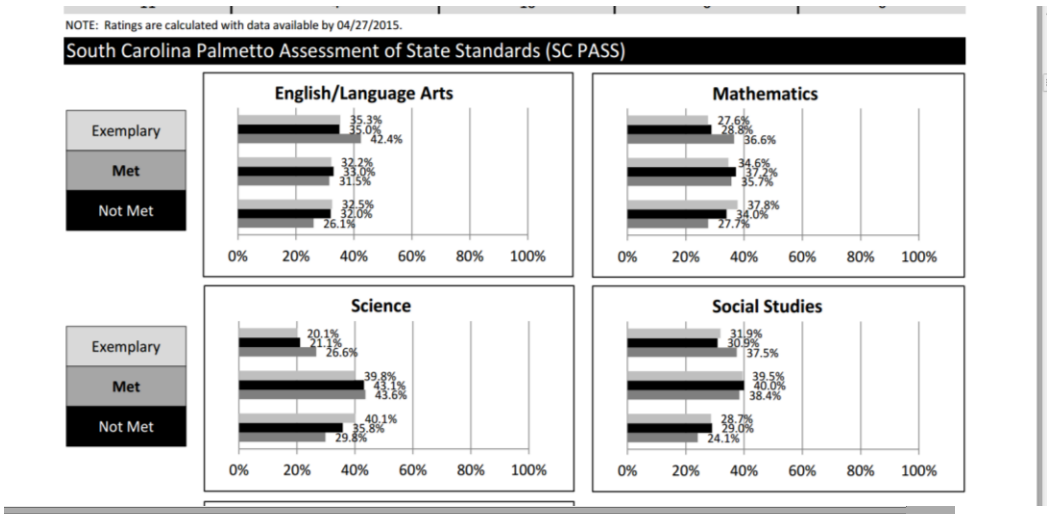
IEP/504 Accommodation	ELA						Math					
	3	4	5	6	7	8	3	4	5	6	7	8
Setting	•	•	•	•	•	•	•	•	•	•	•	•
Timing	•	•	•	•	•	•	•	•	•	•	•	•
Scheduling	•	•	•	•	•	•	•	•	•	•	•	•
Presentation: Oral Administration Script	NS	NS	•	•	•	•	•	•	•	•	•	•
Presentation: Oral Administration CD-ROM	NA	NA	•	•	•	•	NA	NA	•	•	•	•
Presentation: Signed Administration Script	NS	NS	•	•	•	•	•	•	•	•	•	•
Presentation: Signed Administration DVD	NS	NS	•	•	•	•	•	•	•	•	•	•
Presentation: Other	•	•	•	•	•	•	•	•	•	•	•	•
Response Options: Typed/Separate Paper	•	•	•	•	•	•	•	•	•	•	•	•
Response Options: Other	•	•	•	•	•	•	•	•	•	•	•	•
Spelling	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Supplemental Materials or Devices	•	•	•	•	•	•	•	•	•	•	•	•
Extended Response Options	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Alternative Scoring Rubric*	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Calculator	NA	NA	NA	NA	NA	NA	NS	NS	•	•	•	•

Note. • = available, NA = not available or not applicable, NS = available but a non-standard accommodation. Per the PASS Test Administration Manual, the following are considered non-standard accommodations:

1. Oral/Signed Administration for ELA grades 3 and 4
2. Calculator for Math grades 3 and 4

APPENDIX E

PARTIAL 2014 SOUTH CAROLINA DISTRICT REPORT CARD



RICHLAND 1 School District 4/27/2015 4001

Performance By Group - ESEA/Federal Accountability

Subgroups	ELA Mean	Math Mean	Science Mean	Social Studies*/ History Mean	ELA % Tested	Math % Tested	Science % Tested	Graduation Rate
Grades 3-5								
All Students	635.8	627.0	611.2	635.9	99.8	99.9	99.8	N/A
Male	630.1	625.3	610.9	635.0	99.7	99.9	99.8	N/A
Female	641.7	628.8	611.5	637.0	100.0	100.0	99.8	N/A
White	682.9	679.4	660.6	680.4	100.0	99.9	99.7	N/A
African American	624.0	613.1	598.8	624.7	99.8	99.9	99.9	N/A
Asian/Pacific Islander	647.8	653.9	629.2	641.2	100.0	100.0	98.5	N/A
Hispanic	632.0	630.7	607.3	639.2	100.0	100.0	100.0	N/A
American Indian/Alaskan	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
With disabilities	588.7	578.4	573.3	599.2	99.1	99.7	99.2	N/A
Limited English Proficient	622.9	627.4	599.8	625.0	100.0	100.0	99.3	N/A
Subsidized Meals	622.9	613.3	598.1	623.9	99.8	99.9	99.9	N/A
Migrant	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Annual Measurable Objective (AMO)	640.0	640.0	640.0	640.0	95.0	95.0	95.0	N/A
Grades 6-8								
All Students	618.9	620.8	627.9	636.4	100.0	99.9	99.9	N/A
Male	610.3	617.1	625.5	637.0	99.9	99.9	99.8	N/A
Female	627.4	624.4	630.3	635.8	100.0	100.0	99.9	N/A
White	668.1	668.3	683.4	692.3	99.9	99.9	99.9	N/A
African American	605.1	607.0	612.1	620.4	100.0	99.9	99.9	N/A
Asian/Pacific Islander	636.3	650.8	654.6	670.4	98.7	100.0	100.0	N/A
Hispanic	621.0	626.8	630.5	644.4	100.0	100.0	100.0	N/A
American Indian/Alaskan	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
With disabilities	573.5	580.1	581.4	596.7	100.0	100.0	99.5	N/A
Limited English Proficient	603.1	619.3	617.0	630.9	99.4	100.0	100.0	N/A
Subsidized Meals	603.7	606.4	610.3	620.1	100.0	99.9	99.8	N/A
Migrant	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Annual Measurable Objective (AMO)	632.0	632.0	632.0	632.0	95.0	95.0	95.0	N/A
Grades 9-12								
All Students	224.5	216.3	76.2	71.7	99.6	99.3	100.0	74.1

Figure E.1 Partial South Carolina district report card (SCDE, 2015).

APPENDIX F

PARTIAL 2014 SOUTH CAROLINA SCHOOL REPORT CARD

Rosewood Elementary 4/27/2015 4001054

Percent of Student SC PASS Records Matched for Purpose of Computing Growth Rating

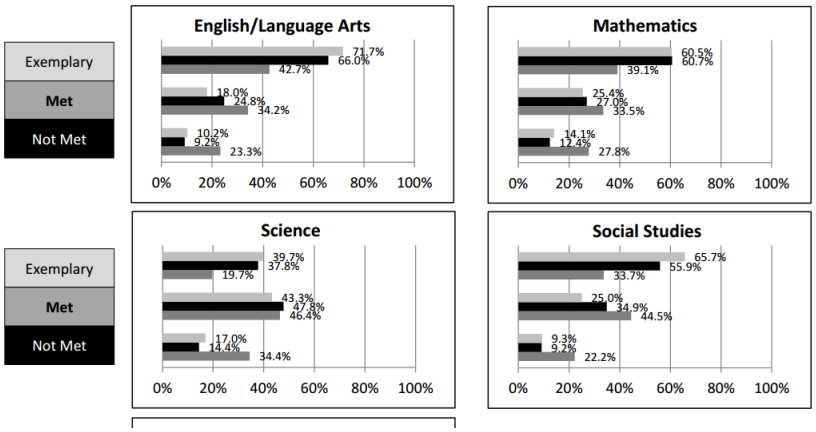
Percent of students tested in 2013-14 whose 2012-13 test scores were located 97.8%

ABSOLUTE RATINGS OF ELEMENTARY SCHOOLS WITH STUDENTS LIKE OURS*

Excellent	Good	Average	Below Average	At-Risk
24	0	1	0	0

* Ratings are calculated with data available by 04/27/2015.

South Carolina Palmetto Assessment of State Standards (SC PASS)



Rosewood Elementary 4/27/2015 4001054

SC PASS Performance By Group - ESEA/Federal Accountability

Subgroups	ELA Mean	Math Mean	Science Mean	Social Studies Mean*	ELA % Tested	Math % Tested	Science % Tested
Grades 3-5							
All Students	678.4	674.7	655.8	683.9	100.0	99.5	99.3
Male	674.6	678.1	655.6	688.4	100.0	99.0	98.7
Female	681.8	671.7	656.0	681.0	100.0	100.0	100.0
White	694.0	691.2	669.6	701.1	100.0	99.3	99.0
African American	617.5	610.8	598.1	N/A	100.0	100.0	100.0
Asian/Pacific Islander	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Hispanic	N/A	N/A	N/A	N/A	N/A	N/A	N/A
American Indian/Alaskan Native	N/A	N/A	N/A	N/A	N/A	N/A	N/A
With Disabilities	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Limited English Proficient	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Subsidized Meals	636.5	627.3	611.3	640.1	100.0	100.0	100.0
Migrant	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Annual Measurable Objective (AMO)	640.0	640.0	640.0	640.0	95.0	95.0	95.0

* Social Studies used as "Other Academic Indicator" for elementary and middle schools.

Figure F.1 Partial South Carolina school report card (SCDE, 2015).

APPENDIX G

PARTIAL ESEA FEDERAL ACCOUNTABILITY SYSTEM COMPOMENT

Sample Elementary School Matrix

Matrix 2 — Elementary School Sample

	ELA Proficiency Met/Improved	Math Proficiency Met/Improved	Science Proficiency Met/Improved	Social Studies Proficiency Met/Improved	ELA Percent Tested (95 % Tested?)	Math Percent Tested (95 % Tested?)
All Students	1	1	1	1	1	1
Male	1	1	1	1	1	1
Female	1	1	1	1	1	1
White	1	1	1	1	1	1
African-American	1	1	0	1	1	1
Asian/Pacific Is	1	1	1	1	1	1
Hispanic	1	1	0	1	1	1
Am Indian/Alaskan						
Disabled	1	0	0	0.2	1	1
Limited Eng. Prof	1	0	0	1	1	1
Subsidized Meals	1	1	0.1	1	1	1
Total # of Points	10	8	5.1	9.2	10	10
Total # of Objectives	10	10	10	10	10	10
Percent of Objectives Met	100	80	51	92	100	100
Weight	0.40	0.40	0.05	0.05	0.05	0.05
Weighted Points Subtotal	35	28	2.55	4.6	10	10

Grade: 90 to 100 = A, 80 to 89.9 = B, 70 to 79.9 = C, 60 to 69.9 = D, < 60 = F

Key: Met=1, Proximity to AMO=.6-.9, Improved= .1-.5, Not Met & Not Improved=0
(Note: Percent Tested may only be Met or Not Met)

Weighted Points Total	90.2
Grade Conversion	A

Step 2— Calculate the means (averages).

- For the "all students" group (no minimum Nsize) and for each subgroup with 30 or more students.

Figure G.1 Partial ESEA Federal accountability system components (SCDE< 2015)

APPENDIX H

EXAMPLE BILOG-MG CODES

Code to obtain Rasch ability esimates using Maximum likelihood estimation and to match SCDE supplied thetas:

```
>GLOBAL DFName = 'Z:\MyDocsAug23\EDRM DIS Research\ela32014.prn',
       NPArm = 1,
       SAVe;
>SAVE PARm = 'Z:\MyDocsAug23\EDRM DIS Research\E32014\parameterml.PAR',
     SCOre = 'Z:\MyDocsAug23\EDRM DIS Research\E32014\scoreml.SCO';
>LENGTH NITems = (36);
>INPUT NTOtal = 36,
      NALt = 5,
      NIDchar = 11;
>ITEMS ;
>TEST1 TNAme = 'TEST0001',
      INUmber = (1(1)36);
(11A1, 0X, 36A1)
>CALIB NQPt = 80,
      CYCles = 40,
      NEWton = 5,
      CRIt = 0.0050,
      ACCel = 1.0000,
      NOSprior,
      RASch;
>SCORE METhod = 1,
      IDIst = 3,
      RSCtype = 3,
      LOCation = (0.8113),
      SCAle = (1.2998);
```

Notes on BILOG Code:

- The rescaling options below were used to obtain a mean of .8113 and standard deviation of 1.2998. These values match the mean and standard deviation of the SCDE supplied thetas
  Location = (0.8113) denotes the desired mean of the theta scale
  Scale = (1.2998); denotes the desired mean of the theta scale
- In order to obtain matching values to Winsteps, the 1PL model, Normal response function metric option must be selected as well as "One Parameter Logistic Model" under calibration options. The Maximum likelihood estimation method was selected to best match Winsteps results.

- 80 quadrature points were selected to match the number of quadrature points selected when using the Expected a posteriori estimation method used with 3PL. However, the option appears to have little or no effect on MLE outcomes.

Score method = 1 indicates the MLE estimation method. MLE finds the θ that maximizes the likelihood function for the examinee (Hambleton, Swaminathan, & Rogers, 1991).

- BILOG uses a Marginal Maximum likelihood (MML) estimation of item parameters. MML is a method of estimating item parameters where the likelihood function is multiplied by a prior on ability, abilities are integrated out of the likelihood function, and the marginal likelihood function is maximized (Hableton, Swaminthan, & Rogers, 1991). No prior options were selected under item analysis.
- (Note that Winsteps, used by the SCDE, uses Joint Maximum likelihood estimation (JMLE). Hableton, Swaminthan, & Rogers, (1991) explain that with JMLE, abilities are estimated and treated as known and them item parameters are estimated; then item parameters are estimated and treated as known and abilities are estimated. These stages are repeated until the estimates do not change (Hableton, Swaminthan, & Rogers, 1991).

Code to obtain 3PL ability esimates using expected a posteriori (EAP) estimation

```
>GLOBAL DFName = 'Z:\MyDocsAug23\EDRM DIS Research\ela32014.prn',
       NPArm = 3,
       SAVe;
>SAVE PARm = 'Z:\MyDocsAug23\EDRM DIS Research\E32014\parameter3ep.PAR',
     SCOre = 'Z:\MyDocsAug23\EDRM DIS Research\E32014\score3ep.SCO';
>LENGTH NITems = (36);
>INPUT NTOtal = 36,
      NIDchar = 11;
>ITEMS ;
>TEST1 TNAme = 'TEST0001',
      INUmber = (1(1)36);
(11A1, 0X, 36A1)
>CALIB NQPt = 80,
      CYCles = 40,
      NEWton = 5,
      CRIt = 0.0050,
      ACCel = 1.0000,
      TPRior,
      GPRior;
>SCORE IDIst = 3,
      RSCtype = 3,
      LOCation = (0.8113),
      SCAle = (1.2998);
```

Notes on BILOG Code:

- The rescaling options below were used to obtain a mean of .8113 and standard deviation of 1.2998. These values match the mean and standard deviation of the SCDE supplied thetas
  LOCation = (0.8113) denotes the desired mean of the theta scale
  SCAle = (1.2998); denotes the desired mean of the theta scale
- Note that MLE ability estimates were found to be unacceptable in the 3PL case. Extreme theta values resulted for low and high abilities and standard errors were unattainable for these extremes. The EAP estimation performed better at the extremes.
- 80 quadrature points were necessary to 'smooth' out 3PL EAP results.
- Prior constraints were selected for item parameters resulting in maximum a posterior (MAP) item estimation method for compatibility with EAP ability estimation method.

APPENDIX I

MODEL FIT CHECKS FOR THE EAP AND MLE ESTIMATION METHODS

In general, checks for model fit in item response theory involve comparing what is observed to model predictions (Swaminathan, Hambleton, &Rogers, 2007). With Rasch, total score is a sufficient statistics for θ and the predicted total score using item and person parameters estimated by the model can be compared to the observed total score. The plots below show that MLE and EAP fit similarly for the Rasch model, with the MLE fitting slightly better low and high scoring examinees.

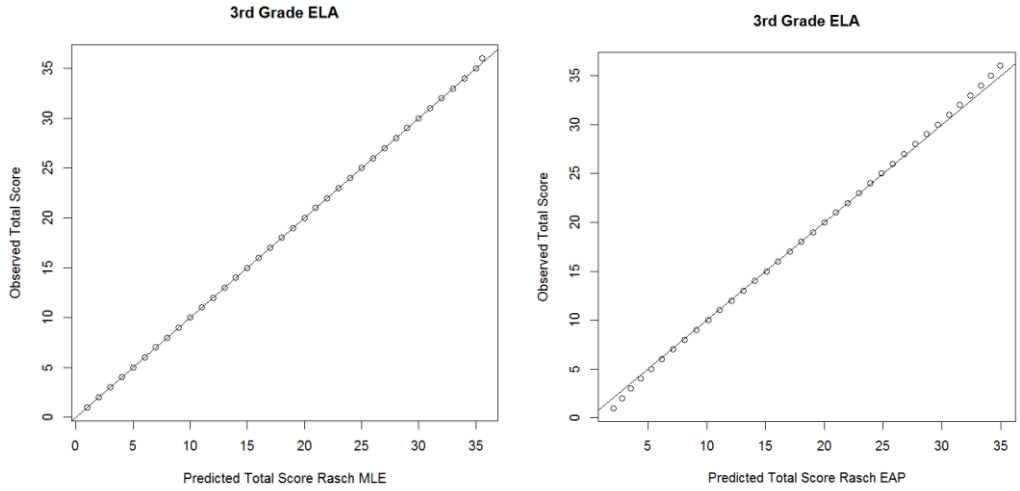


Figure I.1 Plot of predidcted total score versus observed total score EAP and MLE Estimation methods

Additionally, observed values were compared with predicted values at the response level for EAP versus MLE for the Rasch model. For the response matrix, a residual was calculated representing the difference between the scored response (0 or 1) and probability of a correct response based on the model parameters. The differences in the absolute values of the residuals was computed for the two estimation methods. The results are summarized in the following boxplot and indicate the two estimation methods are similar in terms of residuals at the response level. Note that item 13 was removed for this analysis due to an outliers in the residual pattern.

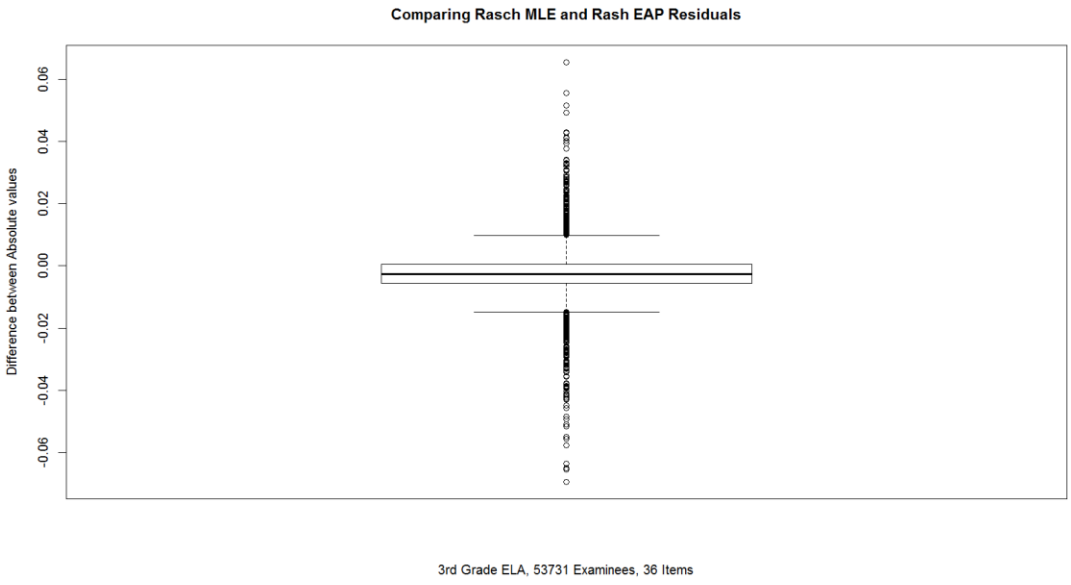


Figure I.2 Boxplot of absolute residual differences for EAP and MLE Estimations

The number of standardized residuals in ranges were also compared for the 2 methods and each method gave similar results:

Table I.1.
Standardized residuals for MLE verusus EAP.

Standardized Residual Range	Rasch MLE %	Rasch EAP %
0-1	75.8	75.8
1-2	19.8	20.1
2-3	3.3	3.1
3-∞	1.1	1

Model checks comparing EAP and MAP for the Rasch model were similar for 3rd grade Math, 8th grade ELA, and 8th grade Math.

Hambleton, Swaminathan, & Rogers (1991)recommend grouping examinees for residual analysis. However with grouping, results were found to be very unstable:

Table I.2.
Standardized residuals for MLE verusus EAP for 10 groups and 15 groups

Number of groups = 10

Standardized Residual Range	Rasch MLE %	Rasch EAP %
0-1	16.1	9.2
1-2	12.8	14.7
2-3	10.3	7.8
3-∞	60.8	68.3

Number of groups = 15

Standardized Residual Range	Rasch MLE %	Rasch EAP %
0-1	18.3	13.3
1-2	14.3	13.5
2-3	10.2	12
3-∞	57.2	61.1

APPENDIX J

GENERAL DATA CHECKS

Table J.1

Counts of zero and perfect scores for the 2014 PASS exam

Grade	Subject	Number of Students	Number of Questions	Number of Zero Scores	Number of Perfect Scores
3rd	ELA	53,731	36	0	422
3rd	Math	53,829	50	0	500
8th	ELA	54,828	50	0	104
8th	Math	54,885	63	0	80

Table J.2

Counts of zero response strings at the end of the exam

Grade	Subject	Number of Students	Count of scored response string '00000' for last 5 items	Percentage of scored response string '00000' for last 5 items
3rd	ELA	53,731	1,791	3.3
3rd	Math	53,829	987	1.8
8th	ELA	54,828	920	1.6
8th	Math	54,885	2,549	4.6

Table J.3

Means and standard deviations of SCDE supplied thetas

Grade	Subject	Theta Mean	Theta s.d.
3rd	ELA	.811	1.30
3rd	Math	.623	1.34
8th	ELA	.234	1.15
8th	Math	.531	1.13

APPENDIX K

ITEM FIT CHECKS

1. A comparison of biserial correlations

The first check involved an analysis to determine if the models were appropriately capturing item discrimination. Classic item analysis was used to calculate biserial correlations for the items in the observed response matrix. (Biserial correlation is a measure of item discrimination in classic item analysis.) Then, parameter estimates from both the Rasch and 3PL were used to simulate response matrices. Biserial correlations from the simulated data were compared with observed biserial correlations. This type of analysis was suggested by Sinhary, Johnson and Stern (2006).

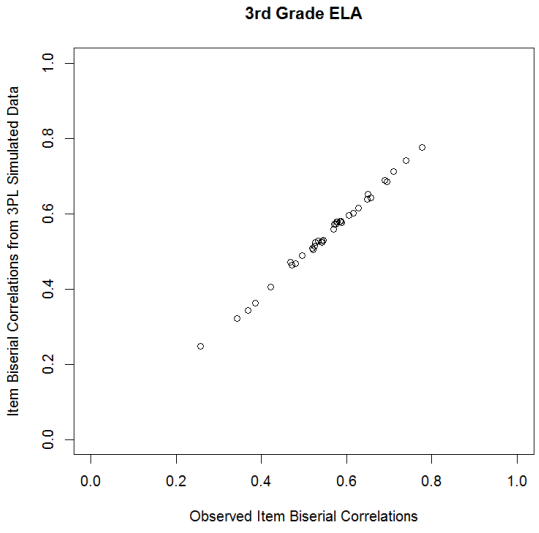


Figure K.1. Plot of observed versus 3PL simulated item biserial correlations

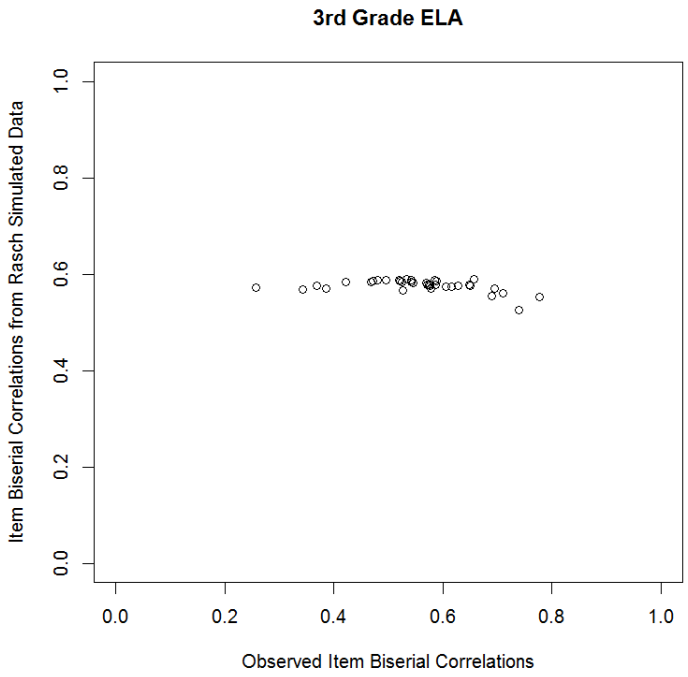


Figure K.2. Plot of observed versus 3PL simulated item biserial correlations

Because the observed item biserials more closely match the results simulated from the 3PL model, this indicates that the 3PL model more accurately described item discrimination. The range of biserial correlations in the simulated Rasch data set was much narrower than the range of observed biserial correlations. This indicates that the Rasch model underestimated item discrimination. Results for other grades and subjects were similar.

A second check also focused on item discrimination. To perform this test, the following steps were followed:

1. Classic item analysis was used to compute the biserial correlation for each of the items on the exam.
2. The standard deviation of the biserial correlations for the items was computed.
3. Model parameters estimated by the Rasch model were used to simulate 10,000 response matrices.
4. The standard deviation of the biserial correlations was obtained for each of the Rasch 10,000 data sets.
5. Model parameters estimated by the 3PL model were used to simulate 10,0000 response matrices.
6. The standard deviation of the biserial correlations was obtained for each of the 3PL 10,000 data sets.
7. The placement of the observed response matrix biserial standard deviation was compared to the Rasch distribution of biserial standard deviations and to the 3PL distribution or 3PL biserial standard deviations. Placement of the observed standard deviation outside of or on the tail end of the simulated distributions is evidence of a poor fit.

Results for 3rd grade ELA:

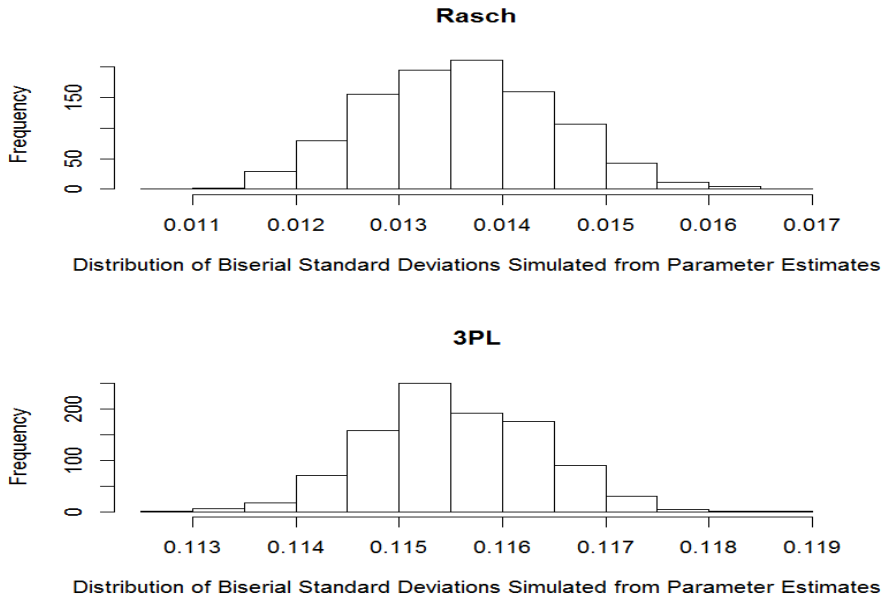


Figure K.3. Histograms of biserial standard deviations simulated from Rasch and 3PL parameter estimates.

The observed biserial correlation for the actual response matrix is **.111** which is at the left tail of the 3PL distribution but is completely outside of the Rasch distribution. This suggests that the 3PL model better describes item discrimination. Results were similar for other grades and subjects.

A third item fit check was based on a Chi-squared goodness of fit index, $S - X_i^2$, which compares the modeled expected proportion of correct responses to an item with the observed proportion of correct responses (Orlando & Thissen, 2000). Given the large sample size, it is expected that standard hypothesis testing would generally indicate misfit, without giving a feel for how severe the misfit was. A traditional measure of fit suggested by Wheaton, Muthen, Alwin, and Summers (1977) is the chi-squared statistic divided by the degrees of freedom. Values greater than 5 indicated significant misfit.

Table K.1

Chi-squared goodness of fit index for Rasch item parameters.

	3rd Grade ELA Rasch			
Item	S-X^2	d.f.	(S-X^2)/d.f.	Significant
1	54	33	1.62	
2	47	33	1.42	
3	559	32	17.46	*
4	2669	32	83.41	*
5	867	33	26.29	*
6	136	33	4.13	
7	1089	33	33.01	*
8	166	33	5.03	*
9	727	33	22.04	*
10	2934	32	91.70	*
11	1519	33	46.02	*
12	119	33	3.61	
13	2089	34	61.44	*
14	98	33	2.96	
15	1042	32	32.56	*
16	68	33	2.07	
17	351	32	10.97	*
18	6656	32	208.00	*
19	498	33	15.10	*
20	267	33	8.10	*
21	2089	33	63.31	*
22	2198	32	68.68	*
23	2503	33	75.86	*
24	1166	33	35.32	*
25	222	33	6.74	*
26	445	33	13.49	*
27	86	33	2.61	
28	625	33	18.94	*
29	222	33	6.72	*
30	800	33	24.23	*
31	229	33	6.93	*
32	308	33	9.33	*
33	593	33	17.98	*
34	450	33	13.63	*
35	98	33	2.96	
36	349	33	10.58	*

Table K.2

Chi-squared goodness of fit index for 3PL item parameters.

3rd Grade ELA
3PL

Item	S-X^2	d.f.	(S-X^2)/d.f.	Significant
1	43	33	1.30	
2	45	33	1.36	
3	74	33	2.23	
4	112	33	3.38	
5	59	33	1.80	
6	39	33	1.19	
7	49	33	1.48	
8	123	33	3.74	
9	49	33	1.49	
10	100	33	3.02	
11	212	33	6.44	*
12	45	33	1.37	
13	299	32	9.36	*
14	41	33	1.25	
15	67	33	2.03	
16	31	33	0.93	
17	82	33	2.47	
18	378	33	11.44	*
19	117	33	3.54	
20	97	33	2.93	
21	325	33	9.83	*
22	410	33	12.43	*
23	156	32	4.87	
24	147	33	4.46	
25	65	33	1.97	
26	131	33	3.97	
27	54	33	1.62	
28	174	33	5.28	*
29	50	33	1.52	
30	63	33	1.92	
31	25	33	0.77	
32	83	33	2.52	
33	119	33	3.59	
34	54	33	1.65	
35	61	33	1.86	
36	39	33	1.17	

Both the 3PL and the Rasch model had several poor fitting items. The Rasch had more misfit items than the 3PL model. Results for 3rd Grade ELA are provided. Similar results were observed for the other data sets.

APPENDIX L

NORMAL QUANTILE PLOTS FOR THETAS

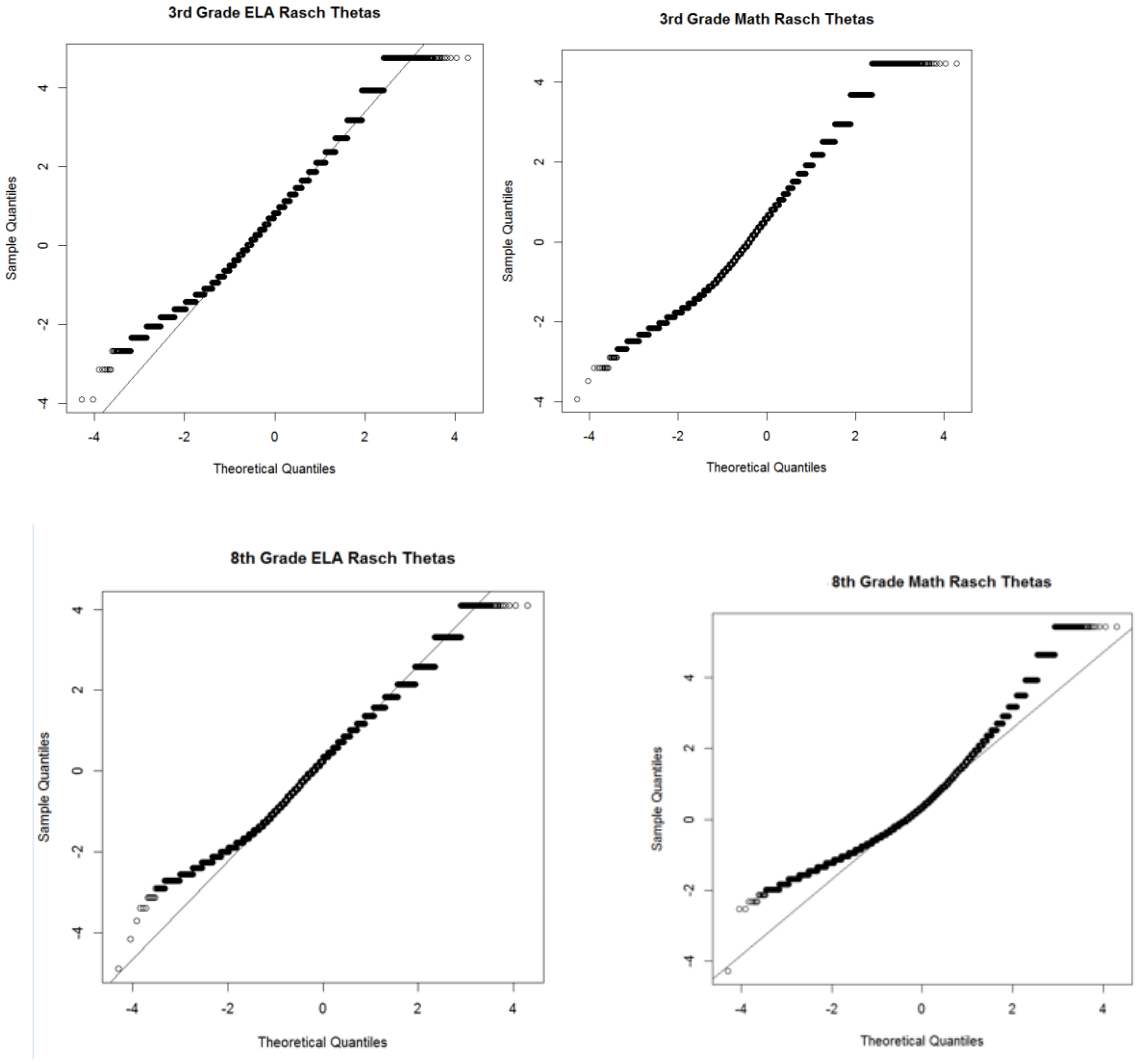


Figure L.1. Quantile plots for Rasch thetas.

APPENDIX M

PERSON FIT QUANTILE PLOTS

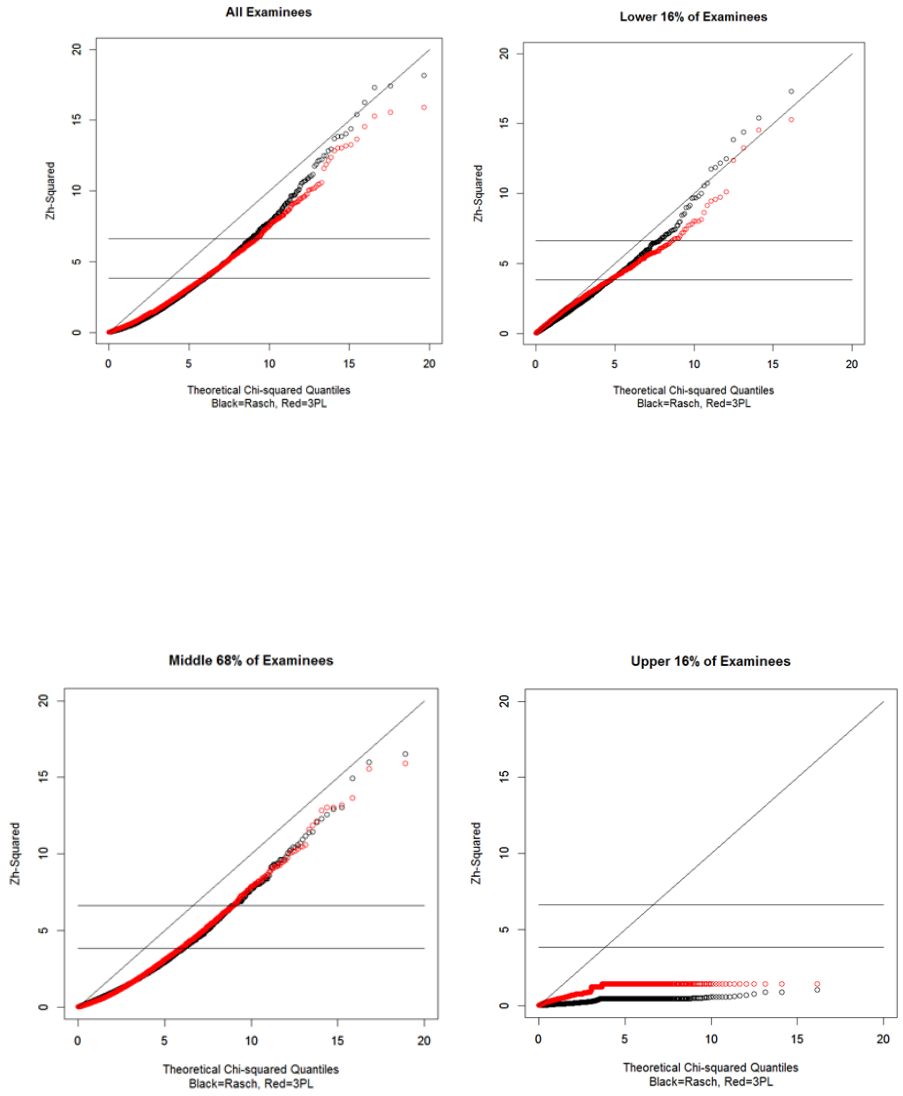


Figure M.1. Person fit quantile plots for 3rd grade Math.

8rd Grade ELA Person Fit Quantile Plots

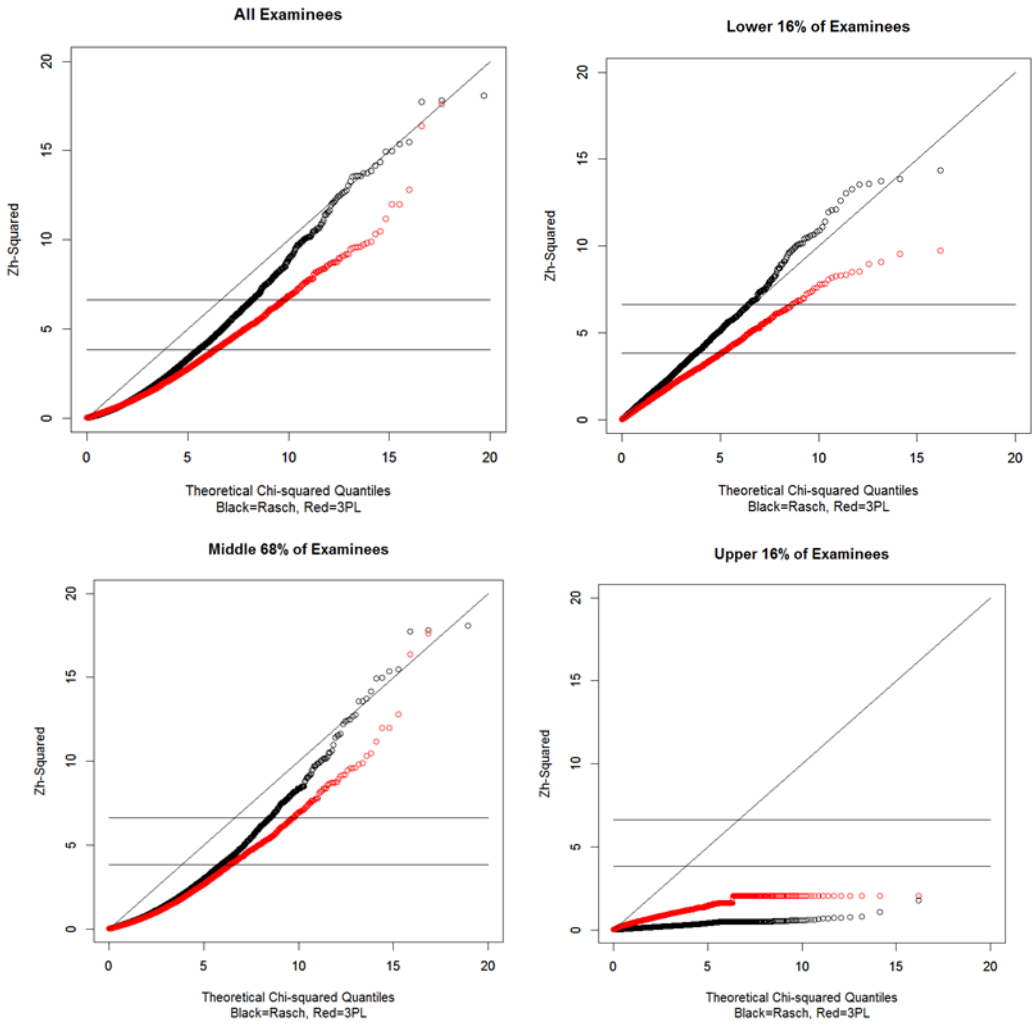


Figure M.2. Person fit quantile plots for 8rd grade ELA.

8rd Grade Math Person Fit Quantile Plots

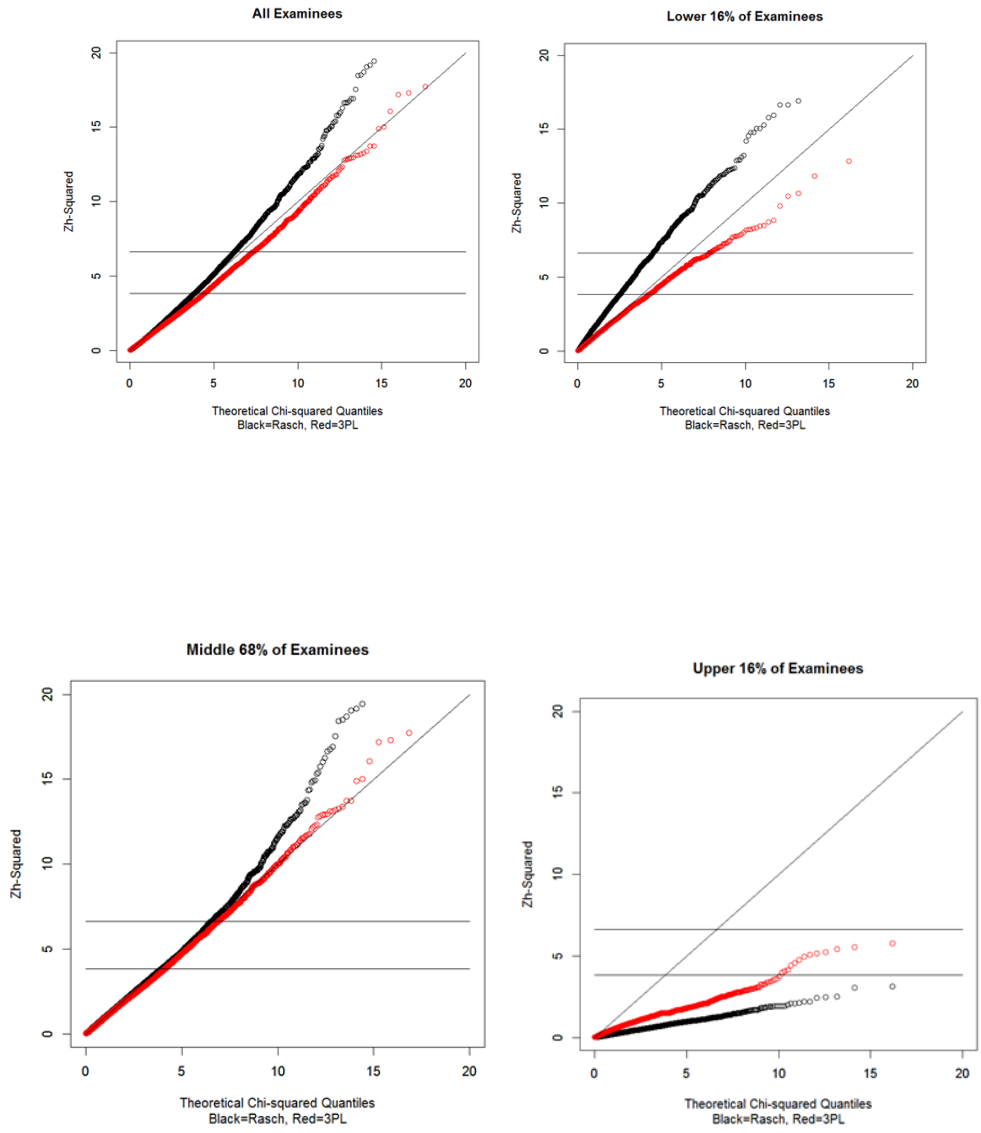


Figure M.3. Person fit quantile plots for 8th grade Math.