Copyright

by

Samuel Heard Haring

2014

The Dissertation Committee for Samuel Heard Haring Certifies that this is the approved version of the following dissertation:

A Comparison of Three Statistical Testing Procedures for Computerized Classification Testing with Multiple Cutscores and Item Selection Methods

Committee:

Barbara G. Dodd, Supervisor

Susan N. Beretvas

Jodi M. Casabianca

Tracey R. Hembry

Tiffany A. Whittaker

A Comparison of Three Statistical Testing Procedures for Computerized Classification Testing with Multiple Cutscores and Item Selection Methods

by

Samuel Heard Haring, BGS

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2014

Dedication

Dedicated to my family.

Acknowledgements

There are a multitude of people who should be recognized for the roles they have filled in my life and I apologize that I cannot mention all of them—know that I try my best to acknowledge them in thought, prayer, and with gratitude as I am fortunate to visit them from time to time.

I first met Dr. Dodd when she and Dr. Beretvas rescued me during an interview at the university. I will never be able to sufficiently thank or repay Barbara for all of her patience and kindness over the past five years. She has been an extraordinary instructor, research colleague, counselor, and friend. Barbara has always being there for me with just the right guidance at the right time—know that you are always a rock star to me.

I would like to thank my committee members for so willingly serving on my committee. I see Tasha sprinting around the halls to make it to all of her meetings, but she has always had great insight and encouragement—you'll never know how much it meant to me a few years ago when you offered me a cup of tea when I was having a really terrible day. For years Tiffany's office door has been open and she somehow always finds time to help me with everything from stats to hip-hop lyrics—thanks for always keeping it real. Jodi agreed to serve on my dissertation committee before she even arrived at the University of Texas. Thank you for taking on so much your first year here. I have to thank Dr. Tracey for not only serving on my committee, but also for being an awesome boss at Pearson—all while carrying her first child. Thank you, Tracey, you've been amazing in so many ways and I can never thank you enough.

I have had a wonderful set of classmates during grad school who have saved me in classes and projects with their insights and discussions—I feel privileged to know and love you. I have also been very fortunate to have been an intern at Pearson for the past two years where I have found a brilliant and wonderfully supportive group of friends.

I have always felt that I am very much the product of my upbringing and therefore feel that I should recognize that I have had a wonderful life. I have had extraordinary teachers, coaches, band directors, friends, mentors, community and church members, and family members who have supported me for decades. Please know that I have learned so much from your words and examples—you have been tremendous. Thank you.

I could not have had a more wonderful family. My siblings have been there for me through so much. You have each, in your own way, taught me to believe in the good in people, self-respect and humility simultaneously, and among other things, about being a superhero. We could not have had a more perfect mother—we love you dearly. There has never been a boy who has loved and admired his mother more than I do you. My grandparents were the best people the world has ever known—I miss you both every day.

I am very proud to say that I finished my dissertation two days before my tenth anniversary. My wife has been amazing through it all. There are still many times I can hardly believe she only turned me down four times before going out on a date with me. I can never repay my wife for her patience, encouragement, love, laughs, and everything else she has given me—thank your. Four years ago my wife gave me our daughter—I have never loved so profoundly. I now have a few people down in the round-tower of my heart, and I will keep them there forever and a day.

A Comparison of Three Statistical Testing Procedures for Computerized Classification Testing with Multiple Cutscores and Item Selection Methods

Samuel Heard Haring, Ph.D.

The University of Texas at Austin, 2014

Supervisor: Barbara G. Dodd

Computerized classification tests (CCT) have been used in high-stakes assessment settings where the express purpose of the testing is to assign a classification decision (e.g. pass/fail). One key feature of sequential probability ratio test-type procedures is that items are selected to maximize information around the cutscore region of the examinee ability distribution as opposed to common features of CATs where items are selected to maximize information at examinees' interim estimates. Previous research has examined the effectiveness of computerized adaptive tests (CAT) utilizing classification testing procedures a single cutscore as well as multiple cutscores (e.g. below basic/proficient/advanced).

Several variations of the SPRT procedure have been advanced recently including a generalized likelihood ratio (GLR). While the GLR procedure has shown evidences of improved average test length while reasonably maintaining classification accuracy, it also introduces unnecessary error. The purpose of this dissertation was to propose and investigate the functionality of a modified GLR procedure which does not incorporate the

unnecessary error inherent in the GLR procedure. Additionally this dissertation explored the use of the multiple cutscores and the use of ability-based item selection.

This dissertation investigated the performance of three classification procedures (SPRT, GLR, and modified GLR), multiple cutscores, and two test lengths. An additional set of conditions were developed in which an ability-based item selection method was used with the modified GLR. A simulation study was performed to gather evidences of the effectiveness and efficiency of a modified GLR procedure by comparing it to the SPRT and GLR procedures.

The study found that the GLR and mGLR procedures were able to yield shorter test lengths as anticipated. Additionally, the mGLR procedure using ability-based item selection produced even shorter test lengths than the cutscore-based mGLR method. Overall, the classification accuracy of the procedures were reasonably close. Examination of conditional classification accuracy in the multiple-cutscore conditions showed unexpectedly low values for each of the procedures. Implications and future research are discussed herein.

Table of Contents

List of Tables

List of Figures

CHAPTER I: INTRODUCTION

The central tenet of testing is that a discrete demonstration of a trait can be quantified and extended to a description of the ability or construct level. In other words, the purpose of testing is to enable accurate generalizations from a small instance to a larger set of phenomena. Testing procedures have been used for centuries to provide evidence that an examinee had achieved a level of ability considered to be related to a pre-specified level competence. For example, some of the earliest records regarding assessments were proficiency measures in the Chan Dynasty (Wainer, Flaugher, Green, Mislevy, Steinberg & Thissen, 1990).

For the past few decades, the focus of many assessments has been to provide a score indicating one's estimated ability level for a given trait. Oftentimes the score achieved on a test is used to then classify the individual into two or more categories. For example, the resulting scale score from a statewide academic achievement assessment may be used to determine if a student receives credit for a course or if they may advance to a subsequent grade level. Due to the nature of the decisions based on the results of the standardized tests, the accuracy of scores and classifications used to make decisions for students and schools as well as the security of the items and tests is of paramount importance.

While much of the public discussion and media interest surrounding high-stakes standardized testing revolves around testing in public schools, high-stakes standardized testing is commonplace in many fields. Standardized testing has become a mainstay in a variety of professional settings. Classification testing methodologies have been successfully implemented in educational settings, employment screenings, and licensure and certification examination programs (Parshall, Spray, Kalohn & Davey, 2000). Medical and clinical settings also utilize testing and classification procedures for determining diagnoses and treatments. Additionally, many higher education entities require placement examinations to best determine the needs of individual students beginning formal education. One of the more notable standardized testing programs using a classification testing scoring procedure for decision making is the COMPASS (ACT, 1993) which has been used in colleges and universities for course placement decisions.

In many situations classifying examinees into one of multiple categorizations (e.g. below average / average / above average) is preferable to simple dichotomous categorizations (e.g. pass / fail). For instance, state testing programs such as the State of Texas Assessments of Academic Readiness (STAAR) use ability scores to categorize students into three categories: Unsatisfactory, Satisfactory, and Advanced. Similarly the National Assessment of Educational Progress (NAEP), used in the Nation's Report Card, also categorizes student performance into four categories: Below Basic, Basic, Proficient, and Advanced. Additionally, the Partnership for Assessment of Readiness for College

and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC) both use more than two cutscores, 3and 4 respectively. Classification testing procedures of several types (e.g. sequential probability ratio test, ability confidence interval, sequential bayes, etc…) have been shown to accurately classify examinees into dichotomous categories as well as into multiple categories (Eggen & Straetmans, 2000; Spray, 1993; Spray & Reckase, 1994).

Increasingly, testing in professional, clinical, and educational settings have been moving to computer-based test administrations as there are several advantages of computer-based administrations over traditional paper and pencil tests. Computer-based administrations allow for greater flexibility in the testing window for examinees while maintaining test security. Computerized administrations also provide the opportunity to make large-scale assessments adaptive to examinee ability based on computerized adaptive test (CAT) algorithms.

The idea underlying computerized adaptive testing is to develop a smart assessment through the use of item selection algorithms. CATs are capable of reducing the number of items needed to obtain examinees' scores as well as improve the accuracy of the ability estimates by tailoring the exam to more closely match the examinee's ability (Reckase, 1989; Wainer, et al., 1990). When assessing an individual's ability level, we gain little information by administering items that are too difficult or too easy for the examinee's ability. Thus by adaptively selecting items based on the examinee

response patterns, such as is done with CATs, the selected items are more informative and useful for obtaining an estimate of the examinee's ability. Unfortunately, the adaptive algorithms common in CATs become less efficient when coupled with certain classification-only scoring methods—consequently the hallmark advantage of CAT becomes nullified.

There is an array of approaches one may take to classify examinees in the context of a CAT. Often an assessment is used to obtain an estimate of the examinee's ability after which the scale score is used to classify them according to a pre-specified performance standard. This, an ability estimate approach, is commonly used in many large state-wide achievement test settings. The estimated ability levels that result from this type of assessment can be used for classification purposes such as admission to a school or educational program, advancement through an education program, and professional licensure or certification. One drawback of this method is that when compared to scoring procedures which are classification-only testing methodologies, the ability estimate approach requires examinees to answer more items.

A technique designed to solely classify examinees without directly estimating examinee ability is a feasible option as well. The prime advantage of the classification-only type methods is that they are highly accurate and very efficient in item usage. The major criticism of such methodologies is that the procedure and outcome do not lend themselves to yield reasonable ability estimates and therefore does not provide insight

into how far above or below an examinee's score is in comparison the performance standard. Even recently published variations of an established classification method still fail to be able to use traditional CAT methodologies and thereby cannot take full advantage of CAT capabilities.

As somewhat of a compromise between ability estimation and pure classification techniques, one may use ability estimates with their accompanying estimated standard errors to determine a classification. As a result both an ability estimate can be provided as well as conserving items by terminating the test upon reaching a classification decision as opposed to obtaining a specific level of precision for the ability estimate. The shortcoming of this technique is that it is not as efficient in conserving items as the classification-only type methodologies.

The focus of this study will be to develop and assess the functionality of a new classification methodology. Previously researched classification procedures, namely the truncated sequential probability ratio test and the generalized likelihood ratio test, will be included in the proposed study to provide baseline measures for comparison to the new classification procedure. This new classification procedure is a modification of the generalized likelihood ratio procedure. These three procedures will be compared to determine the relative merits of each procedure in terms of classification accuracy and test length. Additionally, this study will assess the implementation of an ability-based item selection method for use with the new classification procedure developed in this

dissertation. Therefore, the purpose of the currently proposed study is to (1) extend likelihood ratio test-based methodologies and (2) evaluate the proposed classification methodology in the context of traditional CAT procedures. Additionally, as classification is not limited to dichotomous decisions, the proposed research will incorporate single cutscore, two cutscores, and three cutscores conditions to examine the efficiency of the classification methods.

CHAPTER II: LITERATURE REVIEW

The following literature review provides an in-depth discussion of the three key components forming the foundation for this study. The first section is an overview of item response theory (IRT) presenting three commonly used dichotomous models, namely the one-parameter, two-parameter, and three-parameter logistic IRT models. The second section provides an overview of the essential elements of traditional computerized adaptive testing for dichotomous IRT models including the item pool, testing algorithm, ability estimation procedures, item selection procedures, and test termination methods. The final section presents an introduction to computerized classification testing with dichotomous IRT models including a detailed examination of the truncated sequential probability ratio test, generalized likelihood ratio, and the proposed modified generalized likelihood ratio. Special consideration is also given to the unique item pool requirements and item selection method of computerized adaptive classification testing using IRT models.

ITEM RESPONSE THEORY

IRT is a model-based measurement system that has been used when implementing computerized adaptive testing programs. IRT models have a distinct advantage over classical test theory (CTT) in that IRT analyses are performed at the item level as opposed to the test level as with CTT. As the conventional purpose of CAT is to produce

an estimate of an examinee's ability regarding a construct of interest, the item level analysis of IRT is especially advantageous for CAT as items are selected for administration to a particular examinee based on item statistics to maximize the accuracy of the ability estimate given their previous item responses. The item-level statistics provided by dichotomous IRT models describe the probability of an examinee's response, correct or incorrect, conditional on the examinee's ability level. For any item, the probability of a correct response can be produced for a discrete ability level and may be understood as the probability of a correct response for any randomly selected individual with the specified ability level (DeMars, 2010). The relationship between ability level and the probability of a correct response is a non-linear, monotonically increasing function that is depicted graphically by an item characteristic curve (ICC) (Lord, 1952). The subsequent sections provide a detailed description of three commonly used dichotomous IRT models.

Dichotomous IRT Models

Dichotomous IRT models may be used when assessments utilize only a correct/incorrect scoring system for each item such as multiple-choice items. IRT presents probabilistic measurement models wherein items can be individually characterized by their parameter estimates: difficulty (*b*), discrimination (*a*), and pseudo-guessing (*c*). The item's parameter estimates together with the examinee's responses to a series of items are used to calculate an ability estimate with an accompanying standard

error for the examinee's performance. For the IRT models described in this study the modeling of the item parameters and examinee ability levels require that three core assumptions are met which are: (1) unidimensionality, (2) local independence, and (3) model specified functional form of item probabilities conditional on ability level.

The assumption of unidimensionality implies that examinee responses to test items represent observations of a single latent trait or ability. Therefore all differences between examinee ability estimates is due to actual differences in ability for the single ability dimension being measured (Embretson & Reise, 2000). All other factors that may have influenced the responses to test items are considered random error or nuisance variance distinctive to the individual item and not shared with other items.

The assumption of local independence requires that after conditioning on ability, responses to items are statistically independent of one another. By maintaining the assumption of local independence, we may reasonably conclude that examinee responses to test items represent observations of the unidimensional latent ability of interest. A similar, but weaker form of the local independence assumption assumes that responses to items are uncorrelated after conditioning on ability. In either circumstance, violating the assumption of local independence may cause a misrepresentation of ability estimates as item information is overestimated (Wainer & Lewis, 1990). Violation of the assumption of local independence may suggest that another dimension is influencing the responses to the items and thereby violating the assumption of unidimensionality (DeMars, 2010).

The third assumption concerning correct model specified functional form posits that the empirical data from the examinees' responses to the item follows the expected functional mathematical form specified by the model. A mathematical function can be used to model the expected probability of a given response conditionally across the ability continuum (Hambleton & Swaminathan, 1985). An item's ICC can be used to examine how accurately the model specified functional form corresponds with the observed proportion of correct/incorrect responses based on conditional ability estimates using theoretical versus empirical plots or fit statistics (DeMars, 2010).

A key feature of IRT models is known as parameter invariance which presents two distinct advantages over classical test theory (Lord F. M., 1980). Item parameter invariance means that item parameter estimates are independent of the population used to calibrate the item parameters. Therefore, item parameter estimates should be invariant within a linear transformation to another metric or scale (i.e. scale scores). The stability of item parameters such as difficulty are a distinct advantage over the CTT model where item difficulty is dependent on the examinees who were used to estimate the items. Item parameter invariance is dependent on model-data fit. In situations where model-data fit is poor, parameter invariance cannot be maintained as different populations may produce inconsistent parameter estimates whereas good model-data fit would support parameter invariance. The second advantage of IRT models' parameter invariance provides the capacity for individual ability estimates to be independent of the items used to obtain the

ability estimates for a given unidimensional trait. In comparison the classical test theory model, ability estimates are influenced by test characteristics.

As there are several commonly used dichotomous IRT models, model selection is an important aspect of any application. The IRT models differ based on the number and nature of the item parameters which are estimated. The assumptions required for each model must be considered along with the data being investigated to select a model. Selecting the wrong model will result in poor estimation of the ability levels of examinees. Commonly used dichotomous IRT models are the one-parameter, two-parameter, and three-parameter logistic models.

One-parameter logistic IRT model. The one-parameter logistic model (1-PL) provides estimates of a single item parameter (Rasch, 1960). The item difficulty parameter is represented by *b* and is the only item parameter estimated by the 1-PL model. With the 1-PL, the probability of examinee *j* with a given Θ, where theta represents the examinee's ability level, correctly responding to an item is defined as

$$P(X_i = 1 | \theta_j) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}} \quad (1)$$

where b_i represents the difficulty level, or location of the item along the trait continuum, of item *i*. Theoretically the value of the *b* parameter can range from -∞ to ∞ but it is more common for the range to be closer to -4 to 4.

The 1-PL model is the most restrictive model in that the model only allows for the difficulty, or *b* parameter, value to vary while subsequent models have additional varying item parameters. The 1-PL model includes two additional assumptions pertaining to item discrimination and the probability of an examinee guessing the correct response option. Item discrimination is constant across all items thus assuming that all items discriminate equally at their estimated difficulty locations and exhibiting ICCs that are all parallel. An examinee's ability to correctly guess and obtain credit for the item is not given consideration with this model. Items where guessing may be successfully occurring may be considered for removal from the item pool and therefore the lower asymptote of the ICC for an item is assumed to be zero.

An ICC is used to illustrate the relationship between the examinee's ability, the item difficulty parameter, and the probability of the response to the item. The individual's ability is represented by the Θ scale located on the x-axis. Figure 1 depicts ICCs for three 1-PL items. The item difficulty is also on the same metric and x-axis being represented by the *b* parameter. Having both the ability scale (Θ) and the difficulty scale (*b* parameter) on the same scale facilitates direct comparisons between examinee ability level and item difficulty regarding the probability of a correct response. The probability

of the individual's response to the item ranges from 0.0 to 1.0 and is found along the y-axis.

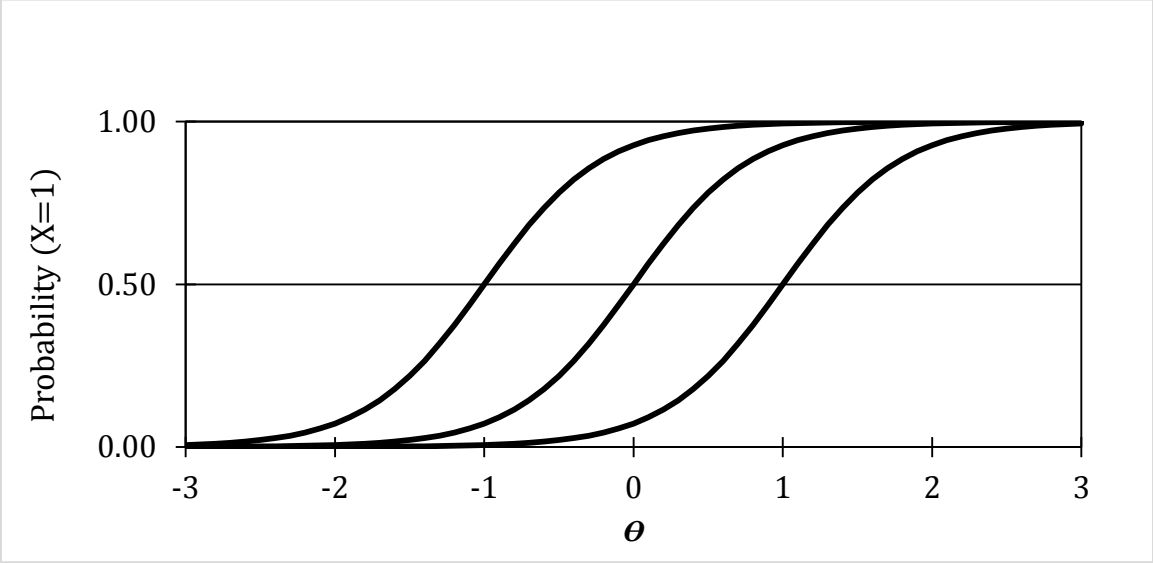


Figure 1: Three item characteristic curves estimated with the 1-PL model.

For the 1-PL model the *b* value of the item is defined by the Θ value on the x-axis that corresponds with the point of inflection or the steepest part of the slope of the ICC. For the 1-PL model the point of inflection will always correspond to a probability of correct response equal to 0.50. All items in a 1-PL model will have the same ICC slopes therefore the curves will never intersect though the curves will eventually converge. The ICCs for the 1-PL model are asymptotic as the probability of the responses at the less proficient end of the ability continuum will become zero to indicate zero probability of guessing.

Two-parameter logistic IRT model. The two-parameter logistic model (2-PL) (Birnbaum, 1968) provides estimates of two item parameters. As with the 1-PL model, the item difficulty is represented by the *b* value. The second parameter estimated by the 2-PL model is the *a* parameter, also referred to as the discrimination parameter or power of the item. With the 2-PL, the probability of examinee *j* with a given Θ, where theta represents the examinee's ability level, correctly responding to an item is defined as

$$P(X_i = 1 | \theta_j) = \frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}} \quad (2)$$

where b_i represents the difficulty level, or location of the item along the trait continuum, of item *i*, a_i represents the discrimination power of item *i*, and D is a scaling constant used to closely align the logistic function with the standard normal ogive. Figure 2 depicts ICCs for 3 items calibrated with the 2-PL model.

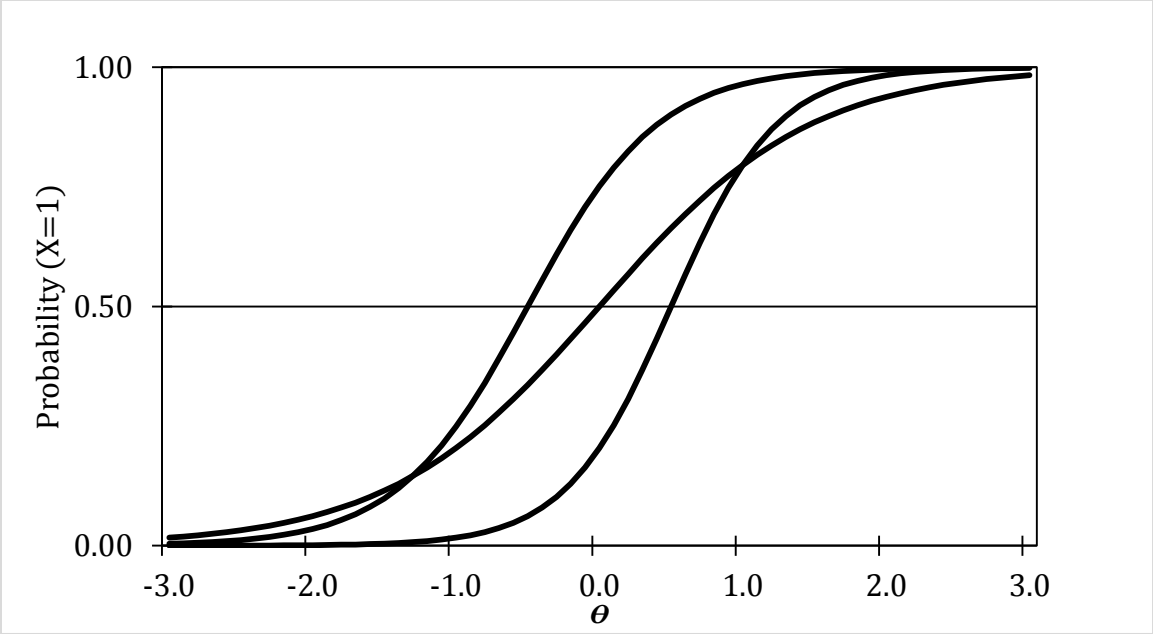


Figure 2: Three item characteristic curves estimated with the 2-PL model.

The value of the *a* parameter is proportional to the slope at the point of inflection; therefore, as the slope of the ICC increases so does the discrimination power of the item. A higher *a* parameter value indicates that the item is better able to distinguish between different levels of examinee ability close to the difficulty level of the item. Similar to the 1-PL model, the 2-PL model does not estimate a parameter to model the probability of guessing the correct response option and therefore has a lower asymptote of zero. The *b* parameter for an item is the ability level that corresponds with point of inflection. The point of inflection of the ICC again is at a probability of 0.50 for correctly responding to

the item. The ICCs for items calibrated with the 2-PL model may cross as the model allows for the discrimination parameters to vary across items.

Three-parameter logistic IRT model. The three-parameter logistic model (3-PL) (Birnbaum, 1968) provides estimates of three item parameters. Building on the previous models, the *b* parameter and the *a* parameter are estimated for each item as well as a *c* parameter which represents a pseudo-guessing parameter. With the 3-PL model, the probability of examinee *j* with a given Θ, where theta represents the examinee's ability level, correctly responding to an item is defined as

$$P(X_i = 1 | \theta_j) = c_i + (1 - c_i) \frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}} \quad (3)$$

where b_i represents the difficulty level, or location of the item along the trait continuum, of item *i*, a_i represents the discrimination power of item *i*, and c_i is a pseudo-guessing parameter to represent the probability of examinee *j* guessing the correct response option for the item. Figure 3 depicts ICCs for two items calibrated with the 3-PL model. For item 1 b = -0.50, a = 0.60 and c = 0.20. For item 2 b = 0.00, a = 0.38 and c = 0.10.

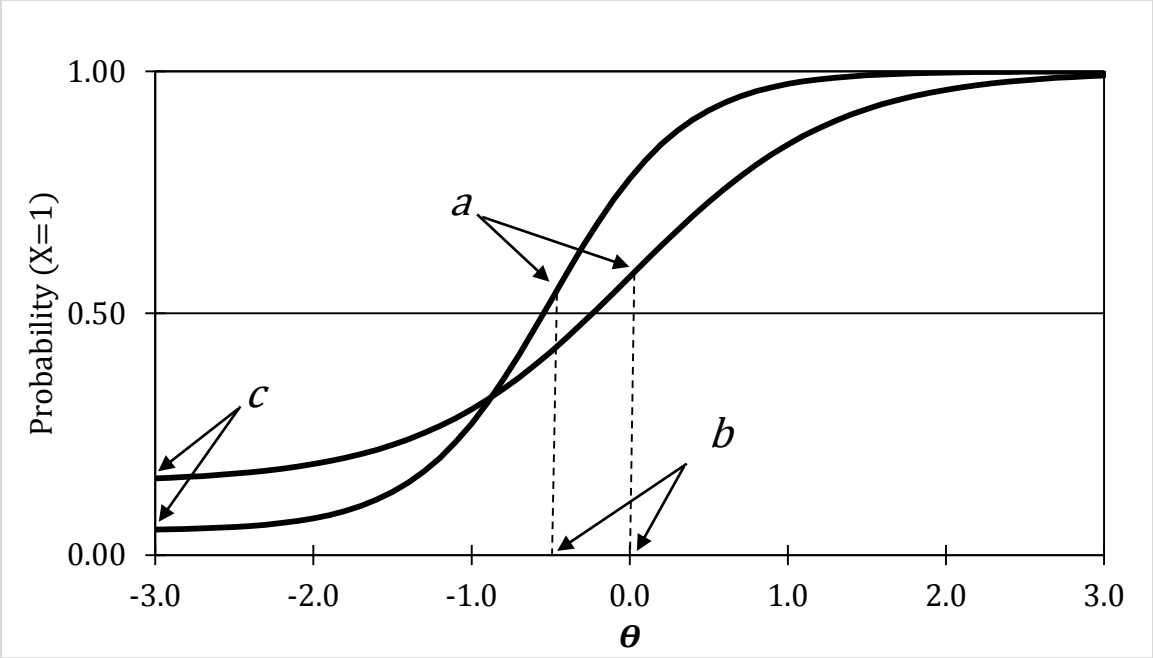


Figure 3: Two item characteristic curves estimated with the 3-PL model.

Unlike the previous models, the 3-PL model includes an estimate of the probability of guessing the correct response. The pseudo-guessing parameter, *c*, is defined as the lower asymptote of the ICC. As the lower asymptote may be greater than zero, the point of inflection will no longer correspond to a probability of correct response being equal to .50, but will be greater. The point of inflection for item *i*, where c_i represents the pseudo-guessing parameter may be defined as

$$\frac{(1+c_i)}{2} \quad . \tag{4}$$

The discrimination parameter is still proportional to the slope at the point of inflection, and the *b* value, or difficulty parameter, is still defined as the Θ value that corresponds to the point of inflection for the 3PL model. However by incorporating the pseudo-guessing parameter with the 3PL, the point of inflection of the ICC differs from the point of inflection produced by using the 1PL and 2PL models. The 3-PL model may reduce into a 2-PL model when the pseudo-guessing parameter for each item is equal to zero. Similarly, the 2-PL model may reduce into the 1-PL model when the discrimination parameters for each item are held constant.

Standard Error and Information for Dichotomous IRT Models. The degree of certainty regarding an individual's estimated ability may be quantified by two related statistics, the standard error (*SE*) and the item's information conditional on the estimated ability. The standard error of the ability estimate will vary across the ability continuum with smaller values indicating greater confidence in the ability estimates and conversely where larger SE indicates less confidence in the ability estimates. The standard error may be calculated as

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}, \tag{5}$$

where $SE(\Theta)$ represents the standard error for a given Θ and $I(\Theta)$ represents the amount of information for the given Θ (Birnbaum, 1968). A small standard error means the

greater is the precision of measurement or information for a given ability level than a large standard error.

The SE represents the level of uncertainty while the information represents the level of accuracy of the measurement. Each item contributes information to the overall ability estimate. The amount of information provided by an item may be calculated as

$$I_i(\theta) = \frac{P_i^{'}(\theta)^2}{P_i(\theta)Q_i(\theta)} \tag{6}$$

where I_i (Θ) represents the information provided by the item, $P_i'(\theta)$ represents the first derivative of $P_i(\theta)$ conditional on Θ, P_i (θ) represents the probability of a correct response to item i, and Q_i (Θ) represents the probability of an incorrect response to item i (Birnbaum, 1968). A major advantage of IRT is that the information can be summed across items to produce the total test information function (TIF) as

$$TI(\theta) = \sum_{i=1}^{n} I_i(\theta) \tag{7}$$

where $TI(\Theta)$ represents the total test information across all items. Larger information values indicate greater precision of measurement for a given level of ability.

COMPUTERIZED ADAPTIVE TESTING

A computerized adaptive test provides several advantages over a traditional fixed-form test. The hallmark advantage of a CAT is the increase in the precision of measurement that may be gained through the adaptive selection of items administered in the test while being able to maintain or decrease test length compared to a traditional linear test format. Through the use of ability estimation procedures and item selection methods, items are selected to improve the precision of measurement in the region of the ability continuum where the examinee's ability is being estimated. Along with the increased precision of estimated examinee ability, CATs offer the ability to terminate a test administration when a pre-specified level of precision pertaining to the examinee's ability estimate has been obtained or a specified number of items have been administered. Either way the test length reduces the testing burden on the examinee while also improving the item exposure of the item pool because only appropriate items are being administered. Items that are too hard or too easy are not administered.

Additionally CAT procedures allow for prompt scoring and reporting as examinee ability is continually recalculated after each item has been administered to the examinee. While some assessments may not have an expressed need for rapid scoring and reporting, other assessments with high-stakes outcomes such as licensure, certification, and assessments for advancement through education may require shorter soring and reporting timeframes as retest opportunities are essential in these settings. Computer-based

assessments also have the ability to improve item security as items can be secured in computer servers while a paper-and-pencil administration allows for the opportunity to have the physical test forms stolen.

Reckase (1989) outlined four major components of CAT procedures that are essential for the development and maintenance of a CAT system: 1) item pool; 2) item selection procedures; 3) ability estimation procedures; and 4) stopping rules. Additionally, content balancing and exposure control methods will also be discussed as they are significant factors of testing programs.

Item Pool. Analogous to the development of items for traditional linear tests, item pools designed for CAT administrations are developed in accordance with the test specifications including all constraints imposed to control item security. Item calibration can be achieved in several different maximum likelihood estimation manners including marginal maximum likelihood, joint maximum likelihood, and conditional maximum likelihood estimation. The most common calibration method, marginal maximum likelihood (Bock & Aitkin, 1981), calibrates items by assuming a standard normal distribution for the theta scale thus enabling score interpretations relative to a distribution with a mean of zero and a standard deviation of 1.0 (Embretson & Reise, 2000). The average likelihood of item parameters is then estimated based on the response strings of the individual examinees and an ability distribution where ability is treated as a known variable (DeMars, 2010).

When developing and maintaining an item pool to obtain the test information, the additive property of item information functions to ensure the measurement of precision at each ability level. For example, when retiring and replacing items from an item pool, new items are selected based on their information function to maintain or increase measurement precision after the appropriate equating of new item parameter estimates has been performed.

Traditional fixed-form norm-referenced test item pools are developed to enable the assembly of fixed test forms which measure across the full range of the ability continuum. These tests will measure ability most accurately around the average ability level so as to place individuals along the ability continuum relative to one another (Wainer, et al., 1990). This results in an optimal TIF being approximately normally distributed about the average ability level of the examinees. Traditional fixed-form criterion-referenced test item pools are developed to facilitate the assembly of fixed item test forms to measure most accurately at a pre-specified cutscore. The purpose of designing test forms around a cutscore is to best determine whether examinees' abilities are estimated to be above or below the cutscore. The specific shape of the TIF for the item pool would depend on the number of cutscores and the location of the cutscores along the ability continuum. For example, a traditional criterion-referenced assessment designed to facilitate classification into three categories using two cutscores would have a

TIF that is bimodal with each mode being leptokurtic with little to no positive/negative skew.

In contrast to the traditional fixed-form assessment format, which utilizes items organized into pre-established test forms, computerized adaptive tests select individual items for administration based on estimated examinee ability. Previous research provides guidelines suggesting that when implementing CAT procedures with a dichotomous IRT model the item pool should be at least 8 to 12 times the size of the number of expected items per test (Stocking, 1994; Way, 1998). The exact number of items required in the item pool to adequately support the test administration is dependent on several factors including the type of scoring model used such as a dichotomous or polytomous IRT model, the item selection method, content balancing procedures, and the exposure control methods. For example, item selection methods may affect the size of the item pool required due to the imposition of strict exposure control constraints. Content balancing within the same item pool requires items spanning multiple content areas providing a sufficient number of items within each content area to satisfy the requirements of the test specifications. Additionally, consideration must be given to item difficulties within each of the content areas that span the pre-specified range of the ability continuum being assessed. CAT item pools have typically been designed to have item difficulty uniformly distributed to provide appropriate items spanning the ability continuum while also

maintaining suitable psychometric properties to meet the assumptions of the IRT model used to calibrate, administer, and score the assessment (Wainer, et al., 1990; Way, 2006).

Testing Algorithm. The testing algorithms involved in CAT procedures require three fundamental processes which include: 1) the commencement of the assessment which includes the specification of the initial ability estimate and initial item selection procedures; 2) the continuation process of the test including item selection procedures based on interim ability estimates and other constraints such as content balancing and exposure control of the items; and 3) options for the termination of the assessment through pre-specified criteria such as maximum test length or maximum standard error (Thissen & Mislevy, 2000).

Ability Estimation. Typically the purpose of using a CAT is to obtain an estimate of an examinee's ability. The examinee's ability estimate is utilized for item selection procedures as well as for providing a final ability estimate. An ability estimation procedure incorporates the examinee's pattern of responses to items in conjunction with the item parameters in order to produce an ability estimate. Individual item parameters must already have been estimated by means of field testing procedures for item pool development and maintenance. When item parameter estimates are used to determine the examinee's ability estimate, the item parameters are no longer treated as estimates but as known item parameters while the examinee's ability is treated as unknown and estimable.

Upon obtaining an initial ability estimate, successive items are selected to provide maximal information at the current or interim ability estimate. The interim ability estimate of each examinee is updated after the examinee responds to each item and subsequently used to select the next item. Three common trait estimation procedures are the Maximum Likelihood Estimation (MLE), Maximum A Posteriori (MAP) and the Expected A Posteriori (EAP) methods which are discussed in the subsequent sections in the context of a dichotomously scored unidimensional item response theory model.

The MLE procedure uses the examinee's response string to a given set of calibrated items to find the Θ value which maximizes the likelihood function. For any response string to a given set of items a likelihood value in log units can be computed for every Θ value along the ability continuum. The maximum likelihood ability estimate is obtained by aggregating the likelihoods conditional on each Θ value and then determining the mode of the likelihood function through a Newton-Raphson iteration procedure (Lord, 1980; Embretson & Reise, 2000).

MLE provides the advantages of less biased ability estimates when compared to the ability estimates provided by the MAP and EAP estimation procedures (Lord, 1986). One problem with the MLE procedure is that it is incapable of providing an ability estimate when the responses to a series of items are all correct or incorrect. In either case all that is known about the ability estimate of the examinee is that it is more extreme than the difficulties of the items that have been administered thus far in the test. The inability

to obtain an ability estimate is problematic especially at the beginning of an exam where examinees may have strings of all correct or incorrect responses and thus MLE ability estimation is not possible.

Two common approaches, the fixed and variable step procedures, have been implemented to ameliorate the issue of adaptive item selection before an initial ability estimate has been established. The fixed step procedure selects subsequent items based on a predetermined step size value until at least one correct and one incorrect response have been obtained. One problem with the fixed step procedure is that individuals may respond in a single category, correct or incorrect, for a series of responses such that the fixed step procedure would attempt to select an item in one of the extreme regions where there are no items with difficulties as extreme as the procedure requires. The variable step size procedure avoids the issue of attempting to select items which are more extreme than those available in the item pool by varying the step size based on the most recently administered item and the item with the appropriate most extreme b value. Items are selected by finding the midpoint between the item previously administered and the item with the most extreme difficulty parameter depending on whether the examinee's responses have been correct or incorrect (Dodd, 1990; Koch & Dodd, 1989). For example, when an examinee has been answering all the questions correctly the variable stepsize procedure will select a more difficult item than the previously administered item by taking half the distance from the last θ estimate and the most difficult item and using

this point of the ability continuum to select an item for administration to the examinee without stepping out of the item pool.

Unlike the MLE method, the MAP and EAP estimation procedures have the advantage of being capable of providing an ability estimate after a single item or based on a response string of all correct or incorrect responses. This is especially advantageous for item selection early in the testing procedure when response strings may be all correct or incorrect as MAP and EAP are able to estimate ability and then use the ability estimate to select items. MAP and EAP methods both make use of a prior distribution when estimating examinee ability by incorporating the prior distribution with the log-likelihood function, given the examinee's response to administered items, to produce an ability estimate. A prior distribution is a hypothetical distribution of randomly sampled examinee abilities which is most commonly assumed to be a standard normal distribution (Embretson & Reise, 2000).

Both MLE and MAP involve iterative processes for ability estimation while the EAP procedure (Bock & Mislevy, 1982) is noniterative. The calculations involved in the EAP ability estimation procedure can be performed more rapidly than the MLE and MAP procedures because they require iterative processes for ability estimation. The faster calculation may be especially important in the CAT context to expedite the estimation of the examinee's ability as items are adaptively selected to match interim ability estimates.

MAP and EAP do, however, have a distinct disadvantage in that the ability estimates produced will be biased. Because the mean of the prior distribution is used as the estimator the ability estimates are biased towards the mean of that prior (de Ayala, 2009). The bias due to the regression towards the mean affects values near the extremes more than values which are nearer to the mean of the ability estimates (Lord, 1986). The problem of regression to the mean may be exacerbated when the prior distribution incorporated is not particularly accurate regarding the distribution of true ability in population of interest. Additionally, Parshall, et al. (2002) recommend not using Bayesian estimation for the final ability estimate as there may be an effect based on the order in which the items were administered such that examinees who take the same set of items and provide the same responses to each item but the order of the presentation of the items varied may have differing ability estimates. While MAP and EAP do present some advantages over MLE for provisional ability estimates, it is important that the prior integrated into the calculations be accurate. It is very difficult for MAP or EAP to overcome a misspecified prior unless the test is very long.

Item Selection. The item selection procedure is the fundamental component of a CAT which provides the "tailored testing" ability. The purpose of the adaptability is to provide the most precise measurement regarding the examinee's ability. The precision of measurement is achieved by selecting items which are most informative conditional on the interim ability estimates of the examinee. Depending on the specifications of the

assessment, item selection procedures may operate while incorporating several constraints such as the balancing of the content required by the test specifications as well as providing some degree of security by managing item exposure.

When no prior information regarding the examinee's ability is being taken into account the initial item selected for an examinee is commonly selected at the mean of the theta distribution or the peak of the item pool distribution. When it is reasonable to assume that the ability being assessed is normally distributed, selecting an item with a difficulty corresponding with the mean of the theta distribution is reasonable as one's best guess regarding the ability of examinee with no prior information.

If prior information about an examinee's ability is available, such as a test score relating to the constructs of interest, this information may be used to inform the selection of the initial item. Otherwise, the initial item is selected using an item selection procedure to produce maximal information at the mean of the ability distribution when using MLE estimation procedures. When using EAP or MAP estimation procedures, items are selected to minimize the expected posterior standard deviation which may result in selecting different items when compared to the items which would be selected for MLE estimation (Embretson & Reise, 2000). Upon obtaining an initial estimate of the examinee's ability level, one of the two following common item selection procedures, Fisher's information (Lord F. M., 1980) and Owen's Bayesian procedure (Owen, 1969), are used for the item selection procedure.

The Fisher maximum information procedure selects items to provide the largest amount of psychometric information for examinees given their interim ability estimates. After each item response is obtained, the CAT algorithm calculates a new provisional ability estimate which is used for the selection of the subsequent item. The procedure repeats by calculating provisional ability estimates and selecting items to maximize information after each item is administered until a stopping rule is invoked to terminate the assessment.

When utilizing Bayesian estimation, the procedure is designed to minimize the expected posterior standard deviation or maximize the precision of the posterior ability estimate. After each item response is obtained the CAT algorithm calculates the expected posterior distribution of the trait which distribution is then used to select an item which will then minimize the expected posterior standard deviation or maximize the precision of the posterior ability estimate. This procedure repeats by recalculating the expected posterior distribution of the ability estimate after each item administration until the stopping rule criteria have been satisfied.

Content Balancing. When a test covers multiple content areas it becomes necessary to ensure that all possible forms of the test are unbiased in representing the content areas. One method to manage multiple content areas is to separate the content areas completely and to estimate an examinee's ability for each content area independently. This is appropriate when the test measures multiple dimensions. When the

test measures a single dimension the content areas are not separated but sampled together to form tests which cover all content areas, using content balancing procedures to conform to test specifications. Traditional linear tests follow the test specifications when building the multiple forms to achieve the appropriate sampling of content areas for their fixed forms.

Content balancing for computerized adaptive testing becomes more challenging as the test adapts to the examinee's ability and thereby presents a somewhat unique series of questions to each examine. To ensure equity across examinees and to satisfy the test specifications for each examinee, the CAT program must employ some content balancing procedure. A common content balancing procedure is the constrained CAT proposed by Kingsbury and Zara (1989). The constrained CAT operates by comparing the proportions of the items administered by content area to the target proportion values for each content area. The content area with the largest discrepancy between the actual administered value and the target value is used to select the item for administration. Once a content area is identified, an item is selected to provide maximum information when using MLE or to minimize the expected posterior standard deviation for Bayesian estimation procedures given the current provisional examinee ability estimate. The constrained CAT procedure repeats by calculating the discrepancies between the administered value and target values for each content area, identifying the content area for item selection and selecting the

most efficient item based on the ability estimation procedure until a stopping rule is invoked.

Exposure Control. Item exposure must be given consideration for CATs, otherwise the Fisher maximum information and Bayesian item selection procedures would constantly select the same item for a given ability level due to its maximally informative psychometric properties. The continual selection of the same item, or set of items, would overexpose the maximally informative items and underutilize or possibly fail to ever select certain items. Items that would fail to be selected when exposure control methods are not incorporated into the CAT algorithm are not items with poor psychometric qualities, but rather are simply ranked behind the most useful items. Additionally, exposure control methods may discourage and decrease the effects of examinees banking and sharing items. A general guideline regarding maximum exposure rate of 20%, referring to the percentage of examinees to whom the item is administered, was suggested by Spray (Parshall, et al., 2002).

Way (1998) used two categories, randomization procedures and conditional selection procedures, to describe some of the more common strategies for item exposure control. Randomization procedures seek to control item exposure by using a random selection component in determining which items will be available to be administered to the examinee. The rationale for randomization-type methods is that the random selection provides a means whereby the chances that examinees with the same or similar ability

estimates will be administered the same items or series of items is purely random. It is reasonable that items selected for possible administration will differ even for examinees with the same interim ability estimates (Geeorgiadou, Triantanfillou, & Economides, 2007). The Randomesque procedure proposed by Kingsbury and Zara (1989) is an example of a well-known randomization procedure. The Randomesque procedure selects a prespecified set of items (i.e. five items) which are selected to maximize information at the current ability estimate of the examinee. From the set of items, one item is randomly selected to be administered to the examinee.

Conditional procedures rely less on a random component in item selection and make use of an exposure control parameter to restrict which items are available for administration. A set of items are selected to provide maximum information or to minimize the expected posterior distribution based on the provisional ability estimate. An item is then selected from the set for administration based on the exposure control parameter. The value of the exposure control parameter is predetermined through an iterative simulation process whereby each item is assigned an appropriate value given the frequency at which it is expected to be selected for administration. Two well-known conditional strategies are the Sympson-Hetter (Sympson & Hetter, 1985) and the Sympson-Hetter Conditional (Chang, 1998).

Combination procedures also have been developed to take advantage of the benefits of both the randomization procedures and the conditional procedures. A well-

known combination procedure is the Progressive-Restrictive procedure developed by Revuelta and Ponsoda (1998). The Progressive-Restrictive procedure combines a random component with an exposure control limit (Revuelta & Ponsoda, 1998).

Stopping Rules. One of the fundamental benefits of CAT is the ability to terminate the test when a desired level of precision regarding the ability estimate has been obtained (variable-length) or after a specified amount of time has expired (Thissen & Mislevy, 2000). The adaptive item selection based on interim ability estimates enables CATs in some situations to provide tests which are half the length of traditional fixed-form tests while maintaining or improving the precision of the ability estimate (Embretson & Reise, 2000). Often a combination of the maximum number of items and the precision of measurement, $SE(\theta)$, are used to terminate CATs (Davis, 2002; Thissen & Mislevy, 2000; Wainer, et al., 1990; Weiss & Kingsbury, 1984).

Fixed-length exams can be less cumbersome when implementing constraints such as content balancing since the forms are predetermined and fixed as pertaining to the number of items and which items in each content category will be administered given the test specifications. Generally, fixed-length tests do not provide equivalent levels of precision of measurement for examinees across all levels of the ability continuum (Wainer, et al., 1990). Fixed-form assessments terminate upon reaching the maximum number of items.

Variable length CATs are designed to terminate the test once an acceptable level of precision for the ability estimate has been achieved. The logic behind the variable length test is that when examinees respond consistently for their ability level, the standard error of their estimate can be reduced until it achieves a specified $SE(\theta)$ value and terminate the test providing a reasonably accurate ability estimate. When using the MLE estimation procedure the stopping rule uses maximum information to select items to reduce the standard error. The Bayesian procedures result in the stopping rule focusing on achieving a target posterior distribution to achieve the pre-specified level of measurement precision to terminate the assessment. Variable length tests conserve items by administering items which are efficient for estimating examinee ability and terminating the test before reaching the maximum test length. Problems may arise when item selection is inefficient or when examinee responses are inconsistent thereby producing longer exams and exposing more items to achieve the specified level of precision of measurement to satisfy the level of precision needed to stop the assessment. In that case a maximum number of items may be used along with a prescribed standard error to stop the CAT. Adaptive variable length assessments have the advantage of providing final ability estimates with similar levels of precision across all levels of the ability continuum.

COMPUTERIZED CLASSIFICATION TESTING

Typically achievement and aptitude assessments have been developed with the purpose of determining a point estimate of an examinee's ability while classification testing is comprised of a unique set of testing procedures and test specifications where the target outcome is a classification decision. Computer-based testing may be preferred to paper-and-pencil methods for making classification decisions especially when the results of the assessment are high stakes in nature. Computerized classification adaptive tests based on larger item pools with reasonable exposure control and content balancing procedures provide improved item exposure and security measures when compared to fixed-form assessments. When the classification procedure requires an ability estimate to determine classification status, CAT methodology can provide improved ability estimates while utilizing fewer items than a fixed form. Additionally, some computerized adaptive classification testing procedures which do not require an examinee ability estimate for classification, namely the sequential probability ratio test (SPRT), have been found to provide accurate classification while utilizing fewer items than typical CATs (Parshall, et al., 2002).

The purpose of classification testing procedures is to evaluate an examinee in relation to a pre-specified cutscore and provide a categorical outcome. In past research classification testing has been referred to as criterion-referenced measurement (CRM), mastery testing, and adaptive mastery testing (AMT) (Weiss, 1983; Wainer, 1990). The

purpose of implementing an adaptive mastery testing procedure is to maximize the percent of correct classifications while reducing the number of items required to a classification decision (van der Linden & Glas, 2010). The term ‘mastery testing’ implies that a dichotomous decision (i.e., pass/fail) is the outcome of the procedure. It has become more common to refer to testing procedures used for categorical decisions as classification testing which allows for classification into more than two categories (i.e. below average/average/above average). Using a CAT as opposed to a fixed-form test to produce classification for multicategorical decisions provides the unique ability to select items with optimal properties to achieve classification. It has been demonstrated that accurate classifications can be achieved with as few as half the number of items used for ability estimation tests in a typical CAT (Lewis & Sheehan, 1990). Based on the purposes of the testing and the importance of the classification outcome, different methods may be employed to provide the most accurate and useful classifications.

Computerized classification testing procedures are capable of handling the complex requirements that are imposed on typical CAT procedures. While some of the issues such as content balancing and item exposure are the same as typical CAT procedures, other components and operations of a classification CAT differ with respect to the item pool, the item selection and testing algorithm, scoring, and termination criteria. Additionally, depending on whether the tests are used for simple dichotomous

(e.g. pass/fail) decisions or for multiple categorical classifications (e.g. advanced/pass/fail) the procedures may vary.

Classification testing procedures can be loosely organized according to whether or not the procedure achieves a classification decision based on ability estimation. When using MLE estimation procedures a CAT algorithm developed for classification testing selects individual items based on maximum information as a typical CAT used for ability estimation would. The ability confidence interval (ACI) method is considered an estimation-based classification method. The ACI functions identically to typical CAT procedures in all respects with the exception of the stopping rule. The stopping rule for an ACI requires a confidence interval to be calculated for the provisional ability estimate of the examinee. ACI testing procedures terminate the assessment when the confidence interval no longer contains the cutscore which results in a classification. In cases when the maximum number of items have been administered, the examinee's final ability estimate is compared to the cutscore to determine classification status. The ACI method is useful especially when a final ability estimate is needed. The main drawback when using the ACI method is that it tends to produce longer test lengths than competing classification testing procedures.

The computerized master testing (CMT) proposed by Sheehan and Lewis is a testlet-based procedure used for classification testing (Lewis & Sheehan, 1990; Sheehan & Lewis, 1992; Parshall, et al., 2002). Testlets are a grouping of items that "may be

developed as a single unit that is meant to be administered together" (Wainer, Bradlow, & Wang, 2007). Though testlet selection is not based on estimated examinee ability, but rather random selection of the testlets, the procedure does compute a final ability estimate for the examinee.

Finally, some classification testing procedures do not base item selection or classification decision on ability estimates. For example, the likelihood ratio-based procedures utilize a point-hypothesis approach with an accompanying statistical test instead of an ability estimate to determine classification. The sequential probability ratio test procedure is considered a statistical testing procedure as after each item is administered, a set of hypotheses are evaluated with a likelihood ratio. Item selection for SPRT and other likelihood ratio testing procedures is based on maximum information at the cutscore. Typically an ability level has not been estimated for examinees when likelihood ratio-based testing has been used to determine classification. The lack of a final ability level estimate has been regarded as a disadvantage of the SPRT procedure. The advantage of the SPRT procedure has been the conservation of items as classification decisions have been consistently reached with fewer items when compared to the ACI method (Spray & Reckase, 1996; Eggen & Straetsmans, 2000; Thompson, 2011). The SPRT procedure operates quite differently than typical CAT and ACI procedures and as this study is based on a likelihood ratio testing procedure, namely the sequential probability ratio test, the following sections will focus on the SPRT-based methodology.

Sequential Probability Ratio Testing

The sequential probability ratio test (SPRT) procedure was developed by Abraham Wald (1947) as a means of conserving supplies in the quality control testing procedures during the Second World War. As part of the quality control measures large samples were drawn from the supply lines to be tested to examine the quality of the product. Initially the whole sample, or batch, was used in the testing and thereby rendered unserviceable. This presented a problem in that supplies during the war were already difficult to come by and the quality control procedures were consuming part of the products that could have been otherwise used by the military. Wald recognized an opportunity to conserve products by sequentially sampling single products until the likelihood of the batch passing or failing the quality control testing could be determined. Upon determining the likelihood of passing/failing, the remainder of the batch could then be returned to the supply lines for use by the soldiers.

SPRT was eventually suggested for use in psychological testing and extended to computerized classification testing (Ferguson, 1969; Epstein & Knerr, 1977; Reckase, 1983; Kingsbury & Weiss, 1983; Spray & Reckase, 1987, 1994, 1996). In 1969 Ferguson studied the utility of the SPRT procedure in evaluating student mathematics ability. Ferguson's procedure proved to be useful in reducing testing time and test length. Epstein and Knerr applied the SPRT procedure to military testing in 1977 resulting in similar improvements in test lengths as Ferguson's studies. Kingsbury and Weiss (1983)

examined the classification accuracy of traditional classification testing methods and SPRT procedures. When optimal items pools were used where the majority of items have maximal information proximate to the cutscore, the classification accuracy rates for the traditional classification procedures and the SPRT procedures produced very similar results, 87% and 86% respectively. When a less than optimal item pool was used for the SPRT procedure the classification accuracy dropped significantly emphasizing the importance of an appropriate item bank for the procedure. It is important to note that theoretically the SPRT procedure does not terminate a test when it has reached a prespecified number of items though it is often considered to do so in research. The truncated SPRT (TSPRT) procedure however does include the criteria of a maximum number of items to terminate a test and has been used in research (Eggen, 1999; Spray & Reckase, 1996; Vos, 2000).

Outlined by Parshall, et al. (2002) are the steps of an TSPRT computerized classification test procedure including (1) specifying the TSPRT parameters, (2) item pool development for selection and administration, (3) item administration, (4) calculation of the likelihood ratio, and (5) the classification status. Steps 3 through 5 will continually repeat until an examinee is given a final classification resulting in the termination of the test or until the maximum test length is reached at which time a classification decision made.

Specification of TSPRT Parameters

TSPRT procedures continually test a set of basic hypotheses for an examinee with an ability level (θ) which results in a classification decision. The hypotheses are structured as

$$H_0 : \theta_j = \theta_c - \delta = \theta_0$$

$$H_1 : \theta_j = \theta_c + \delta = \theta_1$$

where θ_j represents the ability level of the examinee, θ_c represents the passing score, δ represents the indifference region, θ_0 represents the maximum lower bound decision value for classification, and $\theta_1$represents the minimum upper bound decision value for classification and thus $\theta_0 < \theta_c < \theta_1$ (Parshall, et al., 2002). The null hypothesis, H_0, states that the examinee's ability level, θ_j, is equal to lower-bound point of the indifference region, θ_0, which results in a classification status below the cutscore. The alternative hypothesis, H_1, states that the examinee's ability is equal the upper-bound point of the indifference region, θ_1, resulting in a classification status above the cutscore.

To begin the TSPRT procedure a cutscore, θ_c, must first be defined as the upper and lower bound of the indifference region, θ_1 and θ_0, are both dependent on the cutscore value. An appropriate standard setting procedure should be used to obtain the cutscore based on the nature of the stakes involved in passing and failing the assessment. Because the indifference region is dependent on the cutscore, in certain circumstances it

may be useful to choose values for the indifference region during the standard setting. Assessment programs may elect to set indifference regions after simulation studies have been used to explore various allowable passing and failing scores based on error rates. Otherwise, the indifference region values are set to balance efficiency and accuracy based on the characteristics of the item pool. Commonly an indifference region is selected to be symmetrical around the cutscore (e.g. δ = 1.0, $\theta_c = 0.0,\ \theta_0 = -0.5,\ \theta_1 = 0.5$) but this is not a requirement. As the consequences of false-positives and false-negatives may differ based on the purposes of the test, indifference region boundaries may be selected to be asymmetrical. Selecting the indifference region involves balancing the trade-offs between test length and classification error. Typically the larger the indifference region the shorter the test will be especially for individuals with abilities further from the cutscore. Individuals near the cutscore will have somewhat longer tests as well as increases in classification errors. Smaller indifference regions are better able to maintain classification accuracy but typically produce lengthier tests.

As with any classification procedure, there are inherent errors in classification. TSPRT procedures require that the error rates, α and β, be specified before the procedure is implemented. The false positive classification error rates, errors in which the examinee is classified as passing when their true ability level is less than the cutscore, are represented with α. The false negative classification error rates, errors in which the examinee was classified as failing when their true ability level is actually greater than the

cutscore, are represented by β. Both types of error rates may range from 0 to 1, but are typically fixed at .05 or .10 (Parshall, et al., 2002). The Type I and Type II error rates, *A* and *B* respectively, are used to calculate the upper and lower boundaries for the classification decision.

$$\text{Upper boundary: } A = (1 - \beta) / \alpha \tag{8}$$

$$\text{Lower boundary: } B = \beta / (1 - \alpha) \tag{9}$$

TSPRT for multiple categorizations is similar to the procedure for dichotomous categorization with the exception that multiple sets of hypotheses are being evaluated simultaneously (Eggen, 2000). For example, when two cutscores are used to classify examinees into three categories (e.g. below/average/advanced) there are two sets of hypotheses.

The hypotheses may be represented as:

$$H_{01} : \theta_j = \theta_{c1} - \delta_{01} = \theta_{01} \text{ (level 1)}$$

$$H_{11} : \theta_j = \theta_{c1} + \delta_{11} = \theta_{11} \text{ (above level 1)}$$

$$H_{02} : \theta_j = \theta_{c2} - \delta_{02} = \theta_{02} \text{ (below level 3)}$$

$$H_{12} : \theta_j = \theta_{c2} + \delta_{12} = \theta_{12} \text{ (level 3).}$$

where θ_{c1} represents the lower of two cutscores with the accompanying indifference region boundaries, $\delta_{01}\ and\ \delta_{11}$, and θ_{c2} represents the higher of two cutscores with the accompanying indifference region boundaries, $\delta_{02}\ and\ \delta_{12}$. A set of boundaries for each hypothesis must be specified and for the purposes of the current research it can be assumed that $\alpha_0 = \alpha_1 = \beta_0 = \beta_1 = \alpha,\ \delta_{01} = \delta_{11} = \delta_{02} = \delta_{12} = \delta$, and $\ln(1 - \alpha)/\alpha = A$.

Item Pool Development for Selection and Administration

The TSPRT procedure is more efficient with an item pool where the difficulty of the items closely match the cutscore or cutscores used for evaluating the examinees. This is markedly different from typical CAT item pools where items are developed to span the full range of the ability scale. TSPRT procedures are efficient with item pools the distribution of item difficulty has minimal skewness and minimal dispersion around the cutscore or around each of cutscores being used to classify examinees. The further an item's difficulty is from the cutscore, the less efficient the item is in providing information for classification procedure. Similar to typical CAT procedures, items with a higher discrimination value and a lower guessing parameter are more useful for TSPRT procedures. Items with higher discrimination and lower guessing parameters are more efficient for the scoring method as the probabilities associated with the examinee's response are disparate and thereby more informative. Once the item pool is developed, by calculating the probabilities of correct and incorrect responses for both θ_0 and θ_1 for each

cutscore, the individual items can be prepared for a more rapid scoring process before the actual test administration. It has been generally accepted that classification testing items pools for SPRT procedures do not require item pools that are as large as typical CAT items pool. Parshall, et al. (2002) found that it has been conventional to use an exposure rate of approximately 20% resulting in functional item pools which contain five times that maximum number of items.

Item Administration

For dichotomous categorization the item selection procedure is straightforward as there is only a single cutscore and the item pool should be closely distributed around the cutscore. Provisional ability estimates are not used to select items. Selecting items to provide maximum information at the cutscore has proven to be useful for TSPRT procedures (Spray & Reckase, 1994; Thompson, 2007, 2009). Without additional constraints the item selection procedure would continually select the same series of items so as to most closely match the cutscore and have high discrimination parameter values. Exposure controls are employed to manage overexposure of items while content balancing ensures that the test specifications are met.

Multiple category classification mimics typical CAT procedures while adjusting to accommodate the TSPRT procedure. When using multiple cutscores the TSPRT procedure does not employ an ability estimation procedure but rather selects items based

on the maximum item information for the cutscore deemed closest to the examinee's ability. To determine which cutscore is closest to an examinee's ability when an examinee's interim score is between two cutscores the following are used:

the minimum of

$$\left| \sum_{i=1}^{k} a_i x_i - \left\{ A - \sum_{i=1}^{k} \ln\left(\frac{p_i\left(\theta_{C1} - \delta_{01}\right)}{p_i\left(\theta_{C1} + \delta_{11}\right)} \right) \right\} / 2\delta \right| \tag{10}$$

and

$$\left| \sum_{i=1}^{k} a_i x_i - \left\{ A - \sum_{i=1}^{k} \ln\left(\frac{p_i\left(\theta_{C2} - \delta_{02}\right)}{p_i\left(\theta_{C2} + \delta_{12}\right)} \right) \right\} / 2\delta \right| \tag{11}$$

where a_i and x_i represent item discrimination and item responses respectively while the other variables were previously defined (Eggen & Straetmans, 2000). This form of multiple category classification using TSPRT is considered a cutscore-based item selection procedure for evaluating the examinee wherein items are selected to maximize one of the likelihood ratios being used to classify the examinee.

Calculation of Likelihood Ratio

With the parameters specified and an item administered, the likelihood ratio test is calculated to determine the likelihood of classification. The likelihood ratio, LR, is the ratio of the likelihoods of the response given both θ_0 and θ_1 for each cutscore (Parshall, et al., 2002; Thompson, 2007),

$$LR = \frac{L(\theta = \theta_1)}{L(\theta = \theta_0)} = \frac{\prod_{i=1}^{n} P_i(X = 1 | \theta = \theta_1)^x P_i(X = 0 | \theta = \theta_1)^{1-x}}{\prod_{i=1}^{n} P_i(X = 1 | \theta = \theta_0)^x P_i(X = 0 | \theta = \theta_0)^{1-x}} \quad (12)$$

or equivalently,

$$LR = \frac{L(\theta = \theta_1)}{L(\theta = \theta_0)} = \frac{L(x_1, x_2, ..., x_k | \theta_1) = \pi_1(\theta_1)\pi_2(\theta_1)...\pi_k(\theta_1)}{L(x_1, x_2, ..., x_k | \theta_0) = \pi_1(\theta_0)\pi_2(\theta_0)...\pi_k(\theta_0)} \quad (13)$$

where

$$\pi_i(\theta_1) = P(\theta_1)^{x_i} Q(\theta_1)^{1-x_i} \quad (14)$$

and

$$\pi_i(\theta_0) = P(\theta_0)^{x_i} Q(\theta_0)^{1-x_i} . \quad (15)$$

The item characteristic curve displayed in Figure 4 can be used to demonstrate where the probabilities associated with each of the indifference regions boundaries are obtained.

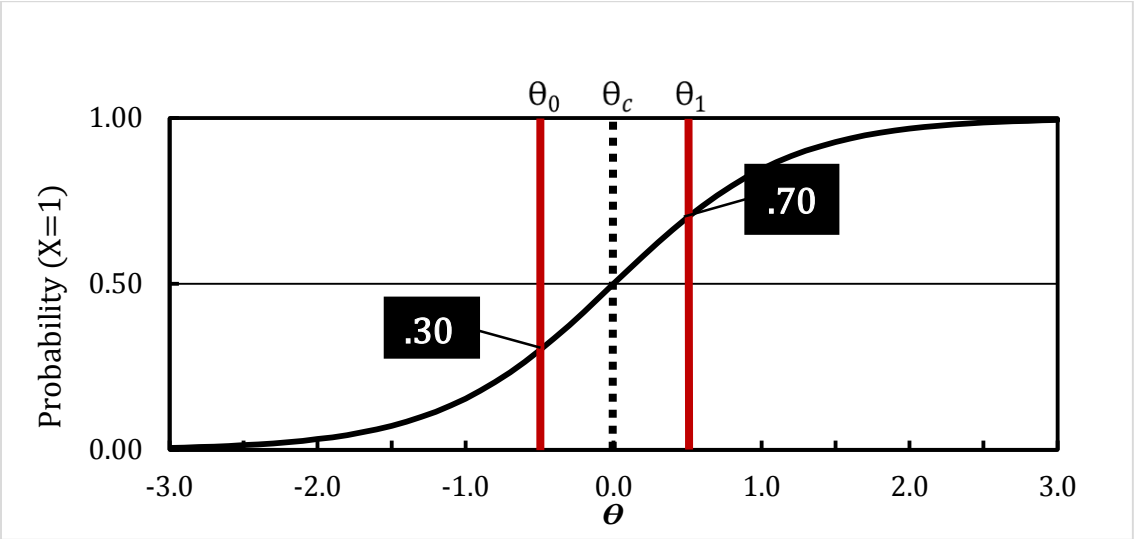


Figure 4: Probabilities of a correct response corresponding with indifference region boundaries where the item difficulty parameter, b = 0.0, matches the cutscore value.

As examinees progress through the test, the likelihood ratio is continually updated where correct responses are incorporated as

$$LR = \frac{L(x_i = 1 | \theta_1)}{L(x_i = 1 | \theta_0)} \tag{16}$$

and where incorrect responses are incorporated as

$$LR = \frac{L(x_i = 0 | \theta_1)}{L(x_i = 0 | \theta_0)} \tag{17}$$

The resulting value of the calculated ratio is then compared to the decision points, *A* and *B*, for a classification or to continue administering test items.

For example, an examinee's response string to five items where the examinee responded correctly to every other item (e.g. 10101) can be expressed as

$$\text{LR} = \frac{P(x_1 = 1 | \theta_1)P(x_2 = 0 | \theta_1)P(x_3 = 1 | \theta_1)P(x_4 = 0 | \theta_1)P(x_5 = 1 | \theta_1)}{P(x_1 = 1 | \theta_0)P(x_2 = 0 | \theta_0)P(x_3 = 1 | \theta_0)P(x_4 = 0 | \theta_0)P(x_5 = 1 | \theta_0)} \tag{18}$$

where substituting probabilities and computing a score would be

$$\text{LR} = \frac{(.70)(.15)(.60)(.20)(.75)}{(.30)(.75)(.20)(.65)(.20)} = \frac{0.0095}{0.0059} = 1.62$$

When both the error rates, alpha and beta, are set to .05 the *A* and *B* values are calculated as

$$A = \frac{1-\beta}{\alpha} = \frac{.95}{.05} = 19$$

and

$$B = \frac{\beta}{1-\alpha} = \frac{.05}{.95} = .05$$

The log form of each value is used when comparing the value from the likelihood ratio to the *A* and *B* values so that when α and β are equal, the boundaries for classification decisions are symmetrical around 0. Thus, the log value of the likelihood ratio is 0.21, the *B* value is -1.28, and the *A* value is 1.28. As the likelihood value does not surpass either boundary value, $-1.28 < 0.21 < 1.28$, the test would continue to administer items.

Because the boundaries of the indifference region are fixed and unchanging with the TSPRT procedure, items with difficulty parameters which do not closely match the cutscore value are less efficient than items with difficulty parameters which closely match the cutscore. To illustrate the inefficiency of items where maximum information is not at the cutscore, Figures 5 and 6 display the probabilities of correct responses associated with the boundaries of the indifference region. The differences in the probabilities for the indifference region boundaries in Figures 5 and 6 is only 0.22

whereas the difference between the probabilities of the indifference region for the item in Figure 4 is 0.40.

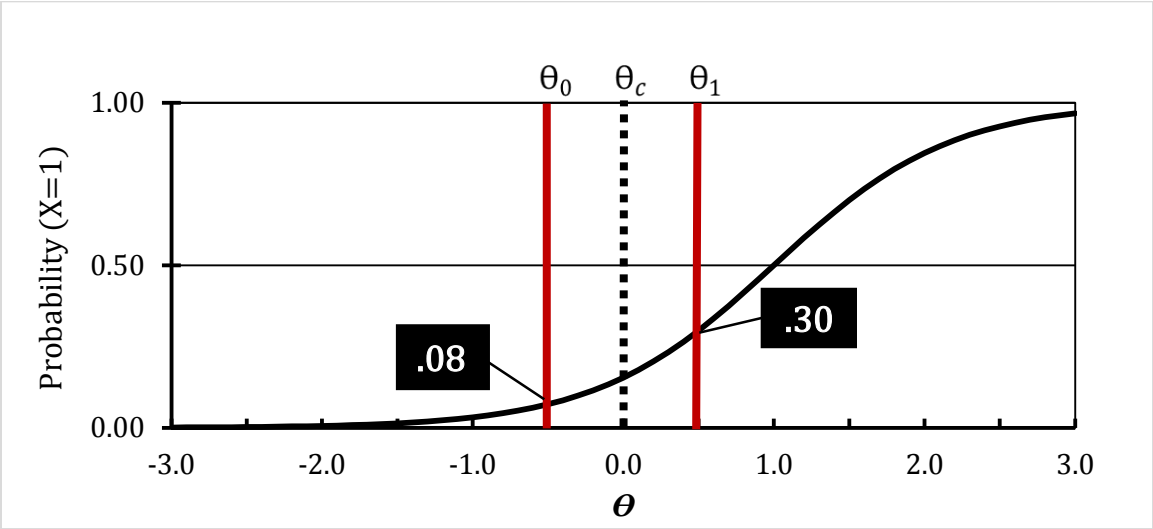


Figure 5: Probabilities of a correct response corresponding with indifference region boundaries where the item difficulty parameter, *b* = 1.0, is above the indifference region.

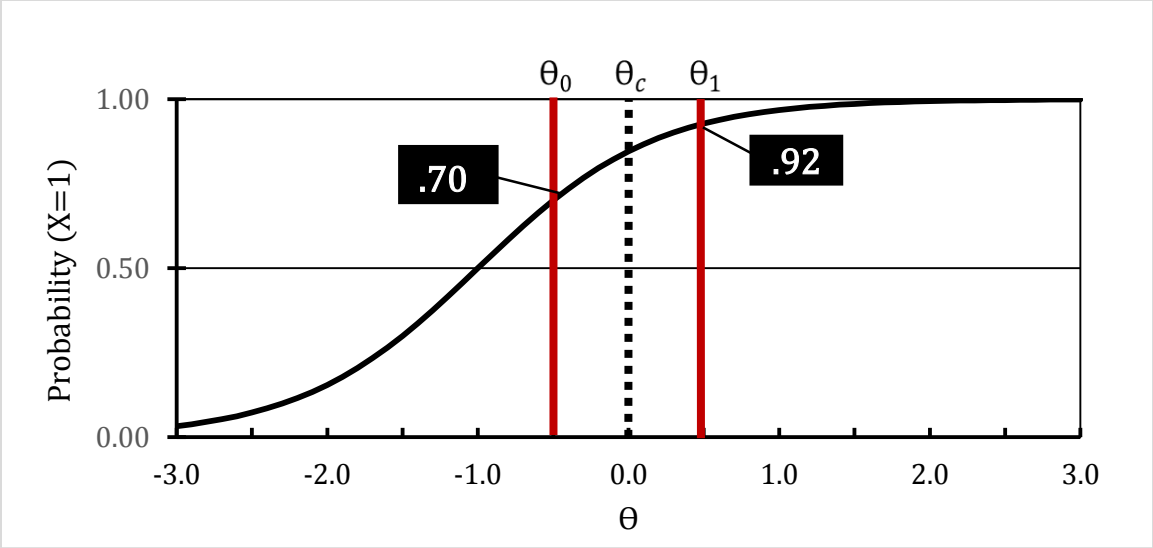


Figure 6: Probabilities of a correct response corresponding with indifference region boundaries where the item difficulty parameter, b = -1.0, is below the indifference region.

Classification

As the guiding principle behind TSPRT procedures is to conserve items while maximizing classification accuracy, the TSPRT procedure will seek to terminate the test before reaching the pre-specified maximum number of items. Termination of a test using TSPRT for dichotomous categorization is simple and straightforward. As examinees progress through the test items the probability ratio is continually recalculated incorporating the new responses and administering additional items until either one of the hypotheses (i.e. the examinee's ability is greater than θ_1 or the examinee's ability is less than θ_0) is satisfied (Spray, 1993). As each item is incorporated into the score, the value

of the likelihood ratio is compared to the *A* and *B* values to determine if the examinee can be classified or if another item should be administered. When the likelihood ratio value is greater than *A* the examinee is classified as passing. When the likelihood ratio value is less than *B* the examinee is classified as failing.

$$L(x_1, x_2, \ldots, x_n, |\theta_0, \theta_1) \geq A, \text{ reject } H_0, \text{ classify as passing.} \quad (19)$$

$$L(x_1, x_2, \ldots, x_n, |\theta_0, \theta_1) < B, \text{ fail to reject } H_0, \text{ classify as failing.} \quad (20)$$

$$B < L(x_1, x_2, \ldots, x_n, |\theta_0, \theta_1) < A, \text{ continue testing.} \quad (21)$$

In cases where examinees do not demonstrate an ability level sufficient to satisfy either hypothesis, a maximum test length is used to terminate the tests. Classification of the examinees is then achieved by calculating the difference between the likelihood ratio and the *A* and *B* values and selecting the classification which minimizes the difference.

Termination of the test when using multiple categorizations is similar to the single cutscore method with the exception that one or all of the statistical tests may have produced classifications. When using traditional TSPRT procedures for two cutscores, if the examinee is above the higher cutscore or below the lower cutscore it is possible that only the likelihood ratio for the cutscore closest to the examinee's ability will classify the examinee and terminate the test. The inability of the cutscore furthest from the examinee's ability to provide a classification decision is a function of the lack of

information used in calculating the likelihood ratio as items are being selected to provide maximum information for the cutscore closest to the examinee's ability level.

Instances in which examinees are between the two cutscores would require that the likelihood ratio for the lower cutscore classify the examinee as above the lower cutscore while the likelihood ratio for the upper cutscore would have to classify the examinee as being below the upper cutscore to terminate the test before reaching the maximum test length. Otherwise the test continues until the maximum numbers of items have been administered. If the maximum number of items has been administered then the likelihood ratios for each cutscore are examined and classification status is decided by calculating the difference between each likelihood ratio and the *A* and *B* values and selecting the classification which minimizes the difference.

Generalized Likelihood Ratio

Optimally item banks used for TSPRT procedures would contain items at or vary near the cutscore. Realistically item pools will be distributed about the cutscore to varying degrees. Items with difficulties that are further from the cutscore are less efficient. Hence the generalized likelihood ratio (GLR) (Huang, 2004; Bartroff, Finkelman, & Lai, 2008; Thompson, 2007, 2009) was proposed to ameliorate the inefficiency of items where the difficulty parameter does not closely match the cutscore value.

The GLR is functionally equivalent to the TSPRT procedure in that the specification of parameters, the calculation of the likelihood ratio, and classification process are all the same. The sole difference between the TSPRT and GLR procedure is that θ_0 and θ_1 are fixed for the TSPRT procedure whereas θ_0 and θ_1 are allowed to vary in the GLR procedure. Based on the parameters specified for the TSPRT procedure, item difficulties may lie outside of the indifference region. The varying of θ_0 and θ_1 in the GLR procedure allows for the maximum of the likelihood function to be incorporated into the calculation when it does not lie within the indifference region.

For example, if a cutscore were 0 with a symmetrical indifference region θ_0 = -0.2 and θ_1 = 0.2, and the item's difficulty were 0.5, the GLR procedure would utilize the maximum of the likelihood function, 0.5, in place of the 0.2 value pre-specified for θ_1. When the maximum of the likelihood function is below or above the boundaries of the indifference region, the values of the boundaries are adjusted for that single item to incorporate the maximum of the likelihood function. Given the results of the research, Thompson (2007; 2009) advocated that the GLR procedure would never be less efficient than the SPRT method and that in it had in certain instances shown evidence of greater efficiency in terms of average test length. His results also suggested that the GLR had no loss in the accuracy of the classifications produced by the procedure.

The TSPRT maintains fixed parameter values for the indifference cutpoints, θ_0 and θ_1, and by doing so maintains a focus on evaluating examinee ability between these

two cutpoints. By allowing the boundaries of the indifference region to vary as the GLR does, it has been demonstrated that the procedure can produce shorter tests. To illustrate the flexibility of the indifference region boundaries provided by the GLR procedure, Figures 7 and 8 include a θ_{G1} or θ_{G0} which is used in place of the original indifference region parameters, θ_1 or θ_0, when the difficulty parameter falls outside of indifference region.

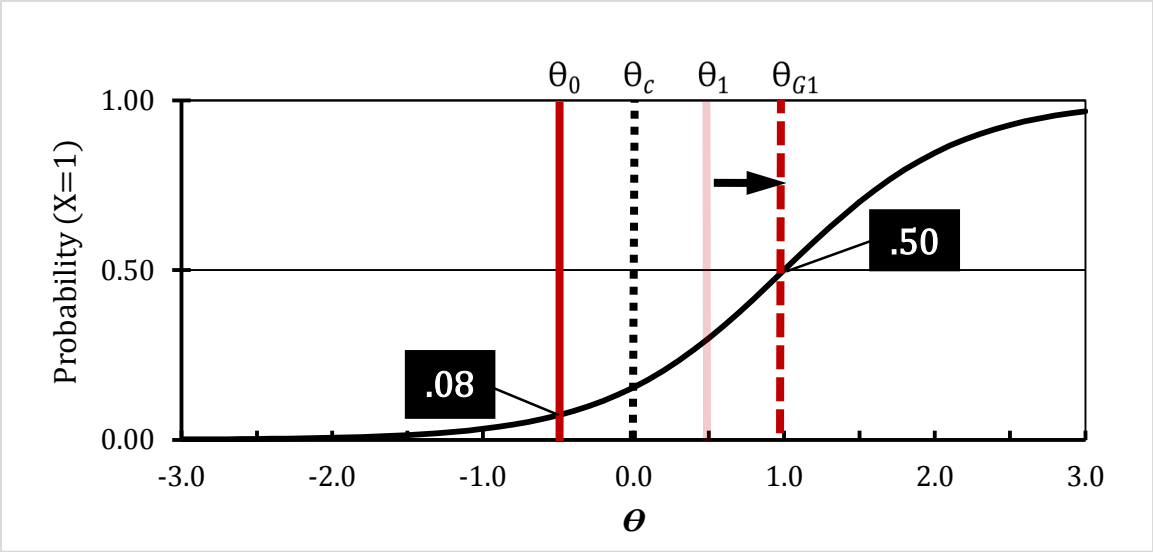


Figure 7: Probabilities of a correct response corresponding with indifference region boundaries of the GLR where the item difficulty parameter, b = 1.0, is above the indifference region.

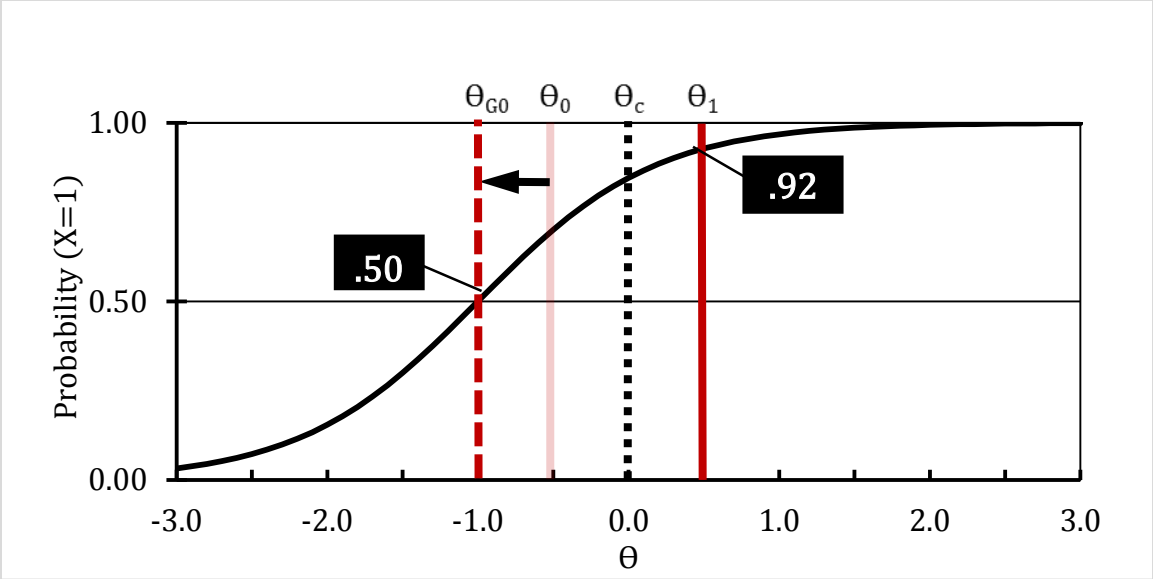


Figure 8: Probabilities of a correct response corresponding with indifference region boundaries of the GLR where the item difficulty parameter, *b* = -1.0, is below the indifference region.

Unintentionally though, the GLR may also include irrelevant error by incorporating into the likelihood ratio calculation portions of the ability distribution which would be irrelevant given the correct/incorrect response of examinee. For example, Figure 9 illustrates the problem which arises when a correct response to a question where the item difficulty is greater than the upper boundary of the indifference region producing a penalty of sorts. The use of the more discrepant values produced by using the GLR indifference region boundaries has a greater negative impact on the examinee's score than if the TSPRT procedure had been used. The examinee has essentially been penalized

for responding incorrectly to an item which is more difficult than the cutscore being used to classify the examinee.

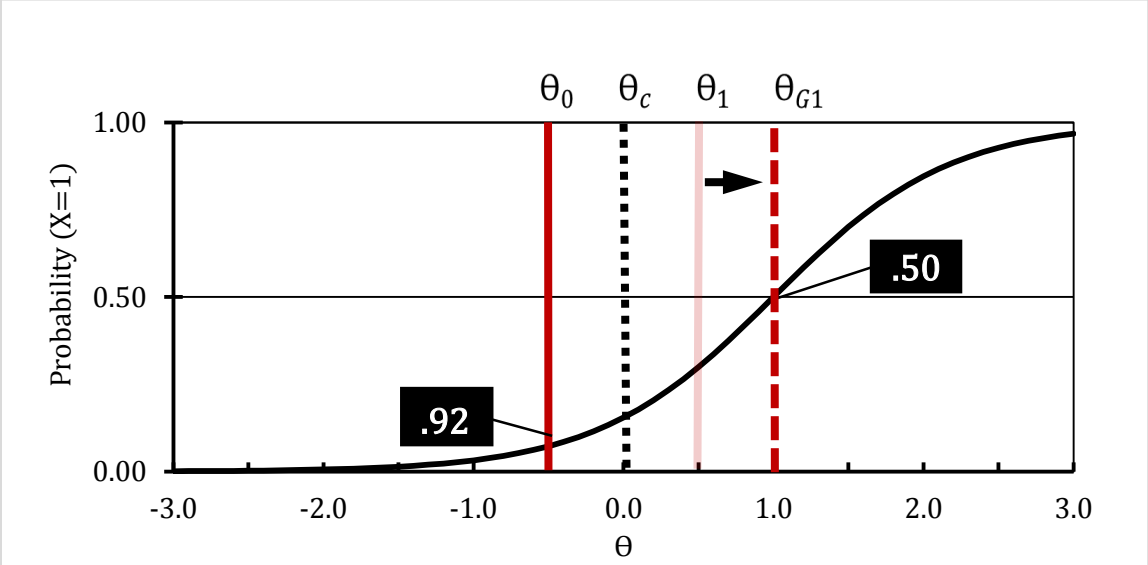


Figure 9: Probabilities of an incorrect response corresponding with indifference region boundaries of the GLR where the item difficulty parameter, *b* = 1.0, is above the indifference region.

Figure 10 illustrates the problem which arises when an incorrect response to a question where the item difficulty is below the lower boundary of the indifference region resulting in a bonus to the examinee's score. The use of the more discrepant values produced by using the GLR indifference region boundaries has a greater positive impact on the examinee's score than if the TSPRT procedure had been used. The examinee has

essentially been given a bonus for responding correctly to an item which is less difficult than the cutscore being used to classify the examinee.

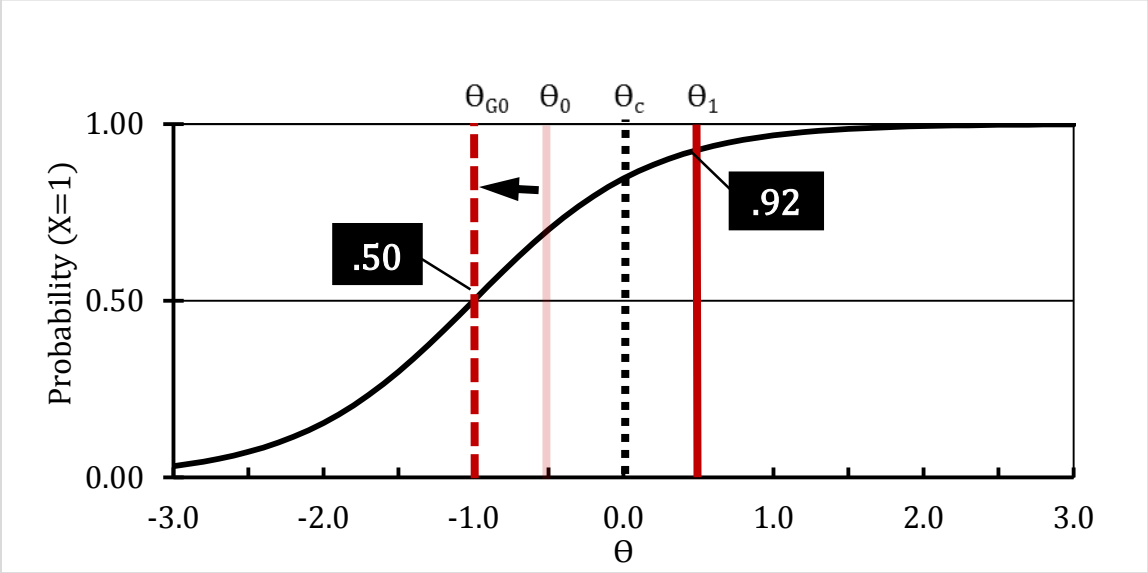


Figure 10: Probabilities of a correct response corresponding with indifference region boundaries of the GLR where the item difficulty parameter, *b* = -1.0, is below the indifference region.

The new procedure being developed in this dissertation, a modification to the generalized likelihood ratio test, ameliorates this problem by only conditionally incorporating the portions of the ability distribution which are outside of the indifference region when the examinee response suggests that the additional information is useful.

Modified Generalized Likelihood Ratio

In the current research, one modification was proposed to the generalized likelihood ratio procedure to continue incorporating the efficiency of the GLR without including the possible additional error. To maintain the efficiency of the GLR, the newly proposed modified-GLR would function similar to the GLR procedure by incorporating the maximum of the likelihood function in instances where it does not lie within the indifference region. Unlike the GLR, the modified-GLR only incorporates the maximum of the likelihood function conditional on whether the examinee answered the item correctly or incorrectly and whether the maximum of the likelihood function were above or below the region. When the maximum of the likelihood value is above the upper boundary of the indifference region the modified-GLR will only utilize the maximum of the likelihood when the examinee selects the correct response. Likewise, the modified-GLR will only utilize the maximum value when it is below the indifference region when the examinee selects an incorrect response. Therefore, the equation for the mGLR is

$$LR = \frac{L(\theta = \theta_{m1})}{L(\theta = \theta_{m0})} = \frac{P(x_1 = 1 | \theta_{m1})P(x_2 = 0 | \theta_{m1})...P(x_n = 1 | \theta_{m1})}{P(x_1 = 1 | \theta_{m0})P(x_2 = 0 | \theta_{m0})...P(x_n = 1 | \theta_{m0})} \quad (22)$$

where θ_{m1} and θ_{m0} represent the conditional boundaries associated with likelihood ratio test. The purpose of proposing the modified generalized likelihood ratio test is two-fold: 1) to include the advantage of the GLR without the major drawback of also including

extraneous error into the likelihood ratio calculation and 2) utilize of ability-based item selection procedures. Figures 11 shows the cutscore, θ_C, the two standard indifference region boundaries, θ_0 and θ_1, and the mGLR indifference region boundaries, θ_{m0} and θ_{m1}.

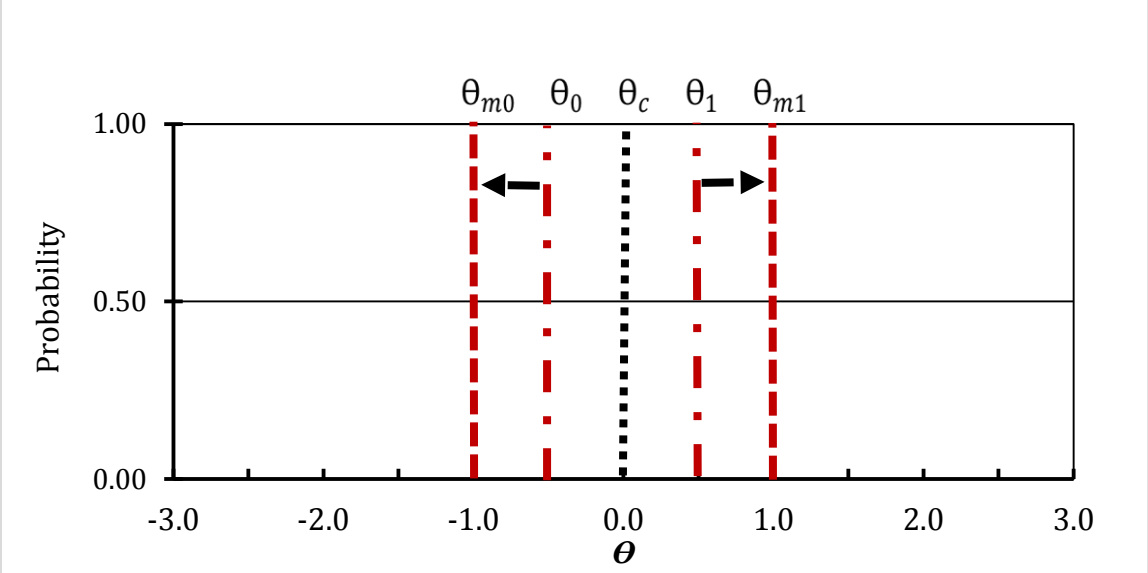


Figure 11: Illustration of cutscore and indifference region boundaries used with the mGLR procedure.

Figure 12 displays the indifference region boundaries with their corresponding probabilities when given a correct response. For the TSPRT procedure, the likelihood ratio would incorporate the 0.08 and 0.30 values. The GLR and mGLR would incorporate the 0.08 and 0.50 probabilities when calculating the likelihood ratio. Because the

examinee correctly answered an item which is more difficult than the cutscore, the examinee's score is increased more than if the TSPRT boundaries were used.

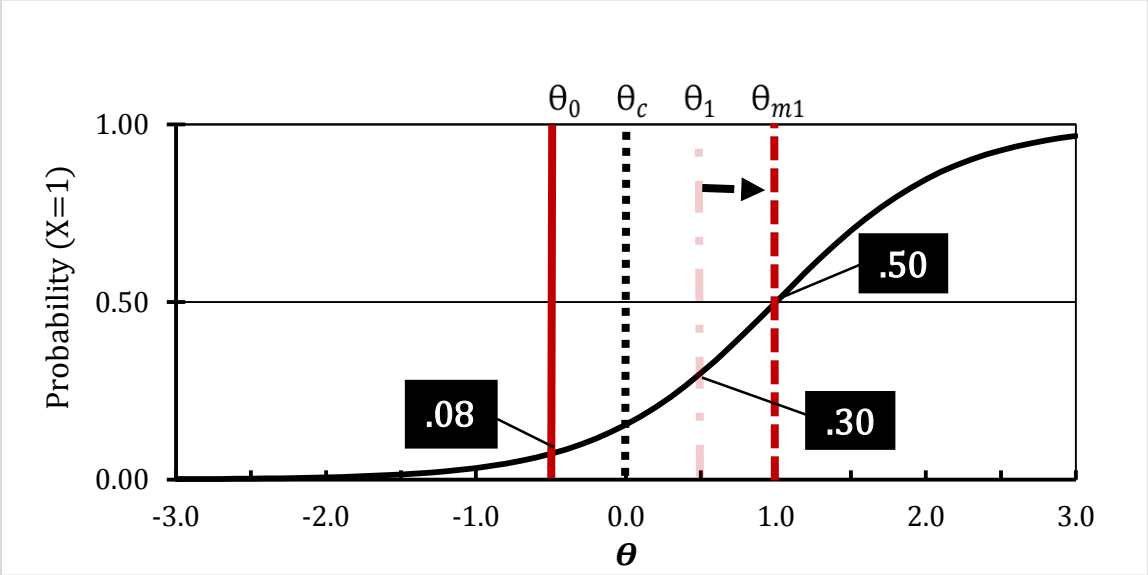


Figure 12: Probabilities of a correct response corresponding with indifference region boundaries of the mGLR where the item difficulty parameter, b = 1.0, is above the indifference region.

Figure 13 displays the indifference region boundaries with their corresponding probabilities when an examinee has incorrectly answered the item. For both the TSPRT and mGLR procedures, the likelihood ratios would incorporate the 0.92 and 0.70 values. The GLR procedure would incorporate the 0.92 and 0.50 probabilities when calculating the likelihood ratio. Because the examinee incorrectly answered an item which is less difficult than the cutscore, the examinee's score is decreased less than if the GLR

boundaries were used. Essentially a penalty has not been included in the score as the item difficult was greater than the indifference region.

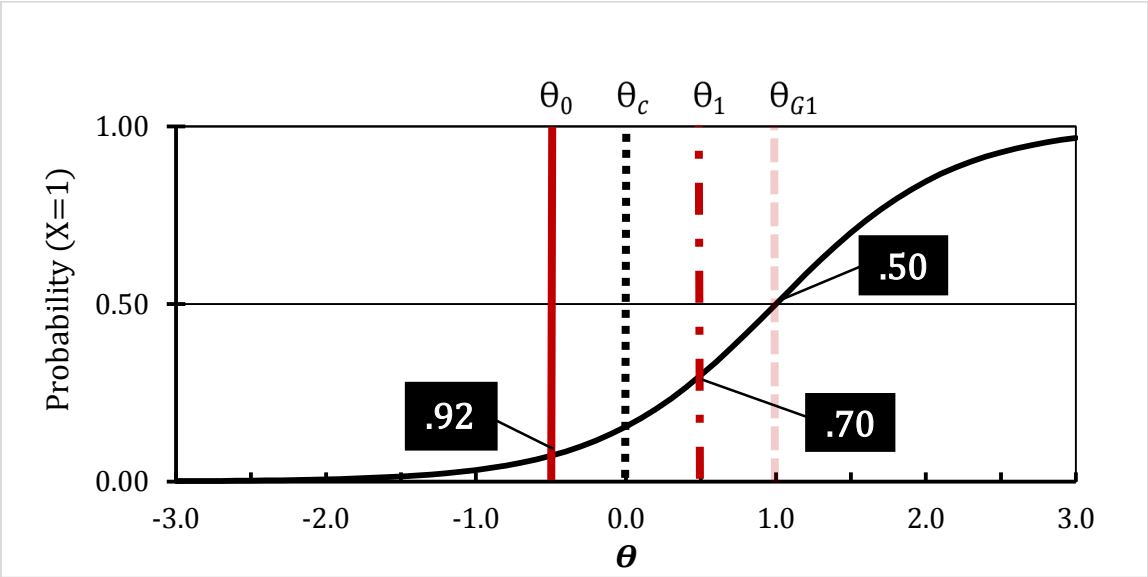


Figure 13: Probabilities of an incorrect response corresponding with indifference region boundaries of the mGLR where the item difficulty parameter, b = 1.0, is above the indifference region.

To demonstrate the possible efficiency and unnecessary error produced by the TSPRT, GLR, and mGLR, Table 1 provides an example of the calculation of the likelihood ratios tests using a single response for an item where the difficulty lies above the indifference region.

Table 1: Likelihood ratio test calculations for the TSPRT, GLR, and mGLR classification methods using a single item where b = 1.0, θ_0= 0.40, and θ_1= 0.60.

Response	Procedure	Likelihood ratio
correct	TSPRT	$LR = \frac{0.30}{0.25} = 1.2$
correct	GLR	$LR = \frac{0.50}{0.25} = 2.0 *$
correct	mGLR	$LR = \frac{0.50}{0.25} = 2.0 *$
incorrect	TSPRT	$LR = \frac{0.70}{0.75} = 0.93$
incorrect	GLR	$LR = \frac{0.50}{0.75} = 0.67 **$
incorrect	mGLR	$LR = \frac{0.70}{0.75} = 0.93$

* Denotes the efficiency produced by incorporating the maximum of the likelihood function. **Denotes the error produced by utilizing the maximum of the likelihood indiscriminately.

When a correct response is given the GLR and mGLR achieve higher ratio values by incorporating a larger portion of the ability distribution and thereby larger discrepancies between the probabilities used in the calculation of the likelihood values.

This increase in the likelihood ratio produced by the GLR and mGLR can be sensibly accepted in the score as the examinee correctly answered an item which was more difficult than the cutscore. We may reasonably assume the examinee's ability lies somewhere about or above the difficulty of the item and thereby use the probability value from the point of inflection in the likelihood ratio calculation. This advantage would help shorten test lengths when the examinee's ability is above the cutscore.

When an incorrect response is observed for the same item, the TSPRT and mGLR produce ratio values higher than the GLR because the probabilities for θ_0 and θ_1 are less discrepant. Yet the GLR still uses needlessly as the point is outside and above the indifference region while the examinee's response is incorrect. Extraneous error is incorporated into the likelihood ratio by utilizing the maximum of the likelihood function resulting in more discrepant probabilities in the likelihood calculation. The extraneous error may also be viewed as a penalty for incorrectly responding to an item which has a difficulty that is higher than the cutscore being used for classification.

The opposite occurs when the item difficulty is below the lower boundary of the indifference region. Figure 14 displays the indifference region boundaries with their corresponding probabilities when given an incorrect response. For the TSPRT procedure, the likelihood ratio would incorporate the 0.08 and 0.30 values. The GLR and mGLR would incorporate the 0.08 and 0.50 probabilities when calculating the likelihood ratio. Because the examinee incorrectly answered an item which is less difficult than the

cutscore, the examinee's score is decreased more than if the TSPRT boundaries were used.

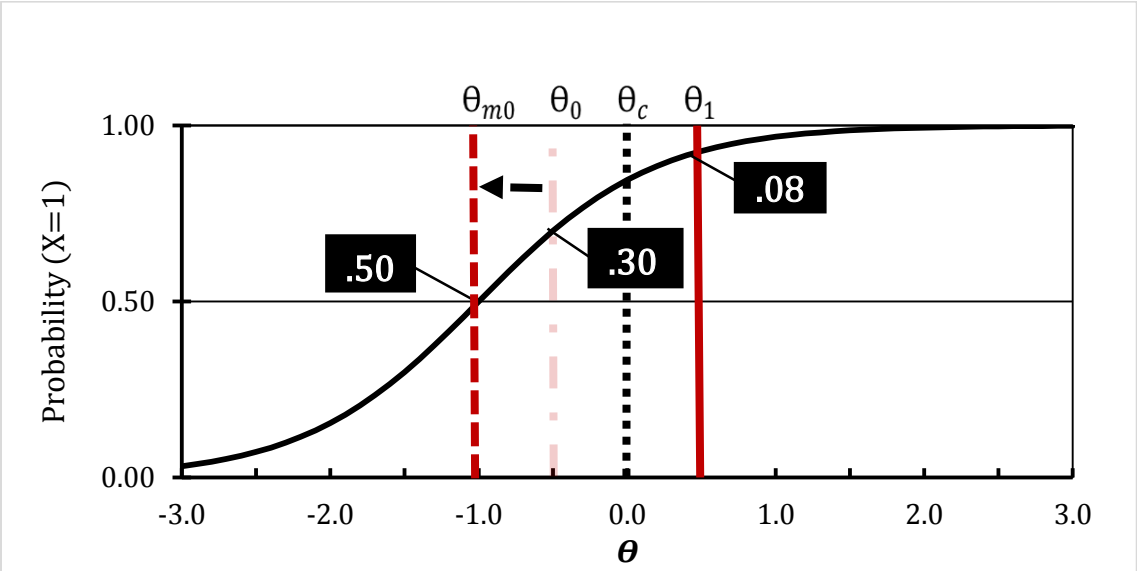


Figure 14: Probabilities of an incorrect response corresponding with indifference region boundaries of the mGLR where the item difficulty parameter, *b* = -1.0, is below the indifference region.

Figure 15 displays the indifference region boundaries with their corresponding probabilities when an examinee has correctly answered the item. For both the TSPRT and mGLR procedures, the likelihood ratios would incorporate the 0.70 and 0.92 values. The GLR procedure would incorporate the 0.50 and 0.92 probabilities when calculating the likelihood ratio. Because the examinee correctly answered an item which is less difficult than the cutscore, the examinee's score is increased less than if the GLR boundaries were

used. By not using the GLR boundaries, a bonus has not been included in the score as the item difficult was lower than the indifference region.

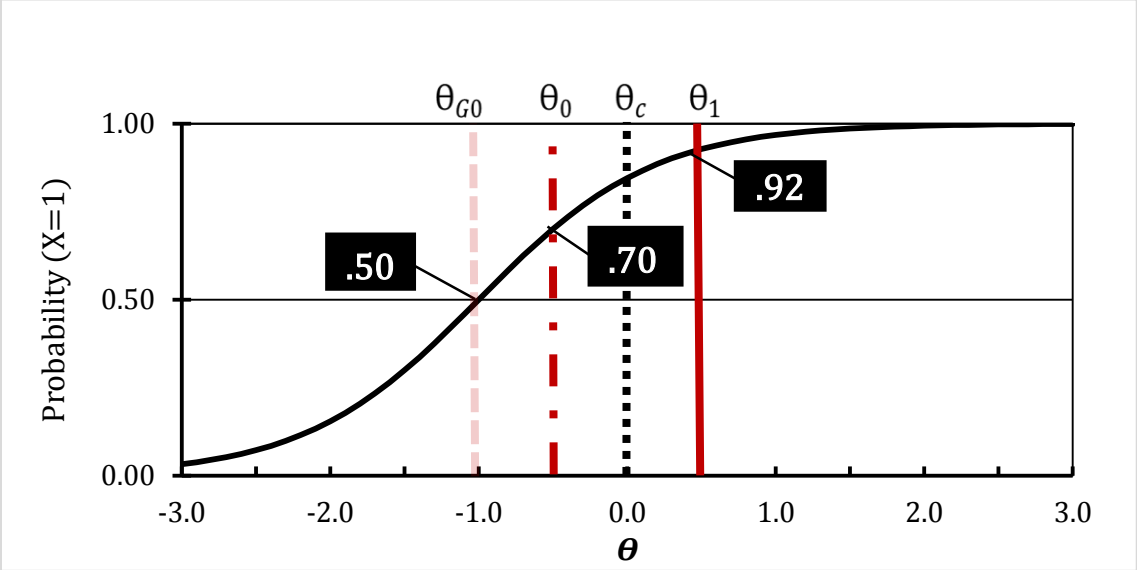


Figure 15: Probabilities of a correct response corresponding with indifference region boundaries of the mGLR where the item difficulty parameter, *b* = -1.0, is below the indifference region.

Table 2 provides an example of the calculation of the likelihood ratio test for the TSPRT, GLR, and mGLR using a single response for an item where the difficulty lies below the indifference region.

Table 2: Likelihood ratio test calculations for the TSPRT, GLR, and mGLR classification methods using a single item where b = -0.50, θ_0= 0.40, and θ_1= 0.60.

Response	Procedure	Likelihood ratio
correct	TSPRT	$LR = \frac{0.90}{0.80} = 1.13$
correct	GLR	$LR = \frac{0.90}{0.50} = 1.80 **$
correct	mGLR	$LR = \frac{0.90}{0.80} = 1.13$
incorrect	TSPRT	$LR = \frac{0.10}{0.20} = 0.50$
incorrect	GLR	$LR = \frac{0.10}{0.50} = 0.20 *$
incorrect	mGLR	$LR = \frac{0.10}{0..50} = 0.20 *$

* Denotes the efficiency produced by incorporating the maximum of the likelihood function. **Denotes the error produced by utilizing the maximum of the likelihood indiscriminately.

When a correct response is observed for an item with a difficulty below the indifference region, the TSPRT and mGLR procedures use the likelihood values from θ_0 and θ_1 to assess the performance of the examinee relative to the area about the cutscore.

The GLR artificially inflates the value from the likelihood ratio by incorporating the maximum of the likelihood thereby using more discrepant probabilities in the likelihood ratio calculation. The inflation of the score is not particularly useful given the item difficulty is below the cutscore which essentially adds a bonus to the examinee's score for answering easier items correctly. Yet when an incorrect response is observed the GLR and mGLR procedures produce lower scores by incorporating the maximum of the likelihood and thereby including the portion of the ability distribution where it is reasonable to assume the examinee's ability lies. The resulting probabilities are more discrepant rendering the item more efficient even though the item difficulty is not within the indifference region. This advantage would help shorten tests when the examinee's ability is below the cutscore.

As additional items are administered under the GLR procedure the resulting benefits and errors will be compounded. Thus while the GLR has recently been proposed as a means of providing the opportunity for shorter tests while being comparable to the TSPRT procedure (Huang, 2004; Bartroff, Finkelman, & Lai, 2008; Thompson, 2007, 2009), the current study sought to evaluate and provide evidence that the modified-GLR is a better means by producing similarly shortened tests without incorporating the extraneous error.

When multiple cutscores are used in a testing procedure each cutscore uses a separate set of indifference region boundaries. A likelihood ratio is calculated for each

classification decision point for each item that administered in the test. The cutscores and indifference regions shown in Figure 16 are symmetrical and equivalent for purposes of simplicity in the example.

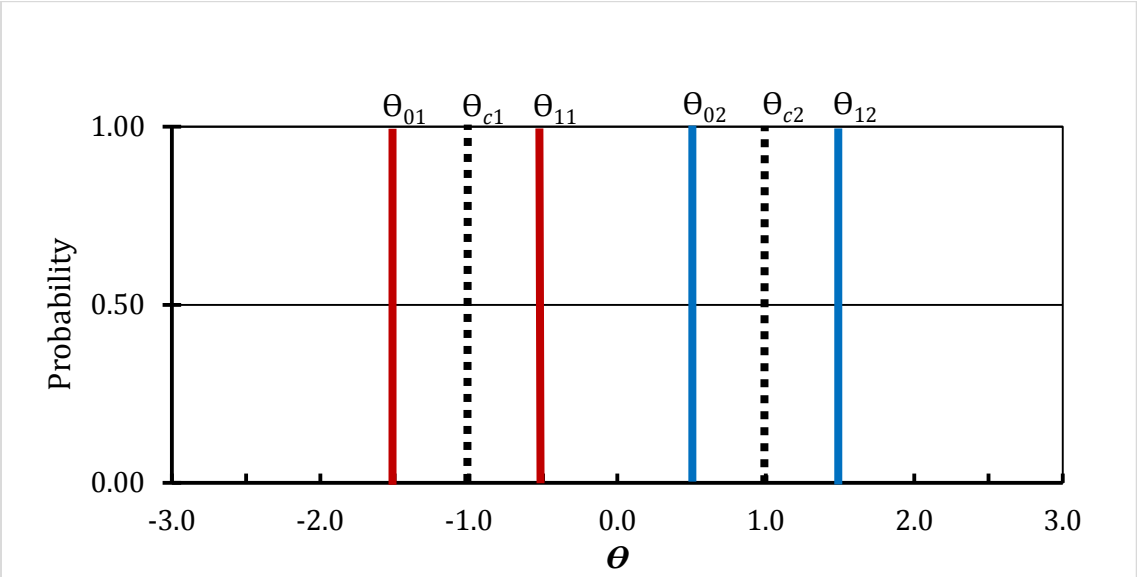


Figure 16: Two cutscores with accompanying indifference region boundaries.

Figure 17 displays the various probabilities associated with the indifference region boundaries for both of the cutscores when the item difficulty is between the two indifference regions. For the likelihood ratio for the lower cutscore, θ_{C1}, the upper boundary of the indifference region has been adjusted to the theta value corresponding with the maximum of the likelihood function, θ=0.0. Neither of the boundaries of the indifference region for the likelihood ratio for the upper cutscore are adjusted because the item difficulty is below the lower boundary of the indifference region.

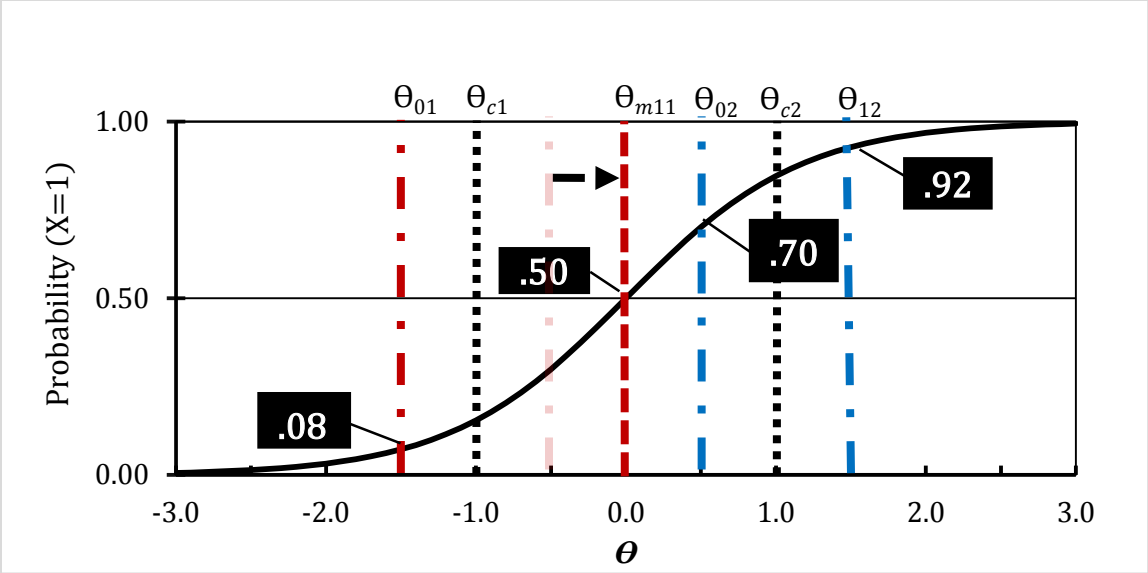


Figure 17: Probabilities of a correct response corresponding with indifference region boundaries of the mGLR where the item difficulty parameter, b = 0.0, is between the two indifference regions.

Figure 18 displays the various probabilities associated with the indifference region boundaries for both of the cutscores using the same item depicted in Figure 17. Given an incorrect response, neither of the boundaries of the indifference region of the lower cutscore change as the item difficulty is above the indifference region. The lower boundary of the higher cutscore, does adjust down to the theta value corresponding with the maximum of the likelihood function.

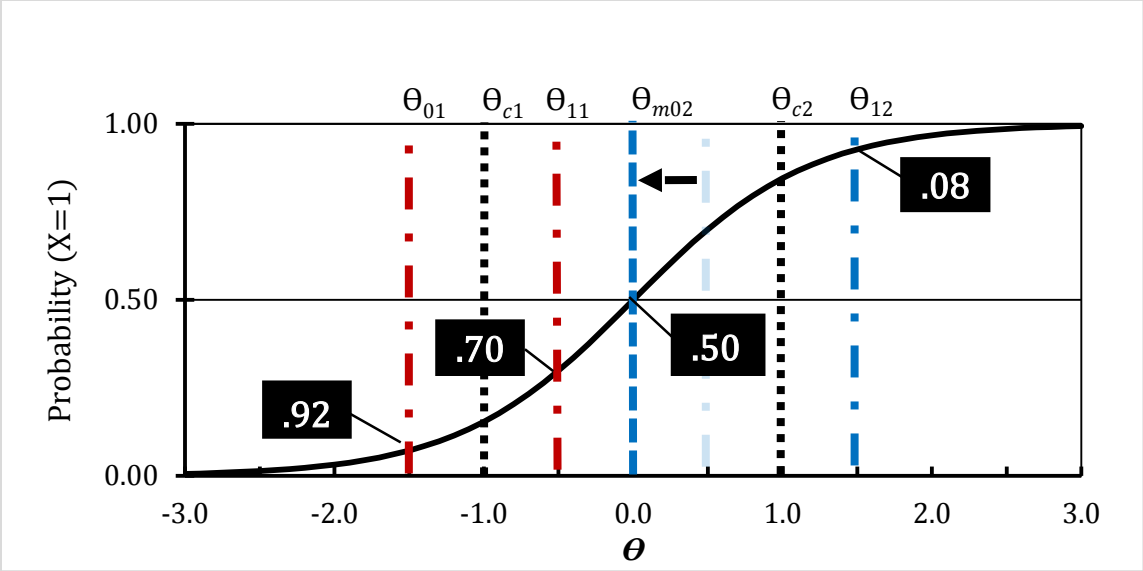


Figure 18: Probabilities of an incorrect response corresponding with indifference region boundaries of the mGLR where the item difficulty parameter, b = 0.0, is between the two indifference regions.

The second purpose of proposing the mGLR procedure is to be able to use ability-based item selection procedures which would then allow for an ability estimate to be determined for examinees and to enable test lengths to be shortened while maintaining a high degree of classification accuracy. As previously discussed, classification procedures based on TSPRT become inefficient when ability estimates are used to select items. The inefficiency of the procedures is due to the inflexibility of the indifference region boundaries. The GLR procedure could work more efficiently by switching from a cutscore-based item selection method to selecting items based on the examinee's interim

ability estimates, but as discussed, the GLR method includes extraneous error. Therefore, the mGLR procedure was developed to provide a flexible TSPRT-based procedure which could utilize ability estimates to select items while still achieving an accurate classification decision more rapidly than the typical ability estimate based classification procedures (e.g. ACI).

Additionally the use of ability-based item selection enables the mGLR to provide an estimate of the examinee's ability. The ability estimate allows examinees and other stakeholders the opportunity to examine how far above or below the examinee's ability is estimated to be. This information could be useful in tracking examinee progress across time. In other circumstances, stakeholders such as potential employers would also be able to rank candidates based on their ability estimates.

Statement of Problem

Over the last two decades studies have investigated various aspects of utilizing TSPRT procedures in educational testing including item selection methods, the number of cutscores used, variations in indifference regions, IRT models (dichotomous/polytomous/mixed models), and the implementation of TSPRT in computerized adaptive testing. Recently variants of the TSPRT procedure have been proposed such as the GLR procedure. Procedurally, TSPRT and GLR are highly similar and thereby it is reasonable that some of what is known about TSPRT would generalize

to GLR as well, but little is known about the extent to which GLR is more efficient and accurate when the procedures are used in CATs when considering multiple cutscores. Additionally, the currently proposed modified-GLR procedure was used in a CAT setting to provide for comparison with the GLR procedure.

To date, research on TSPRT-type and GLR procedures had been relegated to cutscore-based item selection methods. The proposed modified-GLR procedure was developed with the intent to enable a more efficient classification process including the use of ability-based item selection. As a result, the procedure is anticipated to maintain the high classification accuracy of TSPRT procedure while reducing the number of items that are used to classify the examinees. By adopting the ability-based item selection method with the modified-GLR procedure, more efficient items can be selected for administration as well as providing for a final ability level that can be estimated for each examinee.

This dissertation was designed to enable comparison between the classification accuracy rates and average test length across three classification testing procedures, namely TSPRT, GLR, and modified-GLR. Differences between accuracy rates and test length were examined within the context of variations in number of cutscores and the maximum number of items. The modified-GLR procedure was also examined when using two item selection procedures for differences between accuracy rates, average test

lengths, and to examine the recovery of ability estimates. Specifically, this dissertation was developed to examine the following research questions:

1. How do the three classification testing procedures, TSPRT, GLR, and mGLR, using cutscore-based item selection compare to each other in terms of average test length and percent correct classification in the context of multiple cutscore and test length conditions?
2. How does the implementation of an ability-based item selection method with the mGLR procedure compare with the cutscore-based item selection mGLR procedure in terms of average test length and percent correct classification?
3. How well can ability levels be recovered as assessed using bias and root mean square error when an ability-based item selection method is implemented with the mGLR procedure?

To examine these questions, single-cutscore, two-cutscore, and three-cutscore conditions have been simulated for each classification testing method using two different maximum test lengths. Comparisons have been made across classification procedures using the same number of cutscores. Additionally, bias and root mean squared error were calculated to examine the recovery of the ability estimate parameter in mGLR conditions using the ability-based item selection method.

CHAPTER III: METHODOLOGY

Design Overview

A simulation study was performed to evaluate the efficiency and effectiveness of multiple termination procedures in the context of computerized adaptive classification testing. The three classification procedures—the truncated sequential probability ratio test, the generalized likelihood ratio test, and the modified generalized likelihood ratio test—were implemented using items calibrated according to the 3-PL IRT model. The study design was a 3 (classification procedure) x 3 (number of cutscores) x 2 (test length) design yielding 18 conditions. An additional 6 conditions were also included in the study in which the modified-GLR procedure was implemented using an ability-based item selection procedure. Parallel to the cutscore-based item selection conditions, the mGLR with ability-based item selection was examined using variations in the number of cutscores and test lengths. The complete study yielded a total of 24 conditions—18 conditions were simulated using the traditional SPRT-type item selection method which selects items to maximize information at the cutscore while 6 conditions using the mGLR method were simulated using the traditional CAT item selection method which selects items to maximize information at the current ability estimate. It is important to note that comparisons could only be made between classification procedures with the same number of cutscores and with the same maximum number of items. Additionally only the

mGLR procedure was simulated using the ability-based item selection method as previous research has demonstrated the inefficiency of TSPRT procedure when using this item selection method (Spray & Reckase, 1994; Thompson N. A., 2007, 2009). Due to the similarity of the rigid indifference region parameters of the TSPRT, no simulation using ability estimate-based item selection with the GLR was performed. Therefore, only the results from the mGLR with cutscore-based item selection was compared with the results from the mGLR with ability-based item selection.

Item Pool

The item pool used in this dissertation is taken from a national test consisting of 540 multiple choice items calibrated with the 3-PL model. The item pool contains items from six content domains. Table 3 presents descriptive statistics for the item parameter estimates by content domain.

Table 3: IRT item statistics for item pool.

Content Domain		a parameter	b parameter	c parameter
Content I n=127	Mean (SD)	0.839 (0.242)	-0.544 (1.093)	0.204 (0.080)
	Min	0.266	-3.106	0.066
	Max	1.490	5.543	0.489
Content II n=89	Mean (SD)	1.010 (0.302)	0.030 (0.937)	0.190 (0.066)
	Min	0.561	-2.114	0.065
	Max	1.783	2.198	0.352
Content III n=81	Mean (SD)	1.120 (0.322)	0.411 (0.929)	0.187 (0.072)
	Min	0.449	-2.125	0.039
	Max	2.149	3.277	0.382
Content IV n=81	Mean (SD)	1.105 (0.324)	0.552 (0.879)	0.189 (0.086)
	Min	0.465	-2.362	0.056
	Max	1.838	2.807	0.500
Content V n=126	Mean (SD)	1.050 (0.319)	0.552 (0.879)	0.19 (0.087)
	Min	0.481	-2.171	0.058
	Max	1.186	2.428	0.447
Content VI n=36	Mean (SD)	1.286 (0.357)	1.024 (0.745)	0.183 (0.068)
	Min	0.720	-0.193	0.054
	Max	2.317	2.264	0.298
TOTAL n=540	Mean (SD)	1.028 (0.327)	0.161 (1.048)	0.192 (0.079)
	Min	0.266	-3.106	0.039
	Max	2.137	5.543	0.500

Data Generation

A single simulation data set was generated from a uniform distribution ranging from -3.0 to 3.0. A uniform distribution was selected for use in this study to ensure a sufficient number of simulees in the extreme regions of the ability distribution so that classification accuracy, test length, and ability level recovery can be examined across the full range of ability. The same dataset was used in all CAT simulation conditions to enable comparisons to be made between procedures. For the single data set 1,000 simulees with accompanying response strings were generated for each theta value in discrete 0.10 logit increments resulting in a total of 61,000 simulees.

Responses to all 540 items for each individual examinee were generated using the dichotomous 3-PL IRT model. To simulate examinee responses for each of the 24 conditions, an ability level was assigned to each examinee based on a uniform distribution ranging from -3.0 to 3.0—this ability level value will be referred to as the simulated ability level. The probability of responding correctly was calculated for each simulee based on their simulated ability level and the item parameters for the 540 items. Item responses were generated by comparing the probability of a correct response to a random number drawn from a uniform distribution with a minimum of 0 and a maximum of 1. Probability values greater than the randomly drawn number were recorded as correct responses (1) while probability values less than the randomly drawn number were recorded as incorrect responses (0). This procedure was repeated for all items and

simulees until item responses had been generated to build a single simulated response data set. The SAS macro program IRTGEN was used to create the data set (Whittaker, Fitzpatrick, Williams, & Dodd, 2003) by utilizing simulated ability levels based on a uniform distribution and the item parameters from the 540 items into the SAS macro program.

CAT Simulations

The CAT simulations were performed using SAS computer programs that were written to utilize the cutscore-based item selection method proposed by Eggen and Straetsman (2000) for the multiple-cutscore conditions using the TSPRT, GLR, and modified-GLR procedures. An additional CAT simulation program was written for SAS wherein the modified-GLR procedure was designed to operate using an ability-based item selection method. All CAT simulations were programmed to incorporate the same Randomesque exposure control and content balancing constraints. The Randomesque procedure was programmed to draw five items for possible administration each time the simulees was to be administered another item. Content balancing was designed to select items to be proportional to the content category as they were present in the item pool. For conditions using cutscore-based item selection methods where a single cutscore was being used, the initial item selected for administration was selected to maximize information at the cutscore. For conditions using cutscore-based item selection where two cutscores were to be used, the initial item selected for administration was selected so that

the item information was maximized at the midpoint between the two cutscore theta levels. In the study conditions using cutscore-based item selection where three cutscores were to be used, the initial item selected for administration was selected to maximize information at the middle cutscore, the same cutscore used in the single-cutscore conditions. As is common practice for typical CAT simulations, the conditions using the modified-GLR with ability-based item selection were programmed to select the initial item which maximized the information at the mean of the ability distribution.

Type I and Type II error rates were defined by α and β, respectively. Following previous research (Parshall et al., 2002; Lin, 2010), for all single cutscore conditions $\alpha = \beta = .05$ resulting in lower- and upper-bound decision values of A=19, B=.052632, ln$A$=2.944, and ln$B$=-2.944. For conditions in which there were two cutscores $\alpha_0 = \alpha_1 = \beta_0 = \beta_1 = .05$ resulting in identical values for each parameter as described for the single cutscore conditions.

Classification Testing Procedures

Three test termination procedures, TSPRT, GLR, and modified-GLR, have been examined to evaluate the effectiveness and efficiency of the 24 conditions in the proposed study. The TSPRT procedure has been considered a baseline condition as both the GLR and modified-GLR function very similarly but should have been no less advantageous in regards to both average test length and percent correct classification.

Because the procedures are very similar in functionality, a common set of test parameters were used throughout the study to ensure comparability across conditions.

Test conditions in which a single cutscore were used to classify examines utilized a cutscore, θ_c, located at the peak of the test information function. Conditions which implemented two cutscores, θ_{c1} and θ_{c2}, to classify examinees into one of three classifications had cutscores placed one-half standard deviation below and above the peak of the test information function. For the test conditions implementing three cutscores, the lowest cutscore, θ_{c1}, was placed one-half standard deviation below the peak of the test information function, the middle cutscore, θ_{c2}, was placed at the peak of the test information function, and the highest cutscore, θ_{c3}, was placed one-half standard deviation above the peak of the test information function. The locations of the cutscores were determined so that indifference region boundaries for the TSPRT conditions would not overlap. Previous research has commonly used indifference widths ranging from 0.00 to 0.50 in increments ranging from 0.02 to 0.10 (Spray J. A., 1993; Spray & Reckase, 1994; Thompson, 2009).The width of the indifference region, δ, will be fixed for all conditions and be symmetrical with $\delta = 0.20$.

Termination of CATs. Individual tests were terminated when the *A* or *B* boundary test parameters had been surpassed by the value of the likelihood ratio or when the maximum test length had been reached. A maximum number of items per test was set based on the condition, 40 or 60. Examinees which were not classified before reaching

the maximum number of items were classified based on the final value of the likelihood ratio being compared to the cutscore theta, θ_c.

Item Selection Method

The cutscore-based item selection method was implemented with the TSPRT, GLR, and mGLR classification methods. Items were selected to provide maximum information at the cutscore for the single cutscore conditions. For the test conditions with more than one cutscore, the method proposed by Eggen and Straetsman (2000) which selects items to maximize the likelihood ratio of the cutscore nearest the examinee's ability was implemented. Fisher's information was used for the conditions using the ability-based item selection method for the modified-GLR procedure. The aforementioned Randomesque procedure (Kingsbury & Zara, 1989) using a set of 5 items was incorporated into the item selection methods to control item exposure in all conditions. Additionally, the constrained CAT (CCAT) developed by Kingsbury and Zara (1989) was used to provide content balancing.

Number of Cutscores

Ideally, item pools for classification testing would be developed so that item difficulties would match the intended cutscore. As this study used a previously developed item pool, a cutscore was selected for use based on the characteristics of the existing items. For the single cutscore conditions the cutscore was placed at the peak of the test

information function, θ_c = 1.0. Conditions which implemented two cutscores had a cutscores placed 0.50 standard deviations below and above the mean item difficulty for the item pool. The lower cutscore was θ_{c1} = 0.50 and the upper cutscore was θ_{c2} = 1.50. For the conditions which utilized three cutscores, the three aforementioned cutscores were combined into a single procedure. The lowest cutscore was θ_{c1} = 0.50, the middle cutscore was θ_{c2} = 1.00, and highest cutscore was θ_{c3} = 1.50.

Test Length

The maximum test lengths, 40 and 60 items, were selected based on the length of the original test form, as well as previous research using likelihood ratio-based procedures (Spray & Reckase, 1994; Eggen & Straetmans, 1996; Spray & Reckase, 1996; Lau & Wang, 2000; Finkelman, 2008, 2009; Wouda & Eggen, 2009). Selected test lengths also ensure that, given the constraints of the exposure control procedure and content balancing, items would be selected for administration from content domains proportional to the 6 content domains in the item pool. Consideration was also given to the *SE* of the ability estimates produced by the mGLR procedure using the ability-based item selection, thus a longer test, 60 item test, will enable a comparison to the *SE* of the ability estimates to the shorter test length. A minimum test length of 20 items was selected based on previous research (Thompson, 2010, 2011).

Data Analyses

To evaluate the performance of the procedures the average test length (ATL) and the percent of correct classification (PCC) was compared across comparable conditions. Both outcome measures, ATL and PCC, were calculated conditional on the simulated theta values in increments of 0.10 for each study condition. These two methods of evaluation are consistent with previous methods (Finkelman, 2008, 2010; Parshall et al., 2002; Spray, 1993, Spray & Reckase, 1994; Wouda & Eggen, 2009).

The evaluation of the ATL and PCC was a comparison of the descriptive statistics for each variable. For each theta value in each of the 24 conditions for the average test length variable a mean and standard deviation, and minimum and maximum test length were calculated. The percent of correct classification was calculated for each theta level in each of the 24 conditions. To examine the recovery of the simulee's known ability level, conditional bias and root mean squared error (RMSE) plots have been developed for the 6 modified GLR conditions using the ability-based item selection method. The following equations were used to calculate Bias and RMSE:

$$Bias = \frac{\sum_{k=1}^{n}\left(\hat{\theta}_k - \theta_k\right)}{n} \tag{23}$$

and

$$RMSE = \left[\frac{\sum_{k=1}^{n}\left(\hat{\theta}_k - \theta_k\right)^2}{n}\right]^{1/2} \tag{24}$$

where $\hat{\theta}_k$ is the estimated ability level of simulee *k* and θ_k is the known ability level of simulee *k*.

CHAPTER IV: RESULTS

This study was designed to compare three classification procedures (TSPRT, GLR, and mGLR) in the context of multiple cutscore conditions (one, two, and three cutscores) using two test lengths (40 and 60 item maximum) in a CAT setting. Additionally, the mGLR procedure was programmed using an ability estimation item selection methodology. The results from the second part of the study are a comparison between the mGLR procedures using cut-based item selection and ability-based item selection. Results of ability estimate parameter recovery from the mGLR procedures using an ability-based item selection method are also discussed.

Cutscore-based Item Selection Procedures

Average Test Length

This study uses three classification procedures and two maximum test lengths in CAT simulations. All of the classification methods compared in this section used a cutscore-based item selection method. Tables 4, 5, and 6 present the conditional means and standard deviations for all three classification procedures using two maximum test lengths. The accompanying plots, Figures 19 through 24, display conditional average test lengths for each procedure based on the number of cutscores and maximum test length. Table 4 provides the conditional means and standard deviations for the TSPRT, GLR,

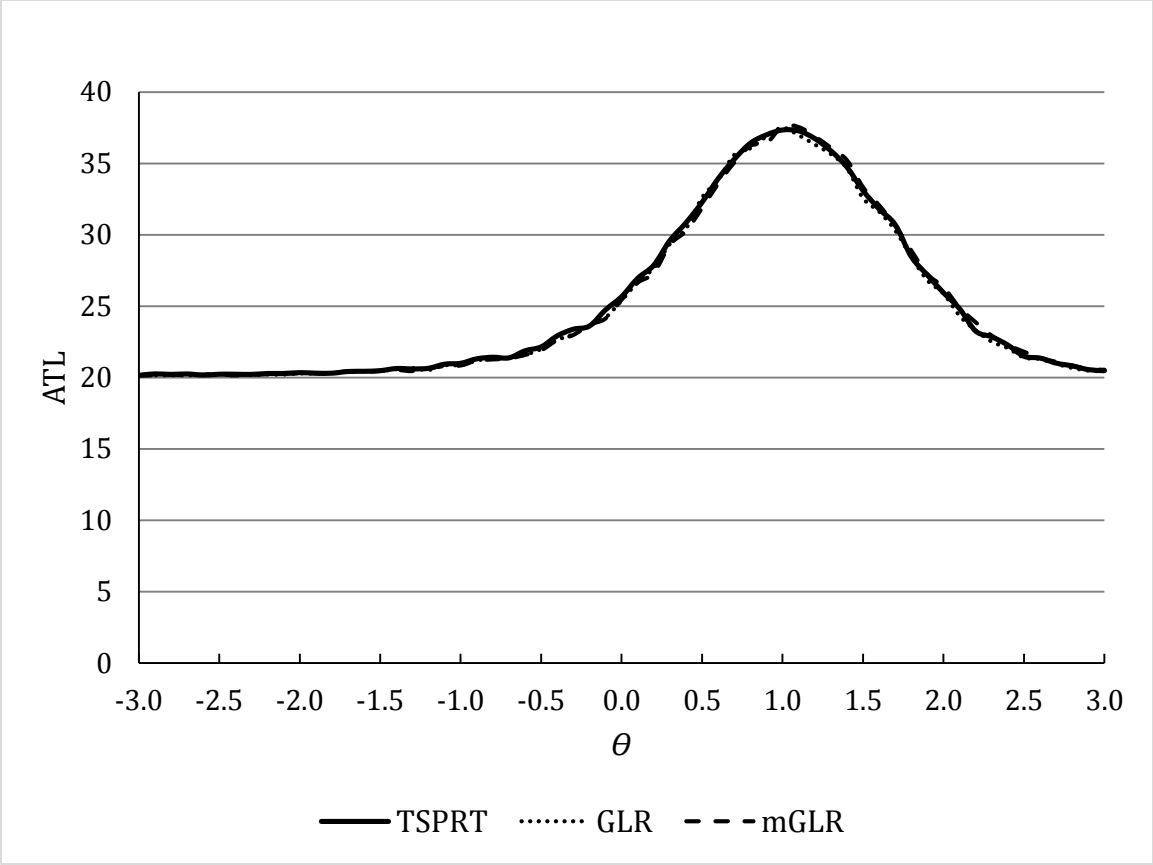
and mGLR procedures when a single cutscore was used. The cutscore for the single-cutscore conditions was placed at the peak of the test information function at the theta value of 1.00.

As evidenced in Table 4 and accompanying conditional plots, Figures 19 and 20, the resulting means and standard deviations for both test length conditions are highly similar. For the conditions with the 40 item maximum test length, ATL range from 37.34 to 37.62. Similarly, for the conditions with the 60 item maximum test length, ATL range from 49.50 to 50.94. Figures 19 presents the conditional ATL plots for the procedures using a single cutscore with a 40 item maximum test length while Figure 20 presents the plots for the 60 item maximum test length.

Table 4: Conditional average test length (ATL) and standard deviation (SD) for the single-cutscore item maximum test length conditions.

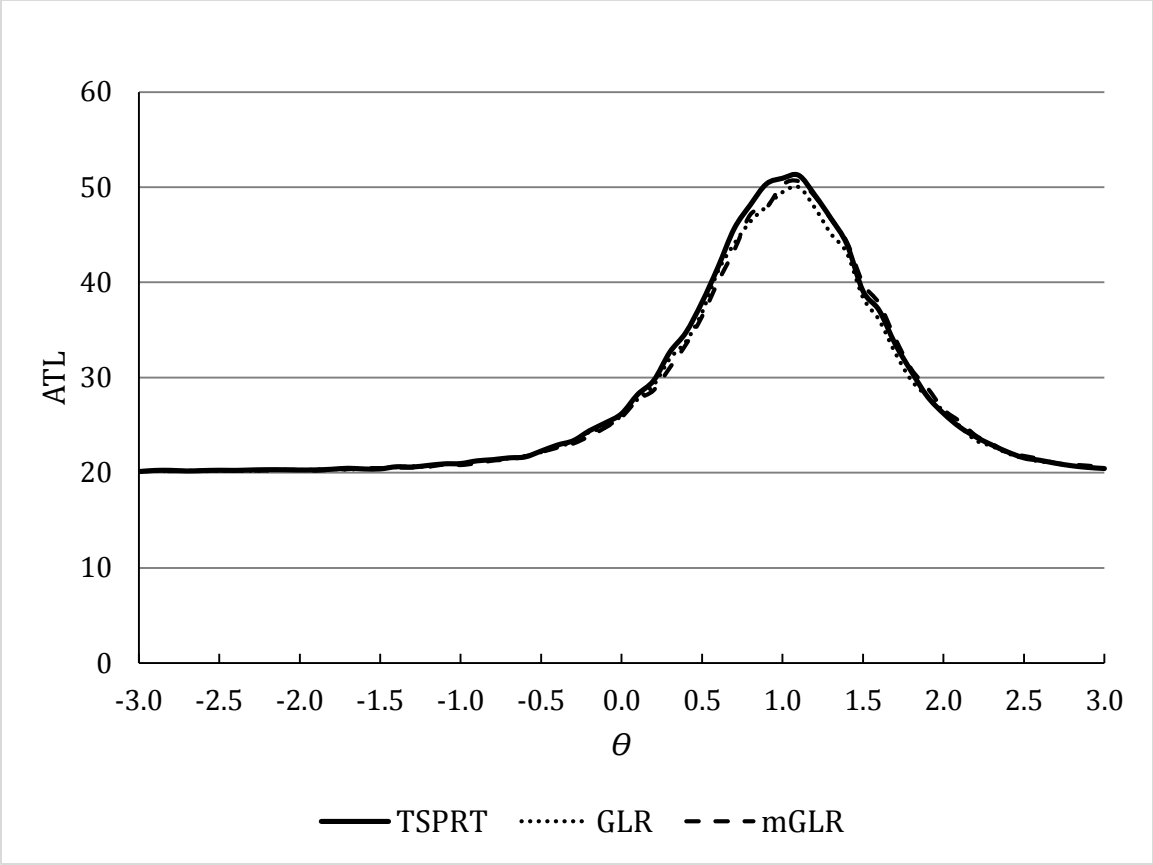
	40 Item Maximum						60 Item Maximum					
	TSPRT		GLR		mGLR		TSPRT		GLR		mGLR	
Theta	ATL	SD	ATL	SD	ATL	SD	ATL	SD	ATL	SD	ATL	SD
-3.0	20.16	0.84	20.15	0.80	20.15	0.85	20.12	0.77	20.12	0.72	20.16	0.81
-2.5	20.23	1.02	20.25	1.14	20.22	1.09	20.26	1.13	20.20	0.92	20.24	1.04
-2.0	20.34	1.50	20.30	1.18	20.29	1.22	20.28	1.15	20.28	1.42	20.27	1.26
-1.5	20.48	1.66	20.47	1.77	20.45	1.60	20.40	1.52	20.47	1.87	20.47	1.69
-1.0	21.00	2.75	20.88	2.45	20.84	2.53	20.96	2.66	20.92	2.55	20.80	2.45
-0.5	22.15	4.17	22.00	3.92	21.90	3.98	22.27	4.98	22.16	4.52	22.16	4.61
0.0	25.67	6.84	25.42	6.83	25.45	6.75	26.21	8.91	26.02	8.53	25.84	8.04
0.5	32.28	7.96	32.61	7.95	31.88	8.02	37.92	14.53	36.85	13.71	36.43	13.59
1.0	37.34	5.60	37.50	5.50	37.62	5.43	50.94	13.12	49.50	13.52	50.27	13.17
1.5	33.10	7.73	32.56	7.81	33.33	7.73	39.08	14.74	38.39	14.45	39.89	14.75
2.0	25.94	6.78	25.88	6.54	26.32	6.86	26.24	8.04	26.47	8.09	26.64	8.46
2.5	21.49	3.19	21.41	2.94	21.80	3.80	21.56	3.17	21.50	3.02	21.76	3.55
3.0	20.48	1.55	20.44	1.52	20.54	1.67	20.44	1.54	20.37	1.21	20.55	1.63

TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.



TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.

Figure 19: Conditional average test length (ATL) for the single-cutscore 40 item maximum test length conditions.



TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.

Figure 20: Conditional average test length (ATL) for the single-cutscore 60 item maximum test length conditions.

Table 5 presents the conditional means and standard deviations for the TSPRT, GLR, and mGLR procedures when two cutscores were used. The lower cutscore for the two-cutscore conditions was placed 0.50 standard deviations below the peak of the test information function at the theta value of 0.50 while the upper cutscore for the two-cutscore conditions was placed 0.50 standard deviations above the peak of the test information function at the theta value of 1.50.

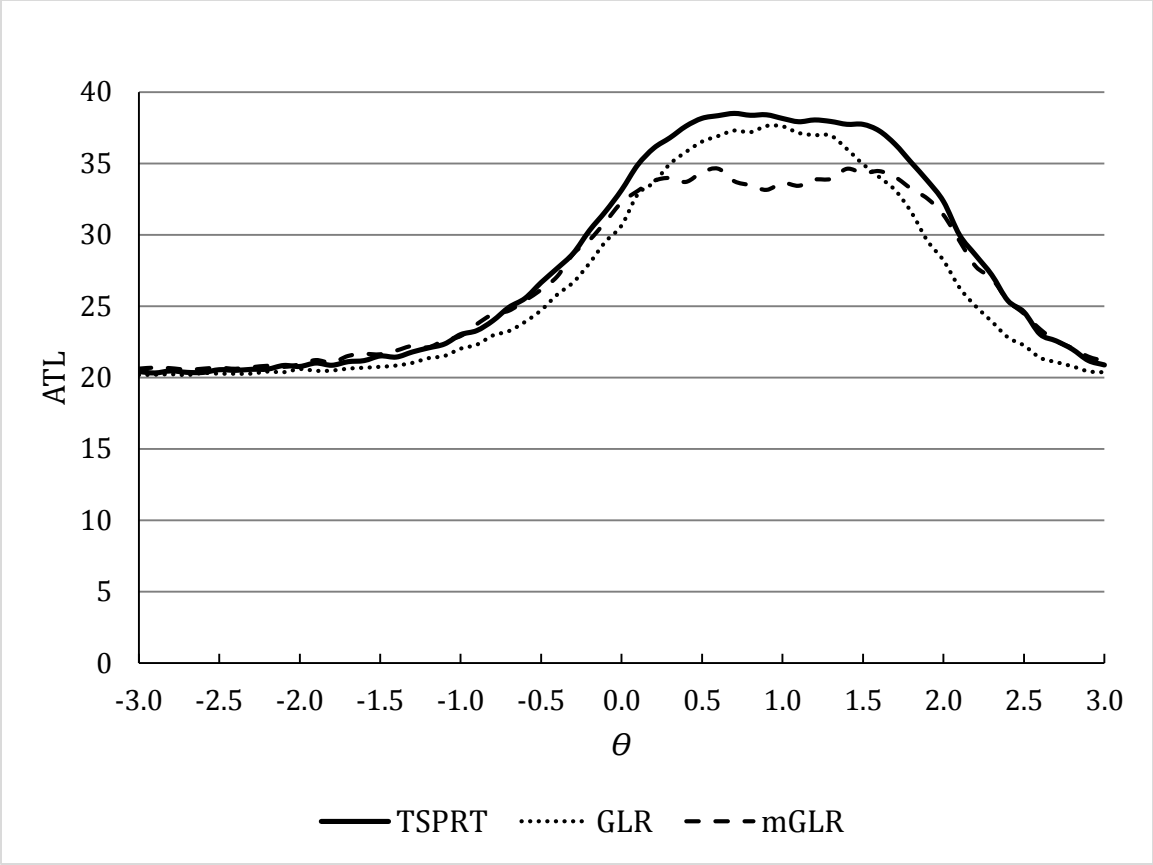
ATL results for the lower cutscore in the conditions with the 40 item maximum test length range from 38.16 to 34.38. The mGLR procedure has the lowest ATL followed by the GLR and TSPRT procedures. Similarly, for the upper cutscore the mGLR has the lowest ATL at 34.31 followed closely by the GLR. For the theta values between the two cutscores, the mGLR yields the best results with lower ATL than both the TSPRT and GLR procedures.

For the conditions with the 60 item maximum test length, the mGLR outperforms the TSPRT and GLR procedures. ATL for the lower cutscore range from 44.91 to 52.12. The upper cutscore yields ATLs ranging from 48.86 to 41.97. The theta levels between the two cutscores also show that the mGLR procedure results in lower ATLs. Figure 21 presents the conditional ATL plots for the procedures using two cutscores with a 40 item maximum test length while Figure 22 presents the plots for the 60 item maximum test length.

Table 5: Conditional average test length (ATL) and standard deviation (SD) for the two-cutscore item maximum test length conditions.

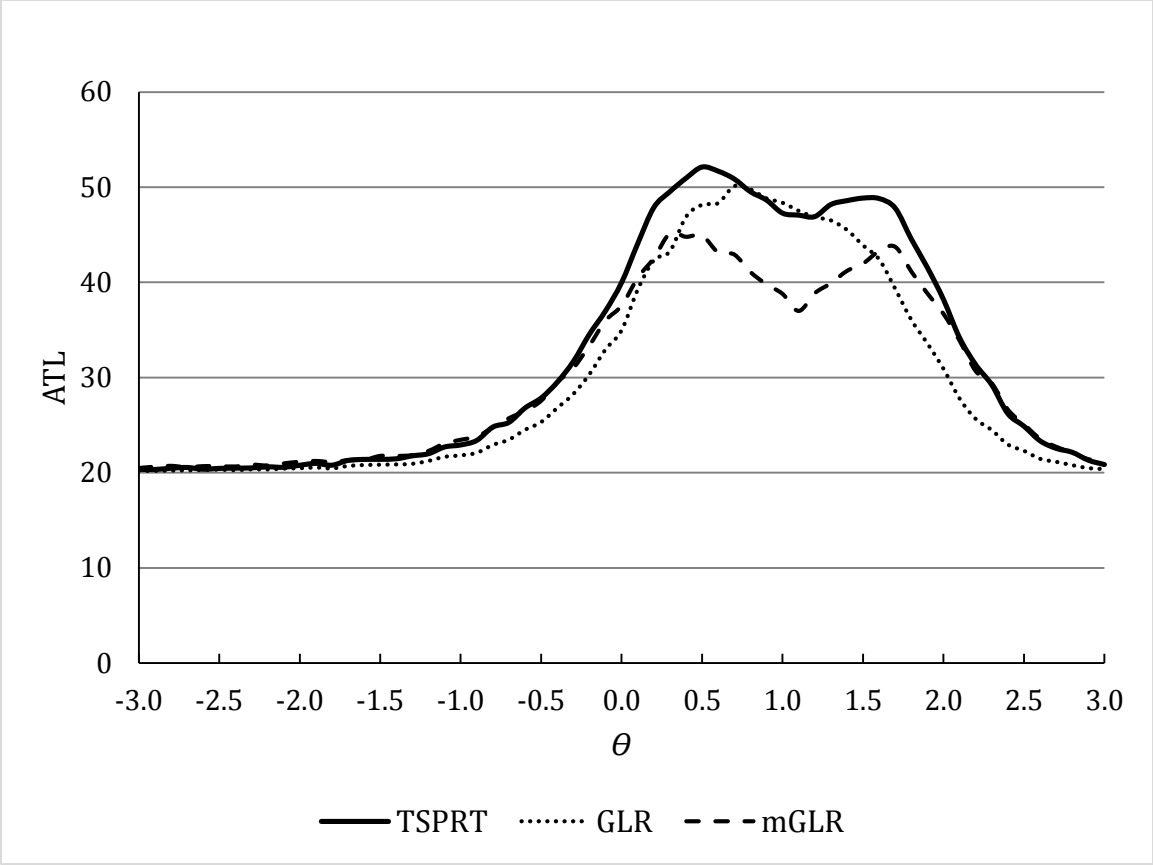
	40 Item Maximum						60 Item Maximum					
	TSPRT		GLR		mGLR		TSPRT		GLR		mGLR	
Theta	ATL	SD	ATL	SD	ATL	SD	ATL	SD	ATL	SD	ATL	SD
-3.0	20.42	1.62	20.17	0.95	20.62	2.19	20.36	1.75	20.19	1.04	20.51	2.01
-2.5	20.55	1.97	20.26	1.35	20.69	2.42	20.45	1.70	20.31	1.39	20.70	2.30
-2.0	20.78	2.49	20.59	2.12	20.93	2.69	20.79	2.46	20.46	1.67	21.16	3.46
-1.5	21.51	3.48	20.74	2.29	21.63	3.69	21.41	3.45	20.83	2.39	21.78	4.28
-1.0	23.01	5.25	22.02	4.24	22.91	5.21	22.92	5.71	21.81	4.01	23.46	6.07
-0.5	26.66	7.43	24.71	6.68	26.15	7.43	27.85	10.88	25.31	8.76	27.56	10.46
0.0	33.14	8.11	30.65	8.50	32.27	8.31	39.99	15.11	34.97	14.66	37.56	15.23
0.5	38.16	4.58	36.53	6.41	34.38	7.87	52.12	11.81	48.16	14.67	44.91	16.34
1.0	38.15	3.42	37.60	5.09	33.62	7.46	47.28	10.72	48.37	12.74	38.78	13.78
1.5	37.74	4.75	34.94	7.28	34.31	7.09	48.86	12.81	43.89	14.57	41.97	14.95
2.0	32.33	8.13	28.19	8.02	31.37	8.32	38.20	15.12	30.89	12.53	36.68	15.41
2.5	24.59	6.43	22.23	4.84	24.47	6.56	24.89	8.25	22.29	4.92	25.09	8.70
3.0	20.88	2.68	20.35	1.65	21.15	3.12	20.86	2.46	20.38	1.71	21.22	3.62

TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.



TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.

Figure 21: Conditional average test length (ATL) for the two-cutscore 40 item maximum test length conditions.



TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.

Figure 22: Conditional average test length (ATL) for the two-cutscore 60 item maximum test length conditions.

Table 6 presents the conditional means and standard deviations for the TSPRT, GLR, and mGLR procedures when three cutscores were used. The lowest cutscore for the three-cutscore conditions was placed 0.50 standard deviations below the peak of the test information function at the theta value of 0.50. The middle cutscore was placed at the peak of the test information function at the theta value of 1.00. The highest cutscore was placed 0.50 standard deviations above the peak of the test information function at the theta value of 1.50.

The ATL results for the lowest cutscore in the 40 item maximum test length range from 38.96 to 37.80. The GLR procedure yields the lowest ATL followed closely by the mGLR and TSPRT procedures. The ATL results for the middle cutscore range from 39.80 to 38.16 with the mGLR providing the lowest ATL. Analogous to the results from the lowest cutscore, the results from the highest cutscore range from 38.43 to 37.29 with the GLR slightly outperforming the mGLR. For the theta values between the two cutscores, the mGLR yields the best results with lower ATL than both the TSPRT and GLR procedures.

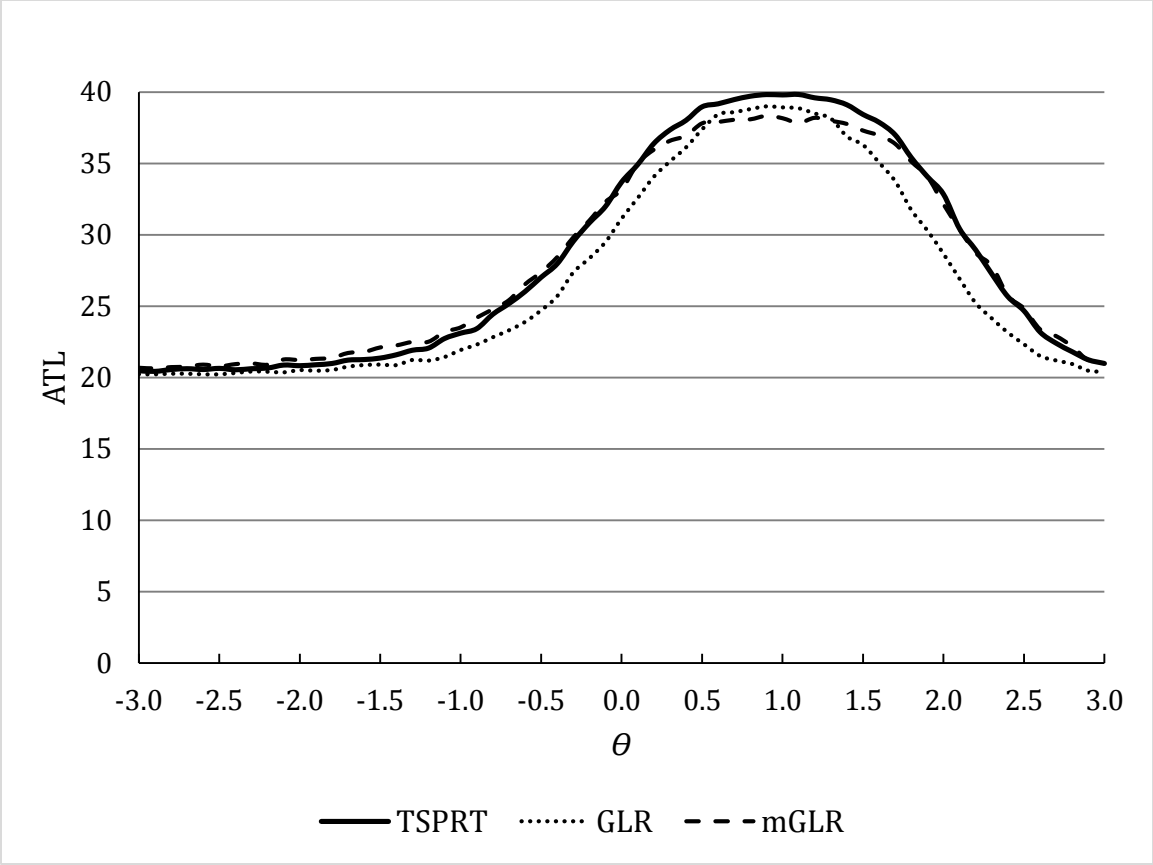
For the conditions with the 60 item maximum test length, the mGLR and GLR were highly similar, outperforming the TSPRT procedure at the lowest cutscore with ATLs ranging from 54.81 to 49.57. ATL for the middle cutscore range from 56.65 to 50.16 with the mGLR producing the lowest ATL. The highest cutscore yields ATLs from 53.52 to 47.34 where the GLR slightly outperformed the mGLR. The theta levels

between the three cutscores also show that the mGLR procedure results in lower ATLs. Figures 21 presents the conditional ATL plots for the procedures using a single cutscore with a 40 item maximum test length while Figure 22 presents the plots for the 60 item maximum test length.

Table 6: Conditional average test length (ATL) and standard deviation (SD) for the three-cutscore item maximum test length conditions.

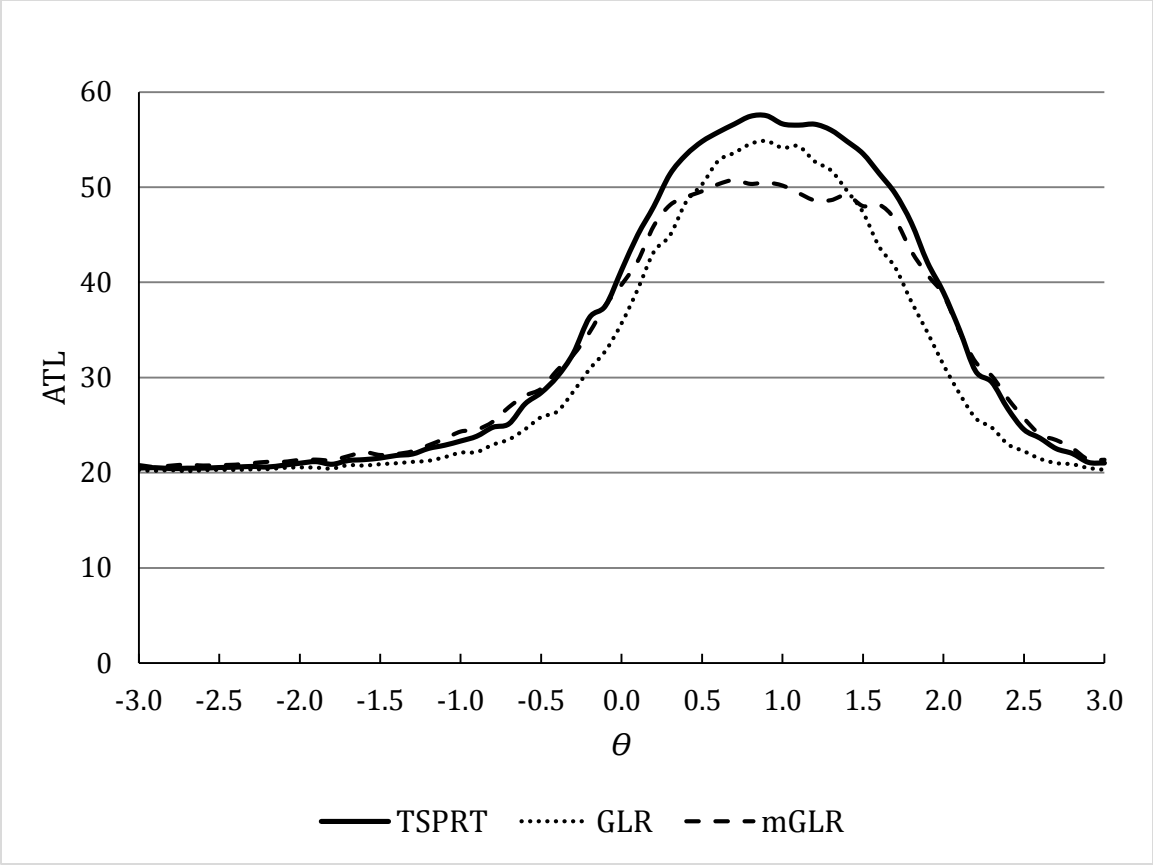
	40 Item Maximum						60 Item Maximum					
	TSPRT		GLR		mGLR		TSPRT		GLR		mGLR	
Theta	ATL	SD	ATL	SD	ATL	SD	ATL	SD	ATL	SD	ATL	SD
-3.0	20.48	2.00	20.22	1.09	20.68	2.43	20.48	1.87	20.27	1.45	20.81	2.97
-2.5	20.64	2.45	20.21	1.05	20.77	2.49	20.54	1.86	20.31	1.50	20.81	2.84
-2.0	20.83	2.61	20.51	1.98	21.22	3.40	21.00	2.92	20.56	2.17	21.34	3.82
-1.5	21.36	3.16	20.90	2.68	22.10	4.39	21.54	3.92	20.90	2.54	21.87	4.42
-1.0	23.11	5.41	21.92	4.21	23.51	5.75	23.33	6.45	22.11	4.62	24.33	7.15
-0.5	27.02	7.73	24.70	6.60	27.37	7.71	28.41	10.66	25.84	9.32	28.83	10.96
0.0	33.69	7.81	31.14	8.41	33.16	7.93	41.30	15.51	35.70	14.58	39.77	14.93
0.5	38.96	3.84	37.39	5.77	37.80	4.90	54.81	10.39	50.29	13.85	49.57	12.80
1.0	39.80	1.77	38.93	3.62	38.16	3.92	56.65	6.63	54.12	10.36	50.16	11.53
1.5	38.43	4.87	36.29	6.77	37.29	5.28	53.52	10.67	47.34	14.57	48.00	12.89
2.0	32.84	8.02	28.66	8.33	32.13	8.23	38.94	15.54	31.31	13.02	38.66	15.24
2.5	24.68	6.43	22.32	4.90	24.83	6.91	24.53	7.85	22.25	5.02	25.72	8.85
3.0	20.99	2.88	20.40	1.77	21.33	3.56	21.02	2.96	20.31	1.39	21.37	3.60

TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.



TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.

Figure 23: Conditional average test length (ATL) for the three-cutscore 40 item maximum test length conditions.



TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.

Figure 24: Conditional average test length (ATL) for the three-cutscore 60 item maximum test length conditions.

Percent Correct Classification

All of the classification methods compared in this section used a cutscore-based item selection method. Tables 7, 8, and 9 present the conditional percent correct classification and an overall accuracy percentage at the bottom of the tables for all three classification procedures using the two maximum test lengths. The accompanying plots, Figures 25 through 30, display conditional percent correct classification for each procedure based on the number of cutscores used and maximum test length for each study condition. Table 7 provides the conditional percent correct classification for the TSPRT, GLR, and mGLR procedures when a single cutscore was used. The cutscore for the single-cutscore conditions was placed at the theta value of 1.00.

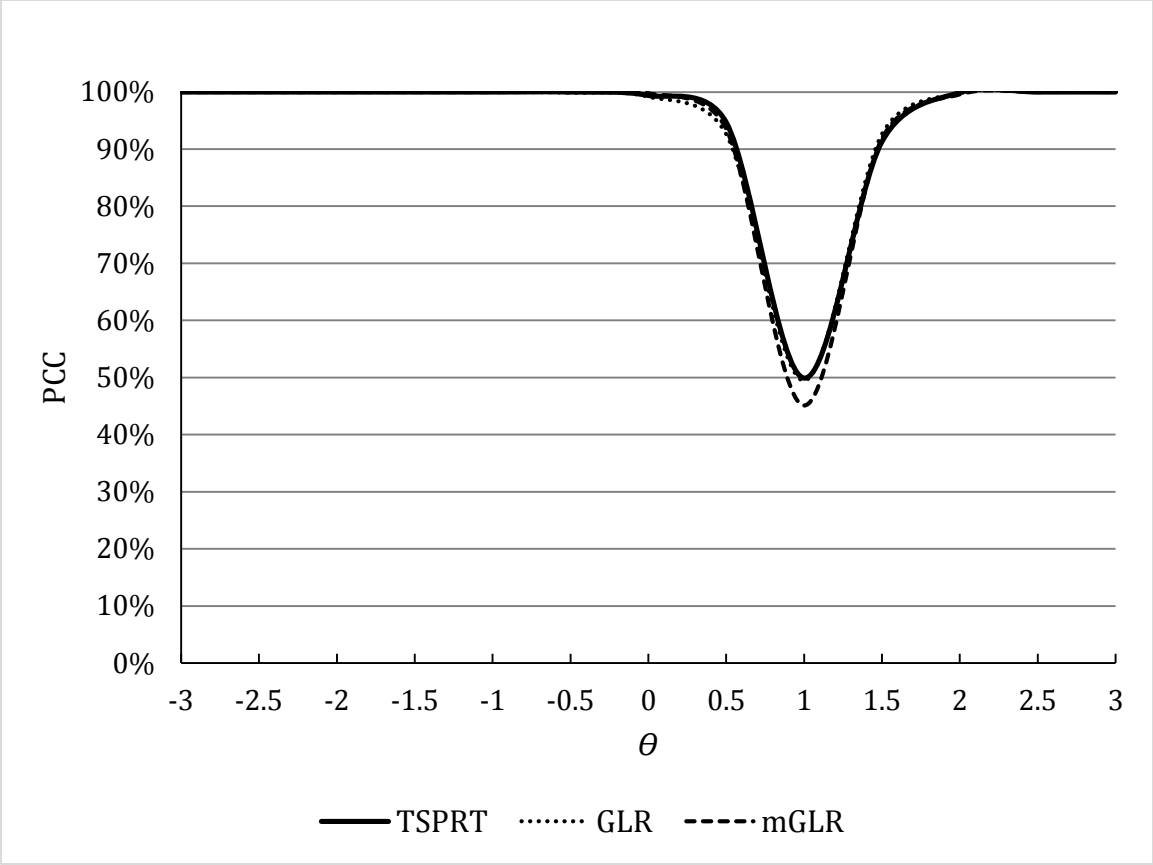
For the conditions with the 40 item maximum test length, at the cutscore the PCC ranges from 45.1% to 49.9%. The mGLR has the poorest performance at the cutscore but exhibits similar PCCs to the other procedures at the remaining theta values. The overall accuracy of classifications is highly similar across procedures.

Similarly, for the conditions with the 60 item maximum test length, at the cutscore the PCC ranges from 40.8% to 49.3%. Here again, the mGLR performs the poorest at the cutscore but provides similar PCC results across the remainder of the theta scale. Figures 25 presents the conditional PCC plots for the procedures using a single cutscore with a 40 item maximum test length while Figure 26 presents the plots for the 60 item maximum test length.

Table 7: Conditional percent correct classification for the single-cutscore conditions.

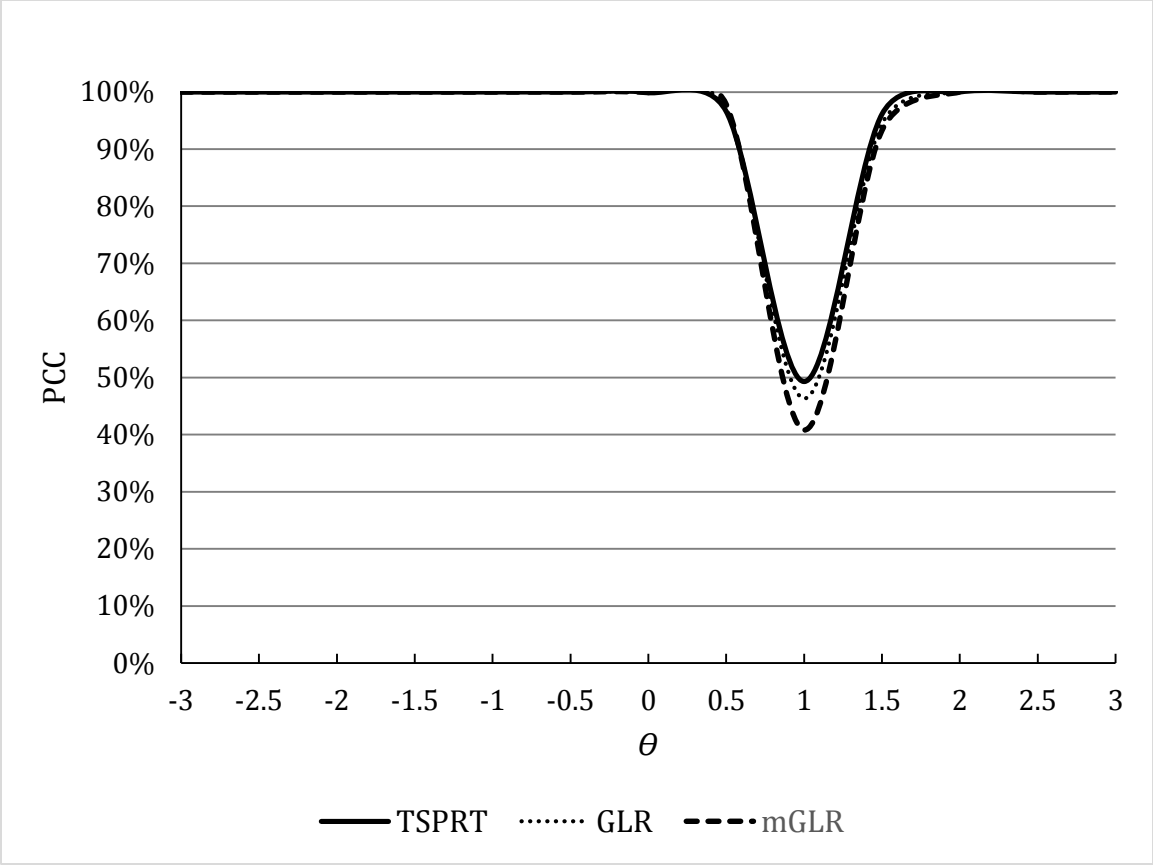
	40 Item Maximum			60 Item Maximum		
	TSPRT	GLR	mGLR	TSPRT	GLR	mGLR
Theta	Percent Correct Classification			Percent Correct Classification		
-3.0	100%	100%	100%	100%	100%	100%
-2.5	100%	100%	100%	100%	100%	100%
-2.0	100%	100%	100%	100%	100%	100%
-1.5	100%	100%	100%	100%	100%	100%
-1.0	100%	100%	100%	100%	100%	100%
-0.5	100%	99.9%	100%	100%	100%	100%
0.0	99.4%	99.2%	99.8%	100%	100%	100%
0.5	94.7%	92.7%	93.6%	96.6%	97.1%	97.9%
1.0	49.9%	49.5%	45.1%	49.3%	46.3%	40.8%
1.5	91.3%	92.6%	91.8%	96.1%	94.4%	93.2%
2.0	99.9%	99.8%	99.6%	100%	100%	100%
2.5	100%	100%	100%	100%	100%	100%
3.0	100%	100%	100%	100%	100%	100%
Overall	95.4%	95.5%	95.3%	96.2%	96.0%	95.9%

TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.



PCC = Percent Correct Classification; TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.

Figure 25: Conditional percent correct classification (PCC) for the single-cutscore 40 item maximum test length conditions.



PCC = Percent Correct Classification; TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.

Figure 26: Conditional percent correct classification (PCC) for the single-cutscore 60 item maximum test length conditions.

Table 8 provides the conditional percent correct classification for the TSPRT, GLR, and mGLR procedures when two cutscores were used. The lower cutscore for the two-cutscore conditions was placed 0.50 standard deviations below the peak of the test information function at the theta value of 0.50 while the upper cutscore for the two-cutscore conditions was placed 0.50 standard deviations above the peak of the test information function at the theta value of 1.50.

For the conditions with the 40 item maximum test length, at the lower cutscore the PCC ranges from 41.9% to 60.8%. The mGLR has the best performance at the lower cutscore and exhibits similar PCCs to the other procedures at the majority of the remaining theta values. The overall accuracy of classifications at the lower cutscore are similar across procedures ranging from 87.7% to 90.4%.For the upper cutscore using a 40 item maximum test length the mGLR exhibits the poorest performance with a PCC of 30.0% while the PCC for the TSPRT and GLR are 40.6% and 53.5% respectively. The PCC for the theta levels between the cutscores show that the mGLR and TSPRT, 90.8% and 92.5%, outperform the GLR procedure, 76.2%.

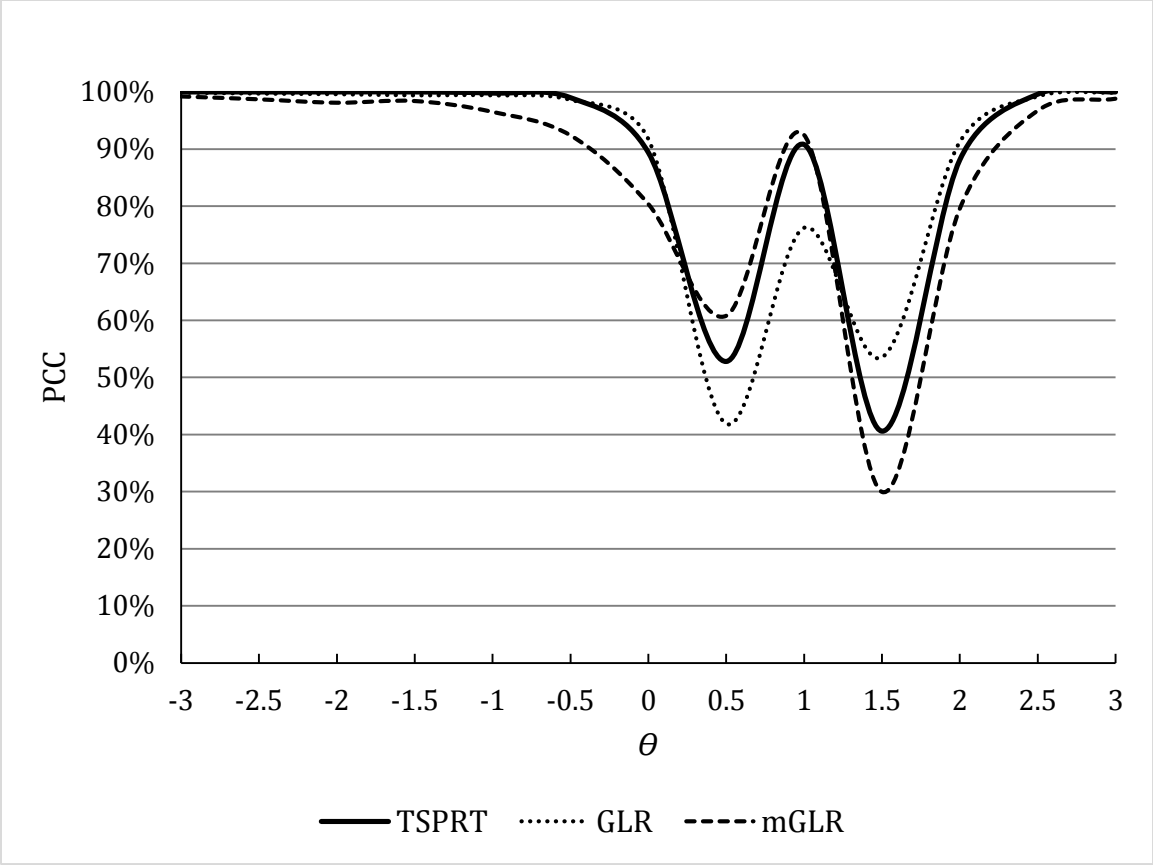
Similar to the results of the 40 item maximum test length, the conditions with the 60 item maximum test length the mGLR performs the best at the lower cutscore, but the poorest at the upper cutscore. Even with the poor performance at the upper cutscore the mGLR provided similar PCC results across the remainder of the theta scale. At the lower cutscore PCC ranges from 43.9% to 60.1%. The PCC results at the upper cutscore range

from 25.4% to 51.8% with the GLR performing the best. Again, the PCC for the theta levels between the two cutscores show that the mGLR and TSPRT outperform the GLR procedure by at least 8.7%. Figures 27 and 28 present the conditional PCC plots for the 40 and 60 item maximum test lengths.

Table 8: Conditional percent correct classification for the two-cutscore conditions.

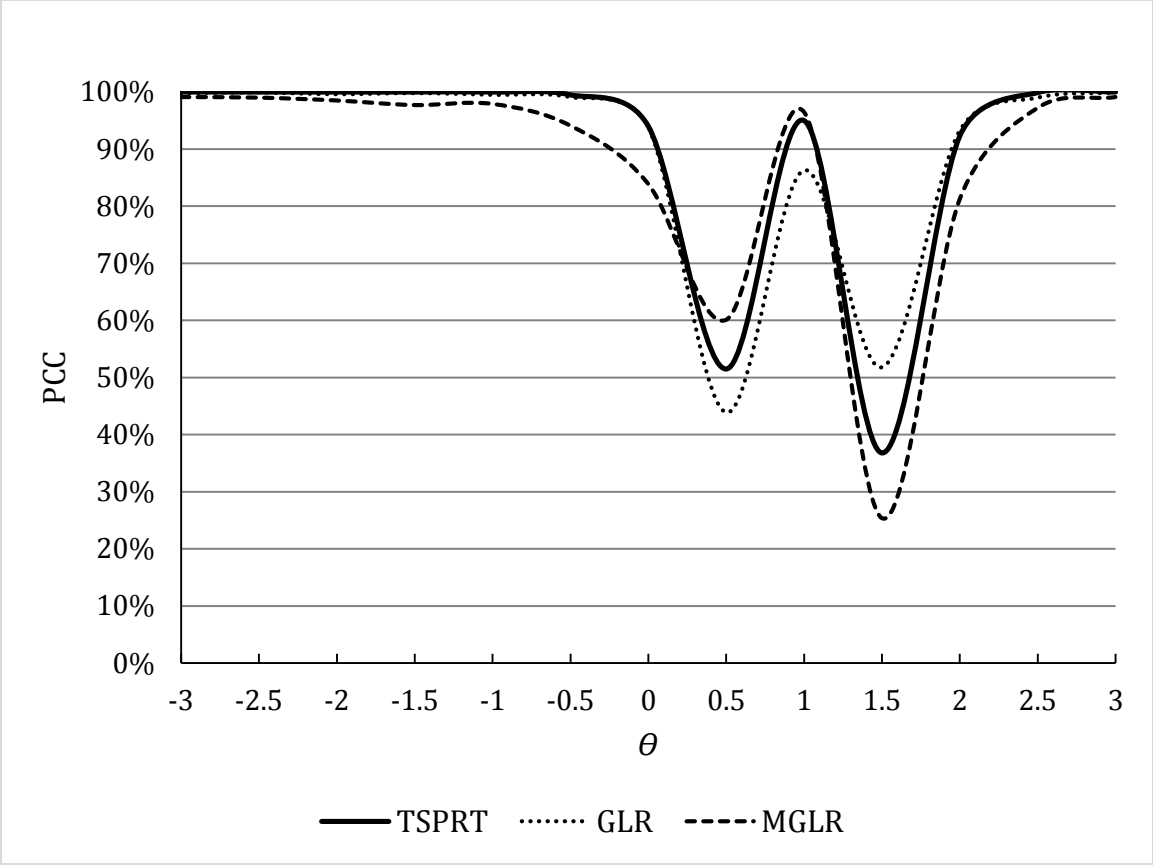
	40 Item Maximum			60 Item Maximum		
	TSPRT	GLR	mGLR	TSPRT	GLR	mGLR
Theta	Percent Correct Classification			Percent Correct Classification		
-3.0	100%	99.9%	99.2%	100%	100%	99.1%
-2.5	100%	99.7%	98.7%	100%	100%	99.0%
-2.0	100%	99.6%	98.1%	100%	100%	98.5%
-1.5	100%	99.4%	98.4%	100%	100%	97.7%
-1.0	99.8%	99.4%	96.5%	100%	100%	97.9%
-0.5	99.0%	98.6%	92.4%	100%	99.1%	94.1%
0.0	89.4%	91.8%	80.4%	94.0%	93.9%	83.9%
0.5	52.8%	41.9%	60.8%	51.5%	43.9%	60.1%
1.0	90.8%	76.2%	92.5%	95.0%	86.3%	96.5%
1.5	40.6%	53.5%	30.0%	36.8%	51.8%	25.4%
2.0	88.1%	91.2%	79.5%	92.3%	93.2%	81.2%
2.5	99.6%	99.2%	96.7%	100%	99.0%	97.2%
3.0	100%	99.9%	98.8%	100%	100%	99.1%
Overall	90.4%	89.3%	87.7%	91.9%	91.0%	88.7%

TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.



PCC = Percent Correct Classification; TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.

Figure 27: Conditional percent correct classification (PCC) for the two-cutscore 40 item maximum test length conditions.



PCC = Percent Correct Classification; TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.

Figure 28: Conditional percent correct classification (PCC) for the two-cutscore 60 item maximum test length conditions.

Table 9 presents the conditional PCC for the TSPRT, GLR, and mGLR procedures when three cutscores were used. The lowest cutscore for the three-cutscore conditions was placed 0.50 standard deviations below the peak of the test information function at the theta value of 0.50. The middle cutscore was placed at the peak of the test information function at the theta value of 1.00. The highest cutscore was placed 0.50 standard deviations above the peak of the test information function at the theta value of 1.50.

The PCC results for the lowest cutscore in the 40 item maximum test length range from 33.6% to 53.5%. The GLR procedure has the highest PCC, 53.5%, at the lowest cutscore while the GLR has the lowest PCC, 33.6. The PCC results for the middle cutscore range from 38.3% to 45.9%. The mGLR provides the highest PCC followed closely by the TSPRT while the GLR performs the poorest. For the highest cutscore, the GLR performs the best with a PCC of 53.7% while the mGLR yields the poorest performance with 28.4% correct classification.

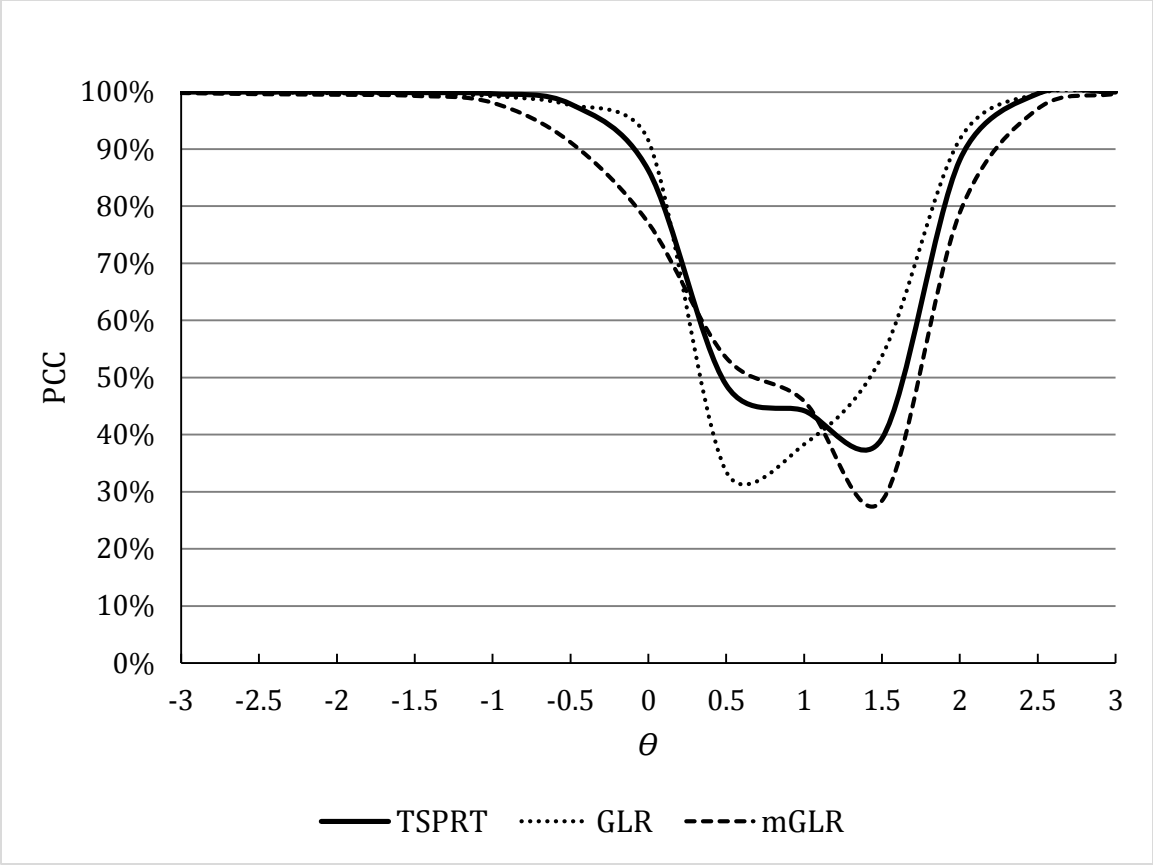
The results for the conditions with the 60 item maximum test length follow a similar pattern as the conditions with a 40 item maximum test length. At the lowest cutscore the GLR procedure has the highest PCC, 58.5%, while the GLR has the lowest PCC, 35.0%. The PCC results for the middle cutscore are similar across procedures ranging from 44.2% to 46.5%. For the highest cutscore, the GLR performs the best with a PCC of 49.8% while the mGLR performed the poorest with 21.4% correct classification.

Figure 29 presents the conditional PCC plots for the procedures using three cutscores with a 40 item maximum test length while Figure 30 presents the plots for the 60 item maximum test length conditions.

Table 9: Conditional percent correct classification for the three-cutscore conditions.

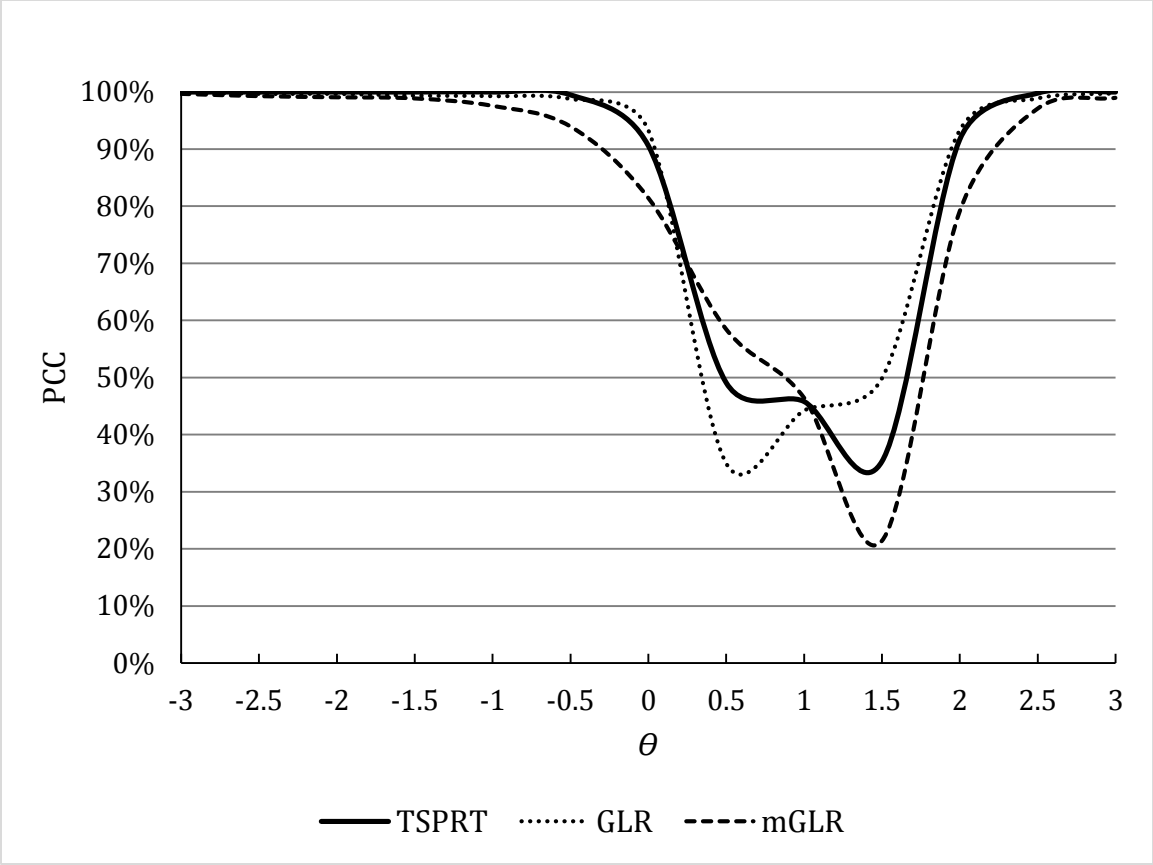
	40 Item Maximum			60 Item Maximum		
	TSPRT	GLR	mGLR	TSPRT	GLR	mGLR
Theta	Percent Correct Classification			Percent Correct Classification		
-3.0	100%	99.9%	99.8%	100%	100%	99.7%
-2.5	100%	100%	99.6%	100%	99.7%	99.3%
-2.0	100%	99.8%	99.5%	100%	99.7%	99.1%
-1.5	99.9%	99.7%	99.3%	100%	99.4%	98.9%
-1.0	99.7%	99.3%	98.1%	100%	99.3%	97.6%
-0.5	97.9%	97.7%	91.2%	99.5%	98.8%	94.0%
0.0	86.3%	91.6%	77.1%	90.6%	93.4%	81.5%
0.5	48.7%	33.6%	53.5%	49.1%	35.0%	58.5%
1.0	44.2%	38.3%	45.9%	45.8%	44.2%	46.5%
1.5	39.3%	53.7%	28.4%	35.3%	49.8%	21.4%
2.0	88.1%	91.7%	78.8%	91.7%	93.3%	79.1%
2.5	99.7%	99.7%	97.0%	99.8%	98.9%	97.1%
3.0	100%	100%	99.7%	100%	99.8%	99.0%
Overall	86.2%	85.7%	82.9%	88.0%	87.1%	84.0%

TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.



TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.

Figure 29: Conditional percent correct classification (PCC) for the three-cutscore 40 item maximum test length conditions.



PCC = Percent Correct Classification; TSPRT = Truncated Sequential Probability Ratio Test; GLR = Generalized Likelihood Ratio; mGLR = Modified Generalized Likelihood Ratio.

Figure 30: Conditional percent correct classification (PCC) for the three-cutscore 60 item maximum test length conditions.

Modified-GLR Procedures

Average Test Length

Results presented in this section are provided to enable comparisons between the mGLR procedures under two item selection methods. The results from the mGLR procedure using cutscore-based item selection method which were presented in the previous section are also presented in this section for comparisons against the mGLR procedure using ability-based item selection. Tables 10, 11, and 12 present the conditional means and standard deviations of the test lengths for the two classification procedures using two maximum test lengths. The accompanying plots, Figures 31 through 36, display conditional average test lengths for each procedure based on the number of cutscores and maximum test length.

Table 10 provides the conditional means and standard deviations for the mGLR procedures when a single cutscore was used. The cutscore for the single-cutscore conditions was placed at the peak of the test information function at the theta value of 1.00. For the 40 item maximum test length conditions, the mGLR procedure using the ability-based item selection method produced a shorter average test length at the cutscore of 29.55 items compared to the mGLR using the cutscore-based item selection average test length of 37.62. Additionally, the majority of the conditional standard deviation

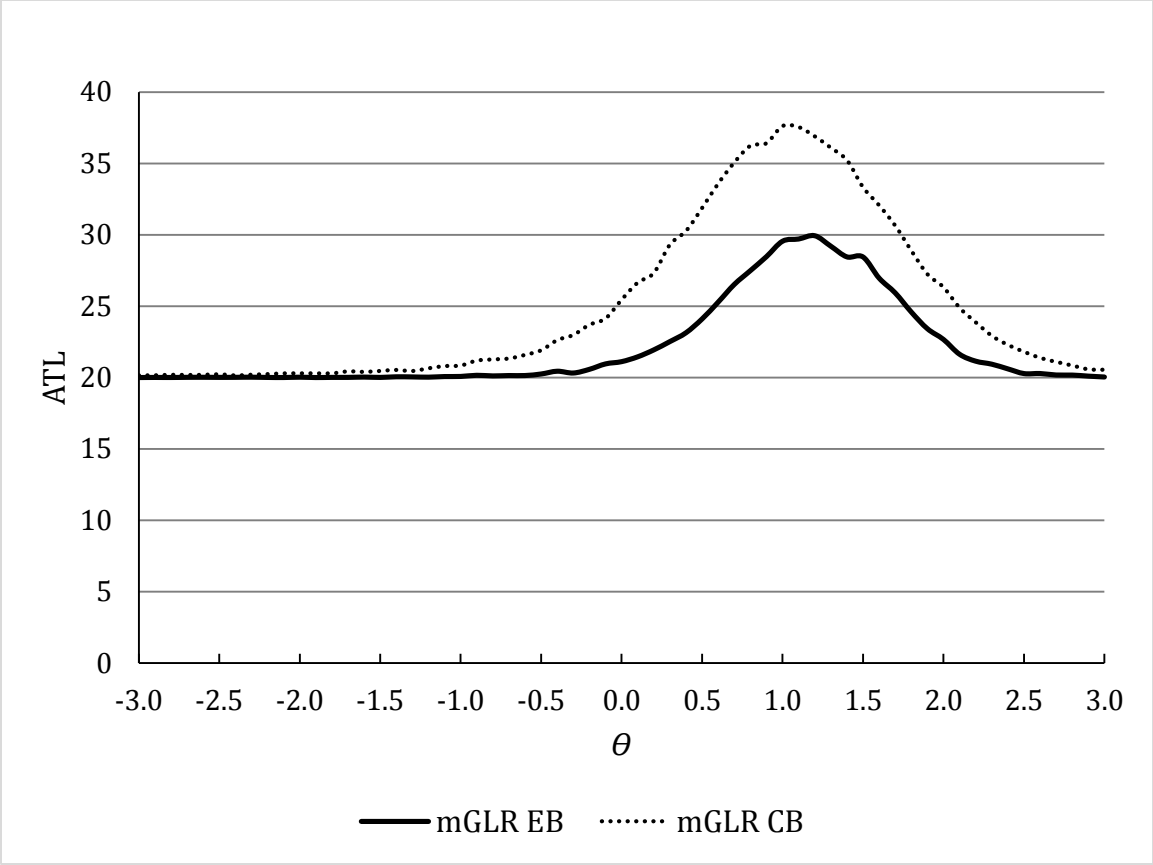
values for the mGLR using the ability-based item selection are much lower than the standard deviation values produced by the cutscore-based method.

For the conditions with a maximum test length of 60 items, the average test length at the cutscore for the ability-based item selection method was 34.38 while the cutscore-based item selection approach results in an average test length of 50.27. Similar to 40 item test length conditions, the majority of the conditional standard deviations produced by the ability-based method are much lower compared to the cutscore-based method. Figures 31 and 32 present the plots for the conditional average test length for the 40 and 60 item tests lengths.

Table 10: Conditional average test length (ATL) and standard deviation (SD) for the single-cutscore conditions using the mGLR procedures with different item selection methods.

	40 Item Maximum				60 Item Maximum			
	mGLR with cutscore-based item selection		mGLR with ability-based item selection		mGLR with cutscore-based item selection		mGLR with ability-based item selection	
Theta	ATL	SD	ATL	SD	ATL	SD	ATL	SD
-3.0	20.15	0.85	20.00	0.00	20.16	0.81	20.00	0.06
-2.5	20.22	1.09	20.01	0.24	20.24	1.04	20.00	0.03
-2.0	20.29	1.22	20.03	0.40	20.27	1.26	20.01	0.19
-1.5	20.45	1.60	20.01	0.16	20.47	1.69	20.03	0.33
-1.0	20.84	2.53	20.08	0.71	20.80	2.45	20.05	0.62
-0.5	21.90	3.98	20.25	1.59	22.16	4.61	20.43	2.24
0.0	25.45	6.75	21.11	3.56	25.84	8.04	21.02	3.62
0.5	31.88	8.02	24.10	7.04	36.43	13.59	25.37	10.02
1.0	37.62	5.43	29.55	8.98	50.27	13.17	34.38	15.77
1.5	33.33	7.73	28.45	8.75	39.89	14.75	31.91	15.21
2.0	26.32	6.86	22.67	5.59	26.64	8.46	22.46	6.66
2.5	21.80	3.80	20.28	1.65	21.76	3.55	20.43	2.38
3.0	20.54	1.67	20.04	0.43	20.55	1.63	20.03	0.34

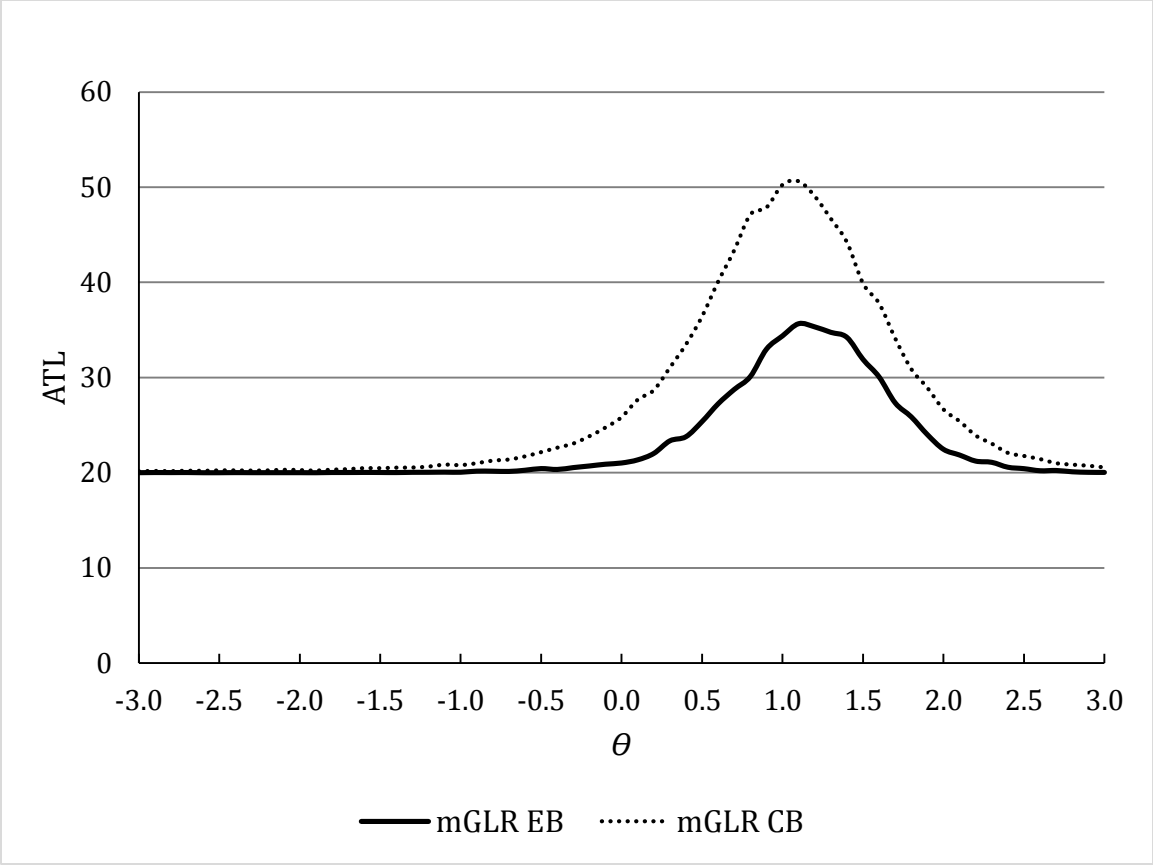
mGLR = Modified Generalized Likelihood Ratio.



mGLR EB = Modified Generalized Likelihood Ratio Estimate-Based item selection;

mGLR CB = Modified Generalized Likelihood Ratio Cutscore-Based item selection.

Figure 31: Conditional average test length (ATL) for the one-cutscore 40 item maximum test length conditions using the mGLR with multiple item selection methods.



mGLR EB = Modified Generalized Likelihood Ratio Estimate-Based item selection;

mGLR CB = Modified Generalized Likelihood Ratio Cutscore-Based item selection.

Figure 32: Conditional average test length (ATL) for the one-cutscore 60 item maximum test length conditions using the mGLR with multiple item selection methods.

Table 11 provides the conditional means and standard deviations for the mGLR procedures when two cutscores were used. The lower cutscore for the two-cutscore conditions was placed 0.50 standard deviations below the peak of the test information function at the theta value of 0.50 while the upper cutscore for the two-cutscore conditions was placed 0.50 standard deviations above the peak of the test information function at the theta value of 1.50.

For the 40 item maximum test length conditions, the mGLR procedure using the ability-based item selection method produced a shorter average test length at the lower cutscore of 28.45 items compared to the mGLR using the cutscore-based item selection average test length of 34.38. The mGLR procedure with ability-based item selection also produced a lower average test length at the upper cutscore, 29.71, compared to the cutscore-based procedure average test length of 34.31. The average test length for the theta values between the cutscores for the ability-based item selection method was also superior to the cutscore-based item selection method. Additionally, the majority of the conditional standard deviation values for the mGLR using the ability-based item selection are much lower than the standard deviation values produced by the cutscore-based method.

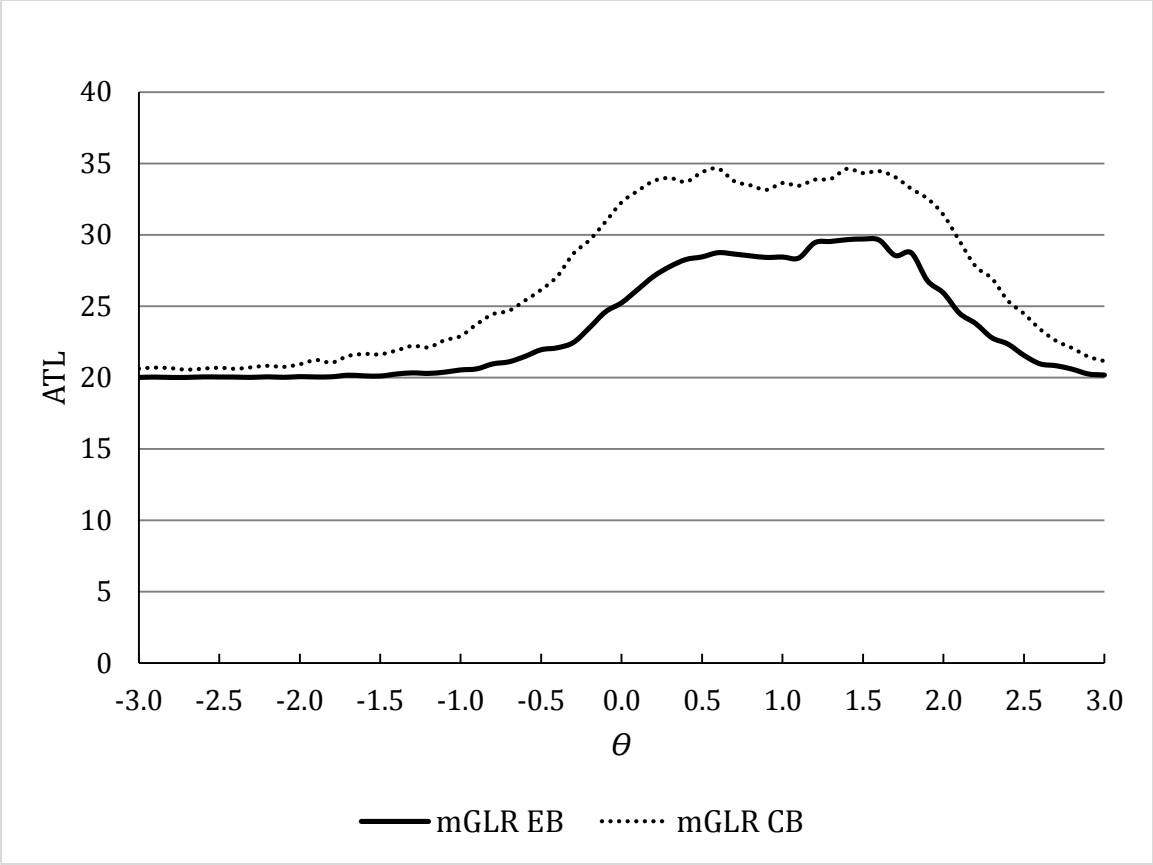
For the conditions with a maximum test length of 60 items, the average test length at the lower cutscore for the ability-based item selection method is 34.08, while the cutscore-based item selection approach produced an average test length of 44.91. The

mGLR procedure with ability-based item selection also produced a lower average test length at the upper cutscore of 34.67 items compared to the cutscore-based procedure average test length of 44.91 items. The average test length for the theta values between the two cutscores for the ability-based item selection method was also superior to the cutscore-based item selection method. Similar to 40 item test length conditions, the majority of the conditional standard deviations produced by the ability-based method are lower compared to the cutscore-based method. Figures 33 and 34 present plots of the conditional average test lengths for the 40 and 60 item maximum test lengths.

Table 11: Conditional average test length (ATL) and standard deviation (SD) for the two-cutscore conditions using the mGLR procedures with different item selection methods.

	40 Item Maximum				60 Item Maximum			
	mGLR with cutscore-based item selection		mGLR with ability-based item selection		mGLR with cutscore-based item selection		mGLR with ability-based item selection	
Theta	ATL	SD	ATL	SD	ATL	SD	ATL	SD
-3.0	20.62	2.19	20.01	0.15	20.51	2.01	20.03	0.51
-2.5	20.69	2.42	20.04	0.51	20.70	2.30	20.02	0.37
-2.0	20.93	2.69	20.06	0.63	21.16	3.46	20.06	0.89
-1.5	21.63	3.69	20.11	0.87	21.78	4.28	20.21	1.32
-1.0	22.91	5.21	20.53	2.43	23.46	6.07	20.63	2.99
-0.5	26.15	7.43	21.95	4.94	27.56	10.46	22.13	6.12
0.0	32.27	8.31	25.24	7.97	37.56	15.23	28.09	13.33
0.5	34.38	7.87	28.45	8.66	44.91	16.34	34.08	16.21
1.0	33.62	7.46	28.44	8.25	38.78	13.78	31.40	13.00
1.5	34.31	7.09	29.71	8.73	41.97	14.95	34.67	15.70
2.0	31.37	8.32	25.92	7.99	36.68	15.41	27.70	12.18
2.5	24.47	6.56	21.56	4.25	25.09	8.70	21.63	4.85
3.0	21.15	3.12	20.18	1.19	21.22	3.62	20.28	1.54

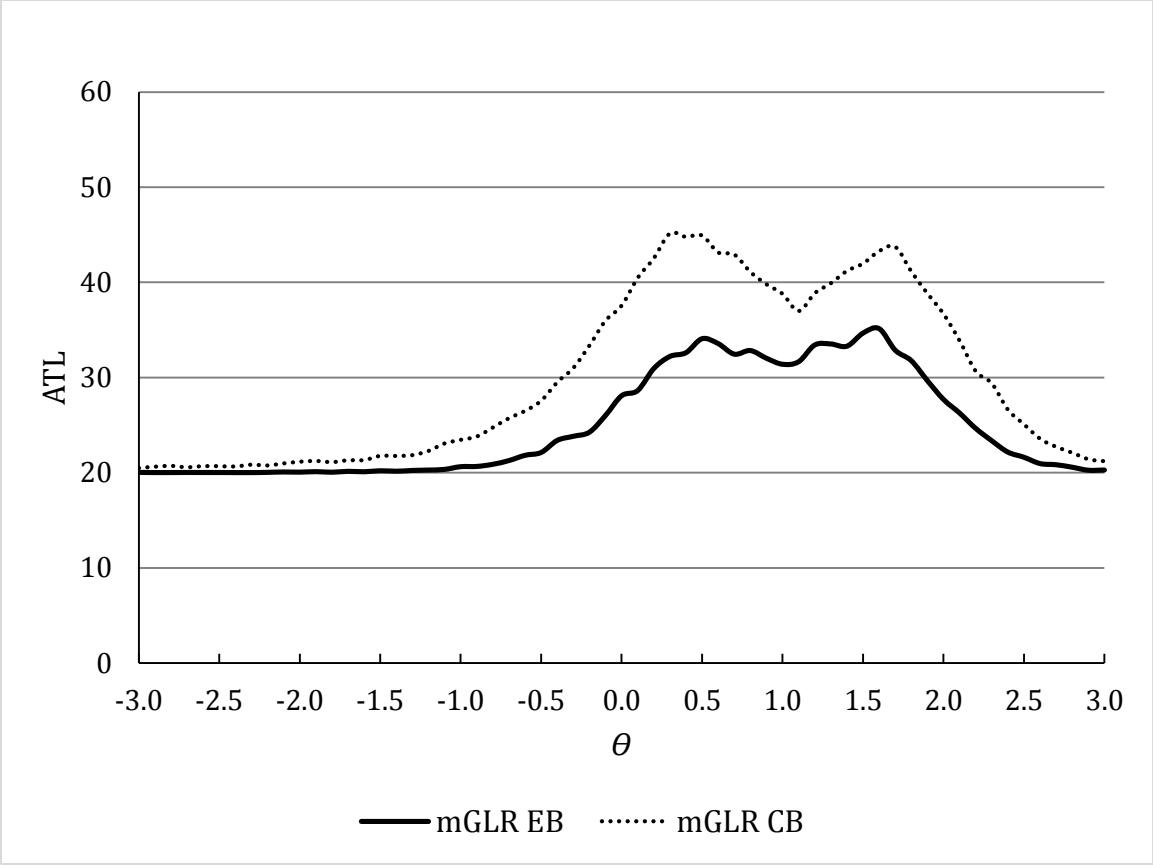
mGLR = Modified Generalized Likelihood Ratio.



mGLR EB = Modified Generalized Likelihood Ratio Estimate-Based item selection;

mGLR CB = Modified Generalized Likelihood Ratio Cutscore-Based item selection.

Figure 33: Conditional average test length (ATL) for the two-cutscore 40 item maximum test length conditions using the mGLR with multiple item selection methods.



mGLR EB = Modified Generalized Likelihood Ratio Estimate-Based item selection;

mGLR CB = Modified Generalized Likelihood Ratio Cutscore-Based item selection.

Figure 34: Conditional average test length (ATL) for the two-cutscore 60 item maximum test length conditions using the mGLR with multiple item selection methods.

Table 12 presents the conditional means and standard deviations for the mGLR procedures when three cutscores were used. The lowest cutscore for the three-cutscore conditions was placed 0.50 standard deviations below the peak of the test information function at the theta value of 0.50. The middle cutscore was placed at the peak of the test information function at the theta value of 1.00. The highest cutscore was placed 0.50 standard deviations above the peak of the test information function at the theta value of 1.50.

The results for the lowest cutscore in the 40 item maximum test length demonstrate that the average test length for the ability-based item selection mGLR procedure, 32.00 items, was better than the cutscore-based item selection method average test length of 37.80. At the middle cutscore, the average test length for the ability-based item selection is 34.37 while the cutscore-based method average test length is 38.16. For the highest cutscore the average test length for the ability-based item selection procedure is 33.08 while the cutscore-based item selection method average test length is 37.29. As presented in the conditional plot, Figure 35, for the theta values between the cutscores, the ability-based item selection method yielded lower test lengths on average when compared to the cutscore-based item selection method. In regards to the conditional standard deviations, the ability-based item selection method produced larger values at the theta values associated with the cutscores. Otherwise, the conditional standard deviations

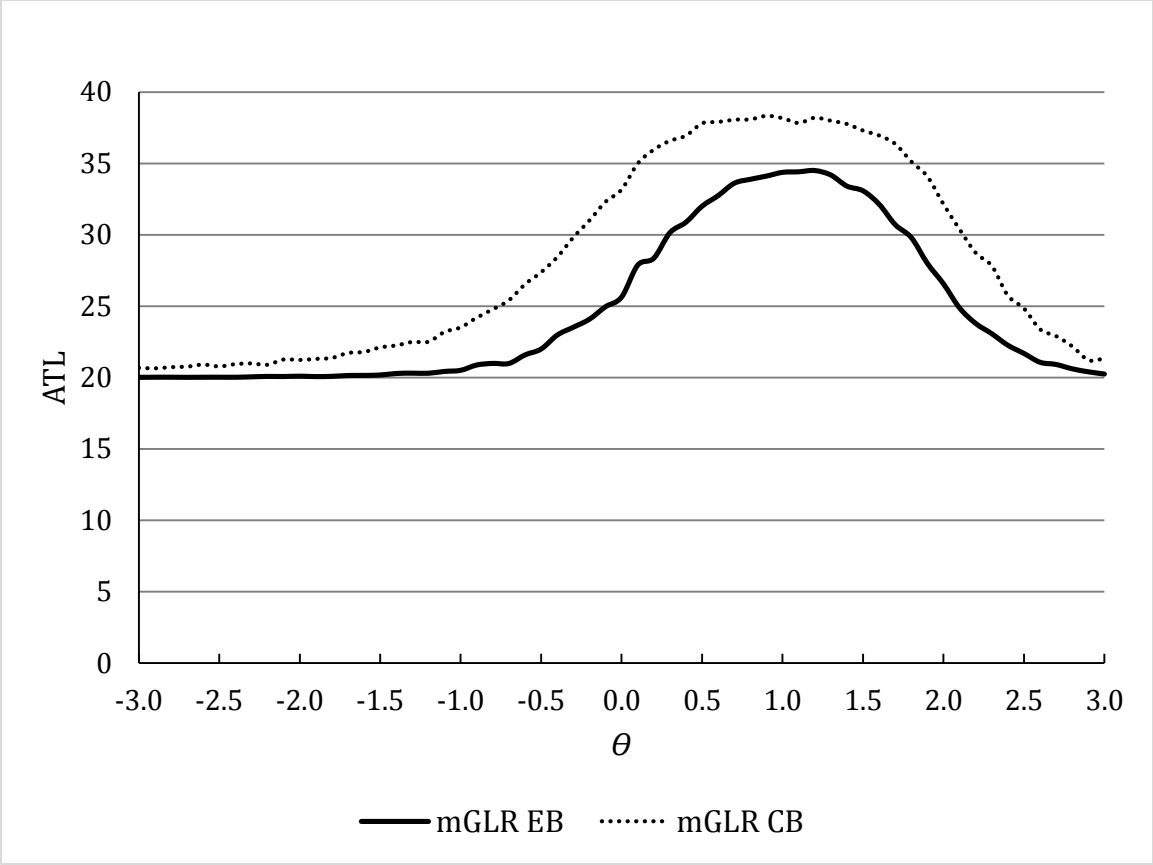
for the ability-based item selection method are noticeably lower than cutscore-based item selection method.

Similar to the results from the 40 item maximum test length, the mGLR procedure using the ability-based item selection method outperformed the mGLR procedure using the cutscore-based item selection at all three cutscores and at the theta values between the cutscores which can be seen in Figure 36. At the lowest cutscore the average test length for the ability-based procedure is 38.90 while the cutscore-based procedure average test length is 49.57. At the middle cutscore the ability-based procedure is 41.68 while the cutscore-based procedure produced an average test length of 50.16. The ATL results for the highest cutscore is 39.77 for the ability-based procedure and 48.00 for the cutscore-based procedure. As seen with the 40 item maximum test length results, the conditional standard deviations of the ability-based item selection methods for the 60 item maximum test length were larger at the theta values associated with the cutscores when compared to the cutscore-based item selection method. The conditional standard deviations for the ability-based item selection method improve so as to be lower than the cutscore-based method when theta values are not equal to the cutscore values. Figures 35 and 36 display plots of the average test lengths for the 40 and 60 item maximum test lengths.

Table 12: Average test length (ATL) and standard deviation (SD) for the three-cutscore conditions using the mGLR procedures with different item selection methods.

	40 Item Maximum				60 Item Maximum			
	mGLR with cutscore-based item selection		mGLR with ability-based item selection		mGLR with cutscore-based item selection		mGLR with ability-based item selection	
Theta	ATL	SD	ATL	SD	ATL	SD	ATL	SD
-3.0	20.68	2.43	20.01	0.19	20.81	2.97	20.01	0.27
-2.5	20.77	2.49	20.02	0.28	20.81	2.84	20.04	0.39
-2.0	21.22	3.40	20.10	0.91	21.34	3.82	20.07	0.72
-1.5	22.10	4.39	20.18	1.19	21.87	4.42	20.26	1.68
-1.0	23.51	5.75	20.51	2.23	24.33	7.15	20.58	2.69
-0.5	27.37	7.71	22.00	4.95	28.83	10.96	22.43	6.35
0.0	33.16	7.93	25.64	7.89	39.77	14.93	28.85	12.95
0.5	37.80	4.90	32.00	8.75	49.57	12.80	38.90	16.28
1.0	38.16	3.92	34.37	7.45	50.16	11.53	41.68	14.34
1.5	37.29	5.28	33.08	8.14	48.00	12.89	39.77	15.35
2.0	32.13	8.23	26.56	8.04	38.66	15.24	28.92	12.92
2.5	24.83	6.91	21.69	4.35	25.72	8.85	21.72	4.95
3.0	21.33	3.56	20.25	1.54	21.37	3.60	20.30	1.72

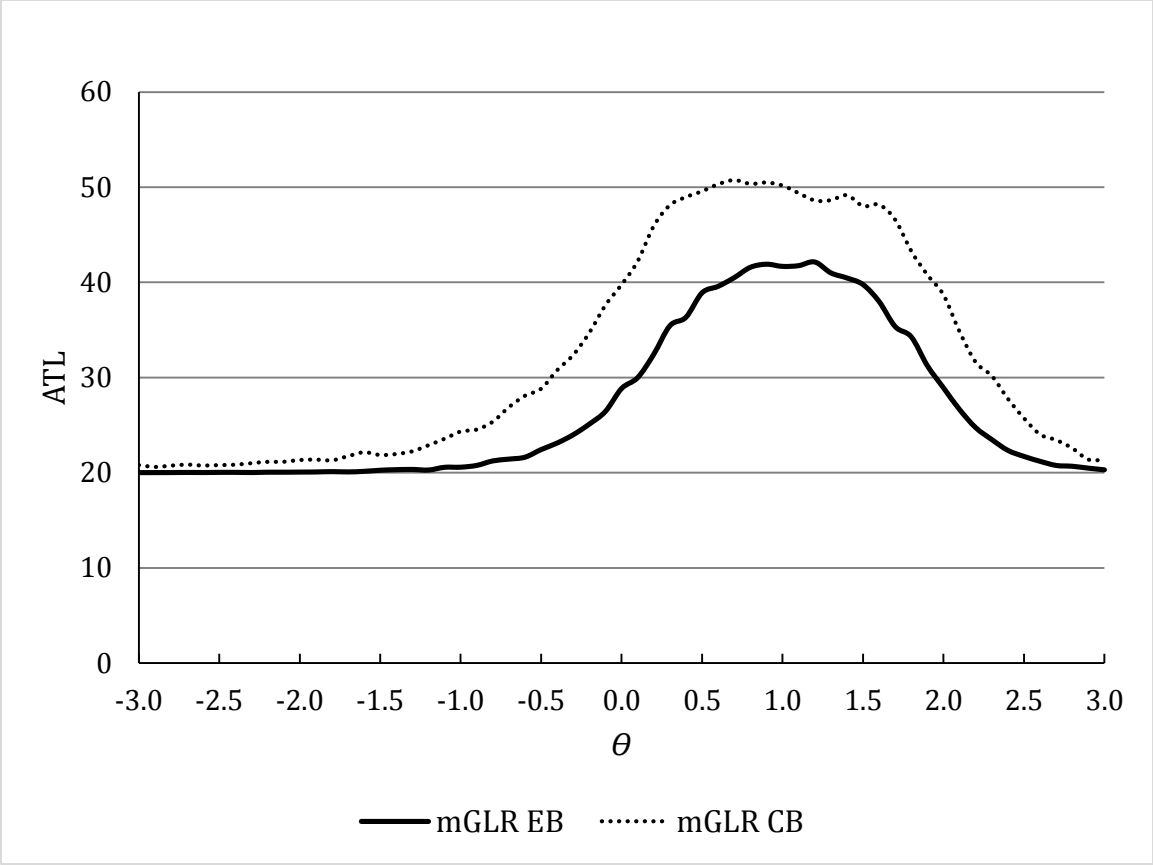
mGLR = Modified Generalized Likelihood Ratio.



mGLR EB = Modified Generalized Likelihood Ratio Estimate-Based item selection;

mGLR CB = Modified Generalized Likelihood Ratio Cutscore-Based item selection.

Figure 35: Conditional average test length (ATL) for the three-cutscore 40 item maximum test length conditions using the mGLR with multiple item selection methods.



mGLR EB = Modified Generalized Likelihood Ratio Estimate-Based item selection;

mGLR CB = Modified Generalized Likelihood Ratio Cutscore-Based item selection.

Figure 36: Conditional average test length (ATL) for the three-cutscore 60 item maximum test length conditions using the mGLR with multiple item selection methods.

Percent Correct Classification

The results in this section are provided to enable comparisons between the mGLR procedures using two item selection methods. The results from the mGLR procedure using cutscore-based item selection method which were presented in the previous Percent Correct Classification section are also presented in this section for comparisons against the mGLR procedure using ability-based item selection. Tables 13, 14, and 15 present the conditional percent correct classification for the two classification procedures using two maximum test lengths. The accompanying plots, Figures 37 through 42, display conditional percent correct classification for each procedure based on the number of cutscores and maximum test length.

Table 13 provides the conditional percent correct classification for the mGLR procedures when a single cutscore was used. The cutscore for the single-cutscore conditions was placed at the theta value of 1.0. For the conditions with the 40 item maximum test length, the percent correct classification at the cutscore for the ability-based procedure is 21.2% while the cutscore-based procedure results in 45.1% correct classification. The mGLR has the poorest performance at the cutscore but exhibits similar PCCs to the other procedures at the remaining theta values. The overall accuracy of classifications for the cutscore-based method is 95.3% while the ability-based method yields 92.9% correct classification.

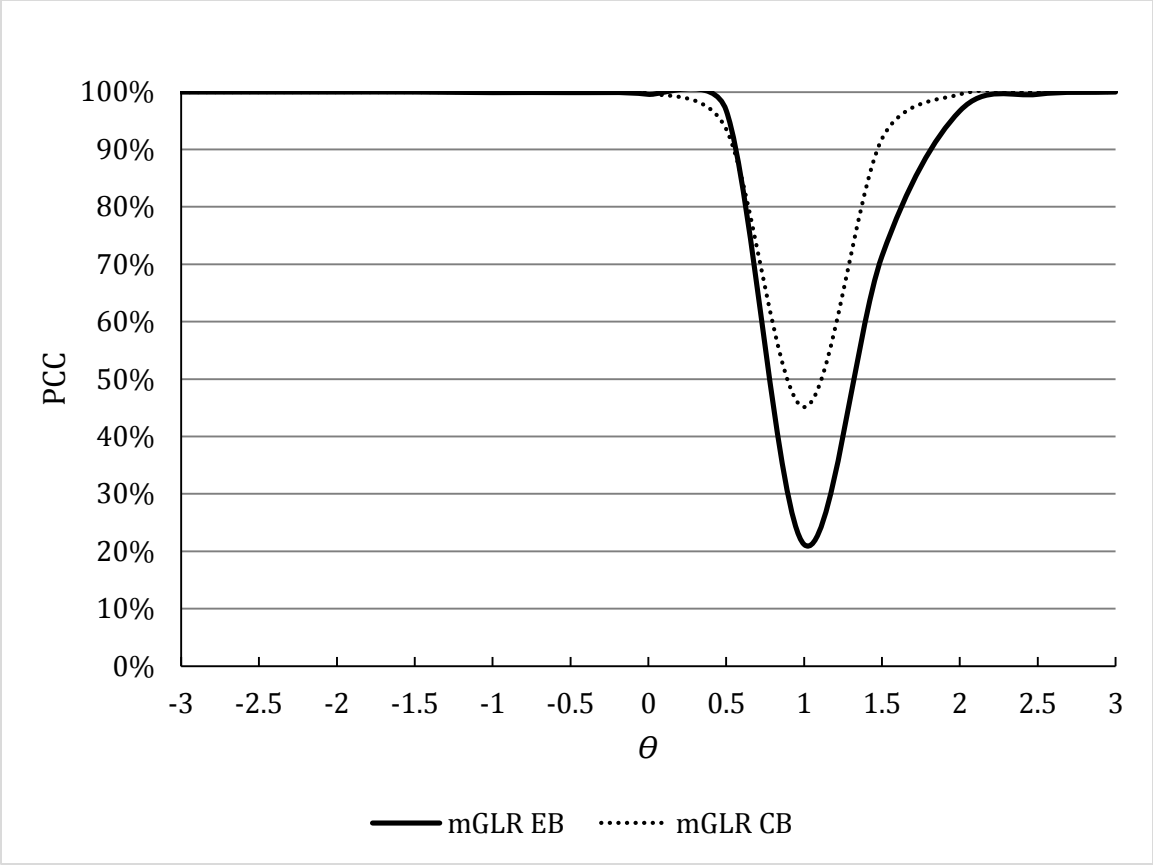
Similarly, for the conditions with the 60 item maximum test length, the PCC at the cutscore is 21.9% for the ability-based item selection while the cutscore-based procedure is 40.8%. Here again, the ability-based procedure performed the poorest at the cutscore but

provided similar PCC results across the remainder of the theta scale. Figures 25 presents the conditional PCC plots for the procedures using a single cutscore with a 40 item maximum test length while Figure 26 presents the plots for the 60 item maximum test length.

Table 13: Conditional percent correct classification for the single-cutscore conditions.

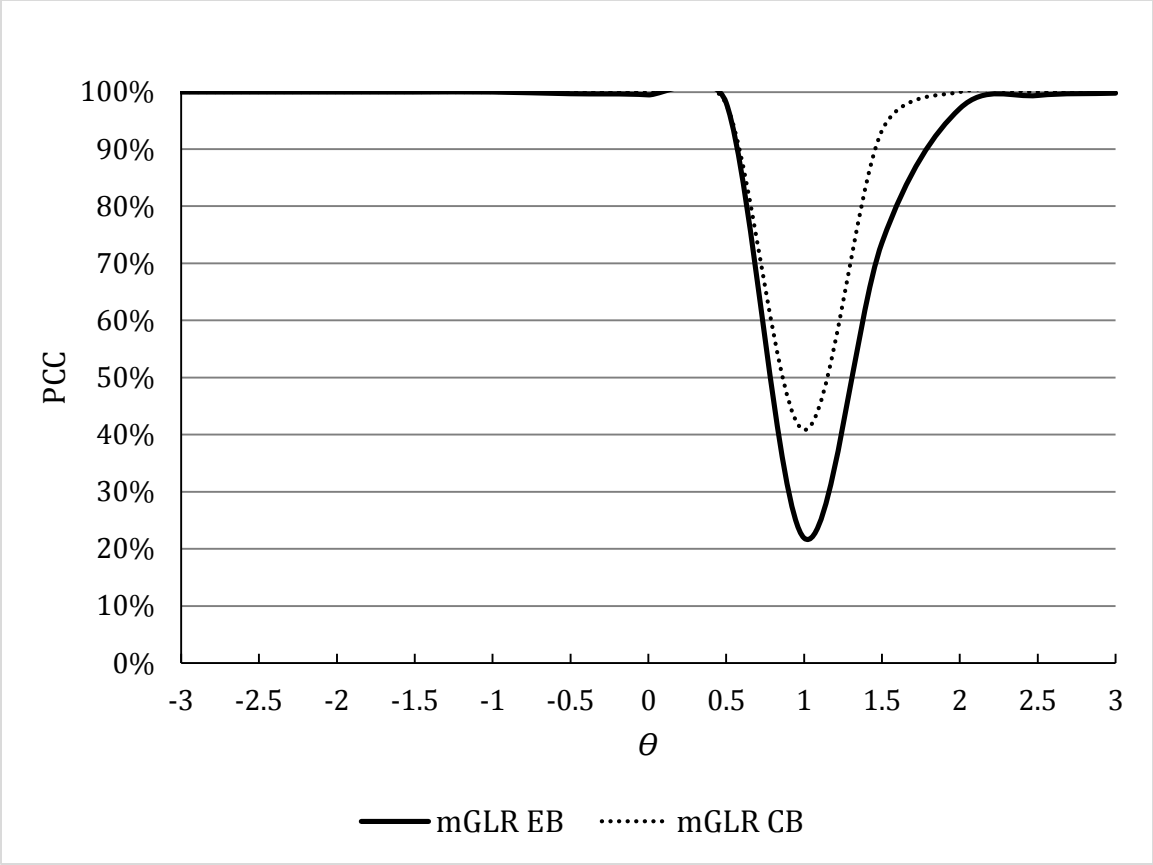
	40 Item Maximum		60 Item Maximum	
	mGLR with cutscore-based item selection	mGLR with ability-based item selection	mGLR with cutscore-based item selection	mGLR with ability-based item selection
Theta	Percent Correct Classification		Percent Correct Classification	
-3.0	100%	100%	100%	100%
-2.5	100%	100%	100%	100%
-2.0	100%	100%	100%	100%
-1.5	100%	100%	100%	100%
-1.0	100%	99.9%	100%	100%
-0.5	100%	99.9%	100%	99.7%
0.0	99.8%	99.6%	99.9%	99.5%
0.5	93.6%	96.8%	97.9%	98.1%
1.0	45.1%	21.2%	40.8%	21.9%
1.5	91.8%	71.4%	93.2%	73.6%
2.0	99.6%	96.7%	100%	97.1%
2.5	100%	99.6%	100%	99.4%
3.0	100%	100%	100%	99.8%
Overall	95.3%	92.9%	95.9%	93.3%

mGLR = Modified Generalized Likelihood Ratio.



mGLR EB = Modified Generalized Likelihood Ratio Estimate-Based item selection; mGLR CB = Modified Generalized Likelihood Ratio Cutscore-Based item selection.

Figure 37: Conditional percent correct classification (PCC) for the single-cutscore 40 item maximum test length conditions using the mGLR with multiple item selection methods.



mGLR EB = Modified Generalized Likelihood Ratio Estimate-Based item selection; mGLR CB = Modified Generalized Likelihood Ratio Cutscore-Based item selection.

Figure 38: Conditional percent correct classification (PCC) for the single-cutscore 60 item maximum test length conditions using the mGLR with multiple item selection methods.

Table 14 provides the conditional percent correct classification for the mGLR procedures when two cutscores were used. The lower cutscore for the two-cutscore conditions was placed 0.50 standard deviations below the peak of the test information function at the theta value of 0.50 while the upper cutscore for the two-cutscore conditions was placed 0.50 standard deviations above the peak of the test information function at the theta value of 1.50.

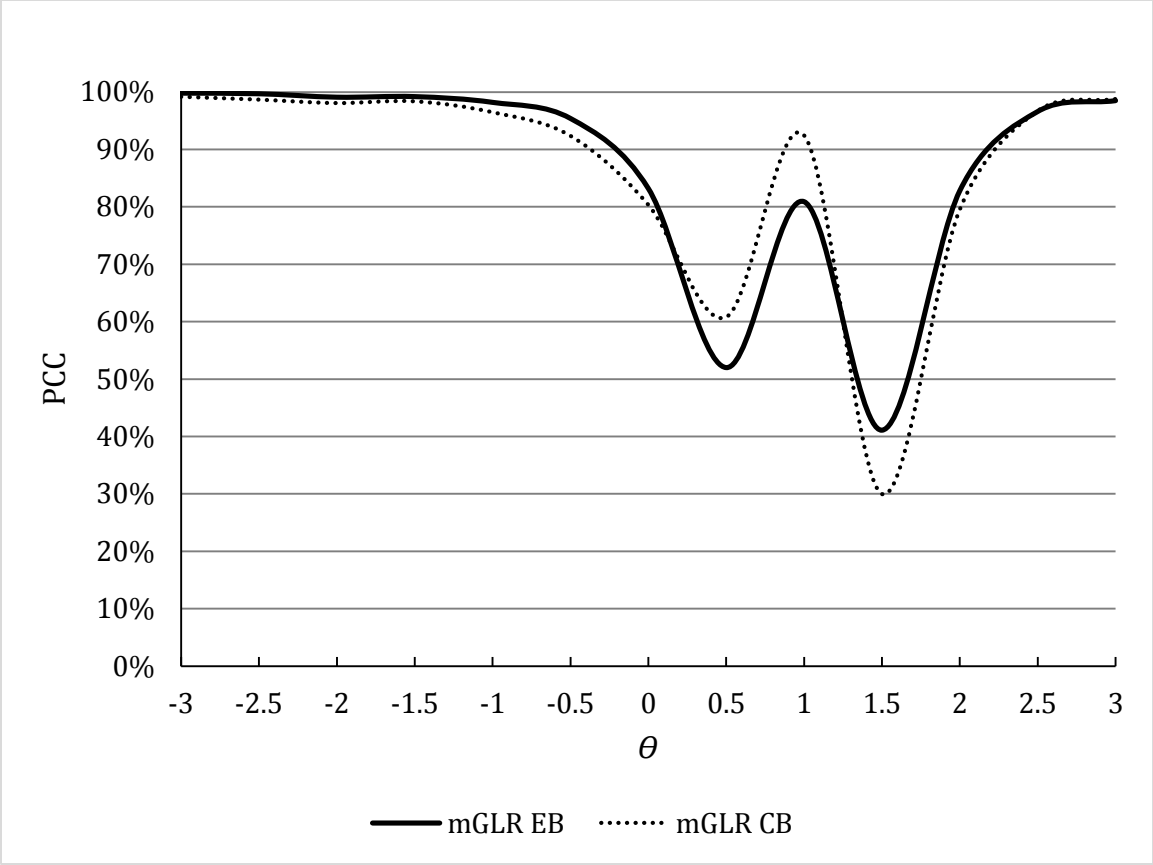
For the conditions with the 40 item maximum test length, at the lower cutscore the PCC is 52.0% for the ability-based procedure while the cutscore-based method correctly classifies 60.8%. At the upper cutscore the ability-based procedure, 41.1% correct classification, performs better than the cutscore-based procedure with 30.0% correct classification. Figure 39 shows that the cutscore-based method outperforms the ability-based method in PCC for the majority of the theta values between the cutscores.

For the conditions with the 60 item maximum test length, at the lower cutscore the PCC is 50.3% for the ability-based procedure while the cutscore-based method correctly classifies 60.1%. At the upper cutscore the ability-based procedure, 35.3% correct classification, performs better than the cutscore-based procedure with 25.4% correct classification. In terms of PCC, Figure 40 shows that the cutscore-based method outperforms the ability-based method for the majority of the theta values between the cutscores.

Table 14: Conditional percent correct classification for the two-cutscore conditions.

	40 Item Maximum		60 Item Maximum	
	mGLR with cutscore-based item selection	mGLR with ability-based item selection	mGLR with cutscore-based item selection	mGLR with ability-based item selection
Theta	Percent Correct Classification		Percent Correct Classification	
-3.0	99.2%	99.8%	99.1%	99.4%
-2.5	98.7%	99.7%	99.0%	99.4%
-2.0	98.1%	99.1%	98.5%	99.7%
-1.5	98.4%	99.2%	97.7%	99.0%
-1.0	96.5%	98.2%	97.9%	97.3%
-0.5	92.4%	95.4%	94.1%	95.8%
0.0	80.4%	83.2%	83.9%	86.0%
0.5	60.8%	52.0%	60.1%	50.3%
1.0	92.5%	80.9%	96.5%	82.8%
1.5	30.0%	41.1%	25.4%	35.3%
2.0	79.5%	82.8%	81.2%	84.6%
2.5	96.7%	96.6%	97.2%	96.3%
3.0	98.8%	98.5%	99.1%	99.3%
Overall	87.7%	87.2%	88.7%	88.1%

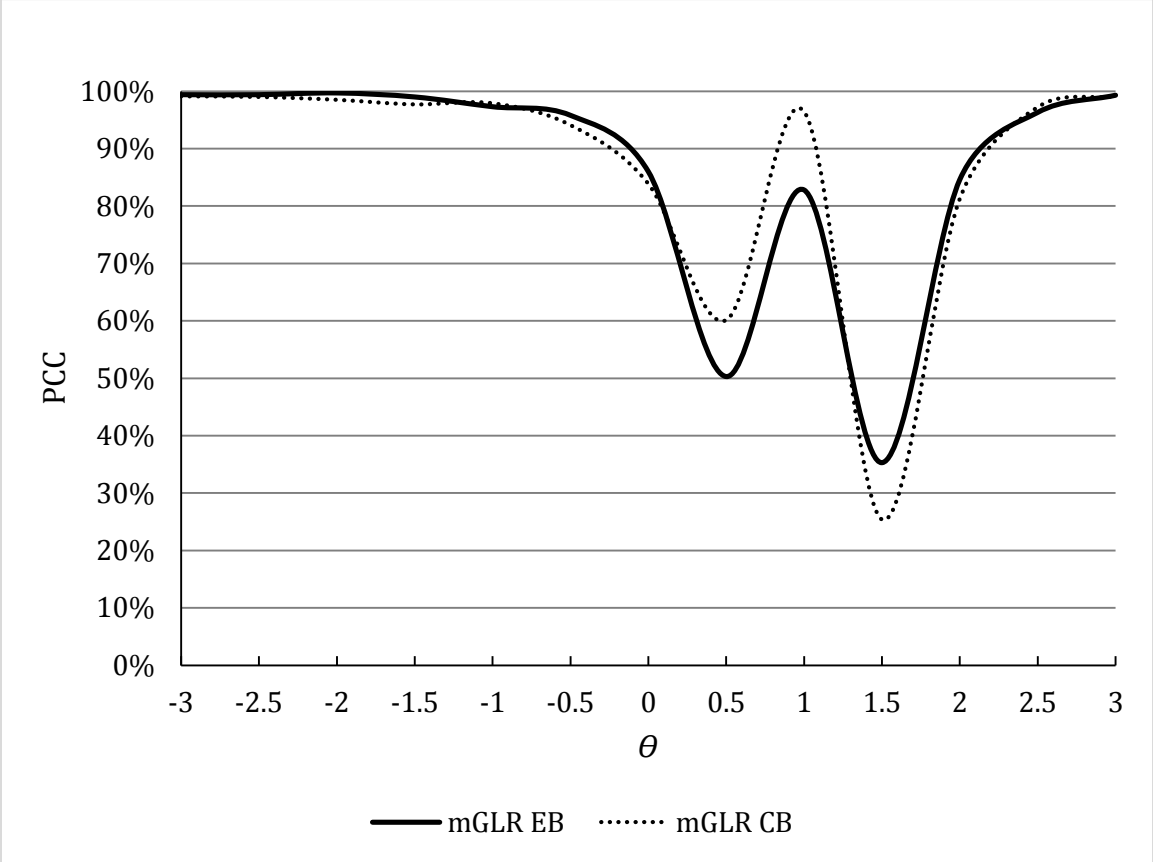
mGLR = Modified Generalized Likelihood Ratio.



mGLR EB = Modified Generalized Likelihood Ratio Estimate-Based item selection;

mGLR CB = Modified Generalized Likelihood Ratio Cutscore-Based item selection.

Figure 39: Conditional percent correct classification (PCC) for the two-cutscore 40 item maximum test length conditions using the mGLR with multiple item selection methods.



mGLR EB = Modified Generalized Likelihood Ratio Estimate-Based item selection; mGLR CB = Modified Generalized Likelihood Ratio Cutscore-Based item selection.

Figure 40: Conditional percent correct classification (PCC) for the two-cutscore 60 item maximum test length conditions using the mGLR with multiple item selection methods.

Table 15 presents the conditional PCC for the mGLR procedures when three cutscores were used. The lowest cutscore for the three-cutscore conditions was placed 0.50 standard deviations below the peak of the test information function at the theta value of 0.50. The middle cutscore was placed at the peak of the test information function at the theta value of 1.00. The highest cutscore was placed 0.50 standard deviations above the peak of the test information function at the theta value of 1.50.

For the conditions with the 40 item maximum test length, the PCC at the lowest cutscore is 38.3% for the ability-based procedure while the cutscore-based method correctly classified 53.5%. The PCC at the middle cutscore is 44.1% for the ability-based procedure while the PCC for the cutscore-based procedure is 45.9%. At the upper cutscore the PCC for the ability-based procedure, 39.8%, is better than the cutscore-based procedure with 28.4% correct classification. The overall PCC for the cutscore-based method is 82.9% while the overall PCC for the ability-based method is 83.0%. Figure 41 shows that the cutscore-based method outperforms the ability-based method in PCC for the theta values between the lowest cutscore and the middle cutscore while the ability-based procedure performes better for the theta values between the middle cutscore and the highest cutscore.

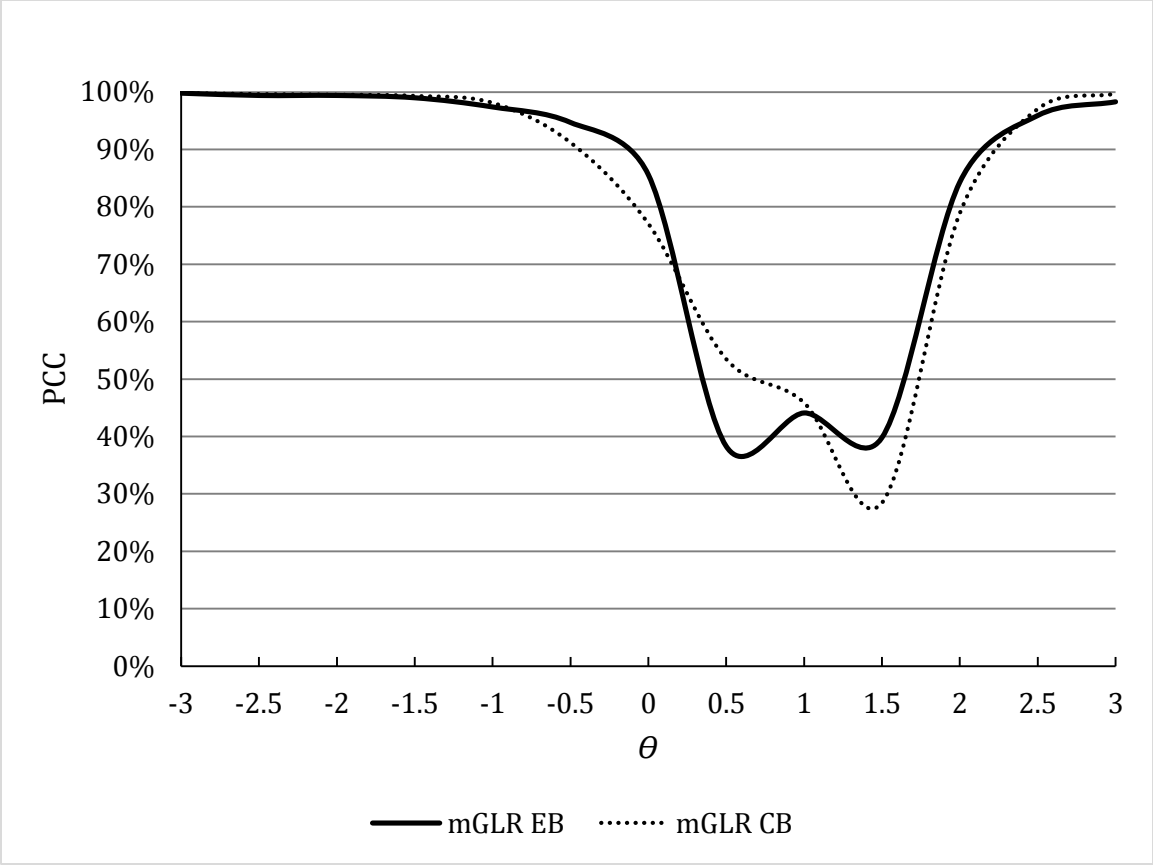
For the conditions with the 60 item maximum test length, the PCC at the lowest cutscore for the ability-based procedure is 38.0% while the cutscore-based method correctly classified 58.5%. The PCC at the middle cutscore is 44.9% for the ability-based procedure while the PCC for the cutscore-based procedure is 46.5%. At the upper cutscore the PCC for the ability-based

procedure, 38.8%, is better than the cutscore-based procedure with 21.4% correct classification. The overall PCC for the cutscore-based method is 84.0% while the PCC for the ability-based method is 84.2%. Figure 42 shows that the cutscore-based method outperformed the ability-based method in PCC for the theta values between the lowest cutscore and the middle cutscore while the ability-based procedure performs better for the theta values between the middle cutscore and the highest cutscore.

Table 15: Conditional percent correct classification for the three-cutscore conditions.

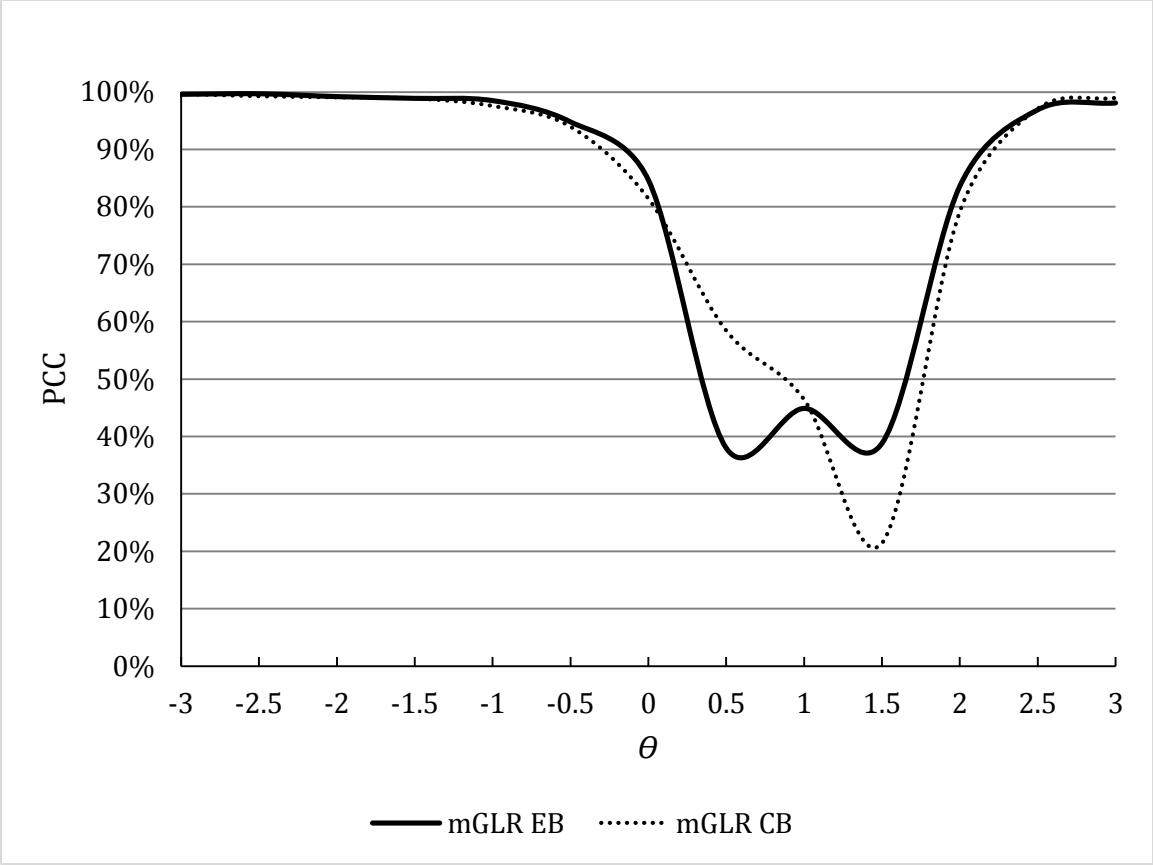
	40 Item Maximum		60 Item Maximum	
	mGLR with cutscore-based item selection	mGLR with ability-based item selection	mGLR with cutscore-based item selection	mGLR with ability-based item selection
Theta	Percent Correct Classification		Percent Correct Classification	
-3.0	99.8%	99.8%	99.7%	99.6%
-2.5	99.6%	99.4%	99.3%	99.7%
-2.0	99.5%	99.4%	99.1%	99.2%
-1.5	99.3%	99.0%	98.9%	98.9%
-1.0	98.1%	97.4%	97.6%	98.5%
-0.5	91.2%	94.7%	94.0%	94.8%
0.0	77.1%	85.6%	81.5%	84.7%
0.5	53.5%	38.3%	58.5%	38.0%
1.0	45.9%	44.1%	46.5%	44.9%
1.5	28.4%	39.8%	21.4%	38.8%
2.0	78.8%	84.3%	79.1%	83.6%
2.5	97.0%	95.9%	97.1%	96.9%
3.0	99.7%	98.3%	99.0%	98.1%
Overall	82.9%	83.0%	84.0%	84.2%

mGLR = Modified Generalized Likelihood Ratio.



mGLR EB = Modified Generalized Likelihood Ratio Estimate-Based item selection; mGLR CB = Modified Generalized Likelihood Ratio Cutscore-Based item selection.

Figure 41: Conditional percent correct classification (PCC) for the three-cutscore 40 item maximum test length conditions using the mGLR with multiple item selection methods.



mGLR EB = Modified Generalized Likelihood Ratio Estimate-Based item selection; mGLR CB = Modified Generalized Likelihood Ratio Cutscore-Based item selection.

Figure 42: Conditional percent correct classification (PCC) for the three-cutscore 60 item maximum test length conditions using the mGLR with multiple item selection methods.

Bias

To examine the accuracy of the final ability estimates, conditional bias was calculated for the six study conditions in which items for the mGLR procedure were selected based on interim ability estimates. Figures 43, 44, and 45 present conditional bias plots for each condition. The figures are grouped by the number of cutscores which were used by the conditions for both 40 and 60 item maximum test lengths. The three conditional bias plots appear seemingly identical with minor differences around the cutscores.

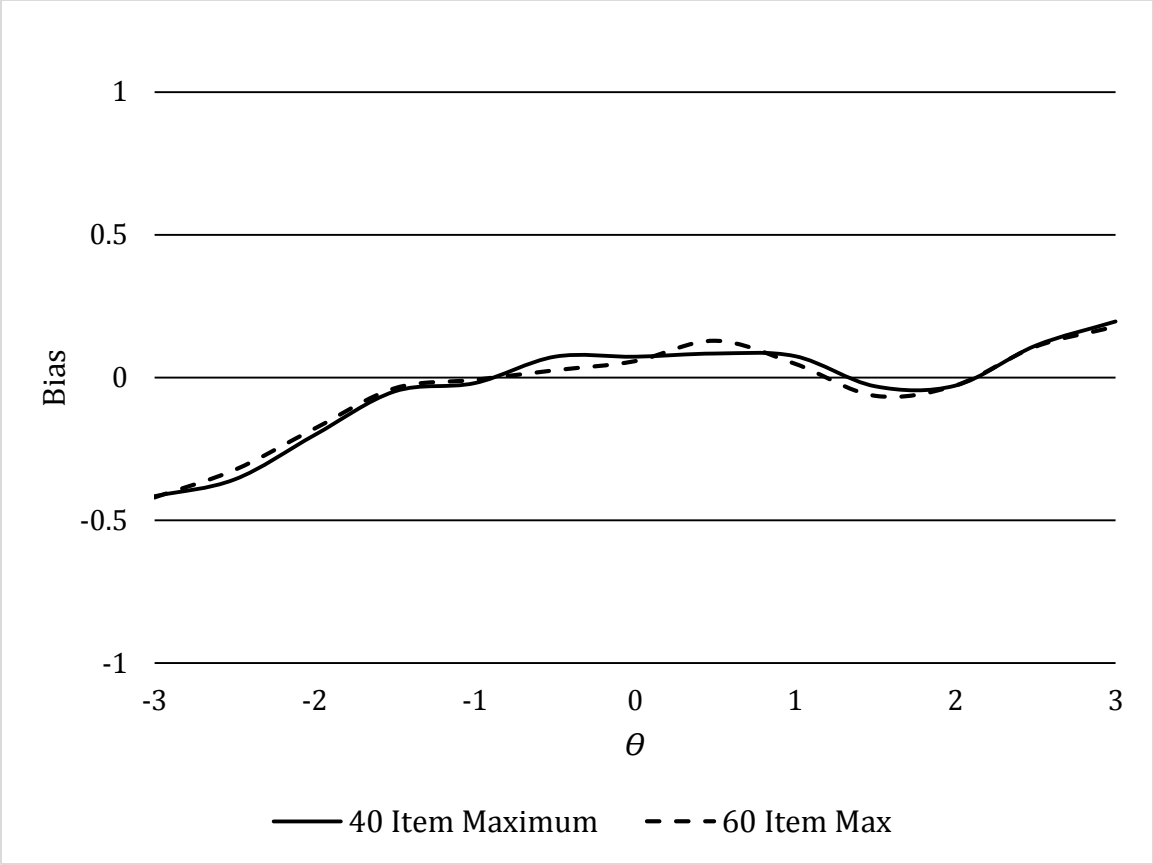


Figure 43: Conditional bias for the single-cutscore mGLR procedure using ability-based item selection conditions using 40 and 60 item maximum test lengths.

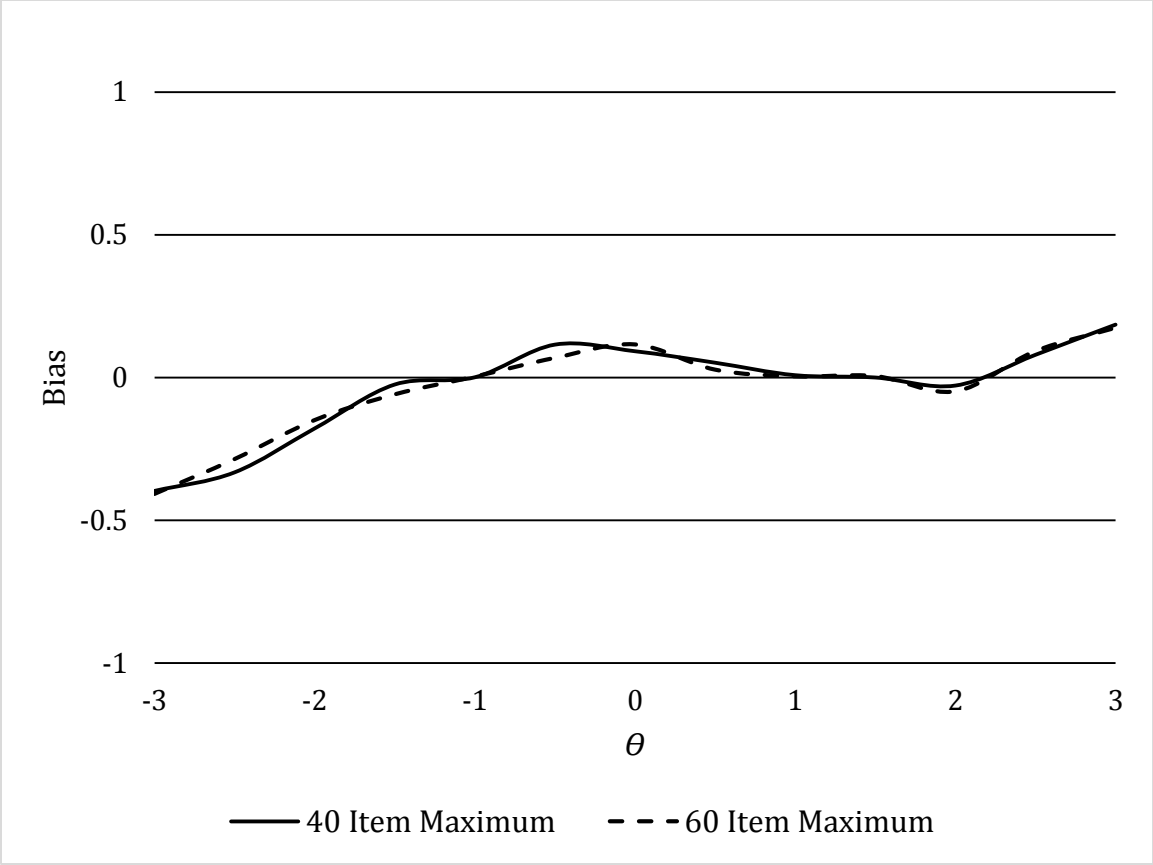


Figure 44: Conditional bias for the two-cutscore mGLR procedure using ability-based item selection conditions using 40 and 60 item maximum test lengths.

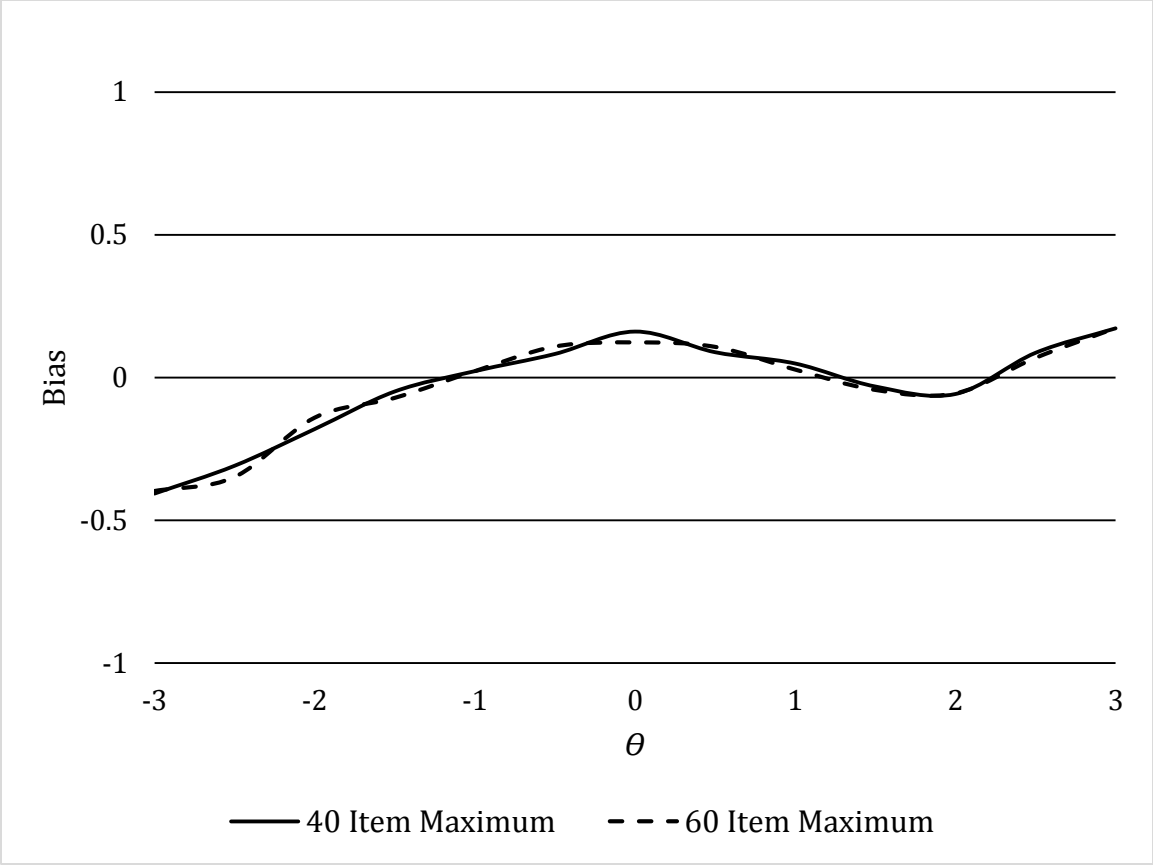


Figure 45: Conditional bias for the three-cutscore mGLR procedure using ability-based item selection conditions using 40 and 60 item maximum test lengths.

RMSE

Additionally, in order to examine the accuracy of the final ability estimates, conditional RMSE was calculated for the six study conditions in which items for the mGLR procedure were selected based on interim ability estimates. Figures 46, 47, and 48 present conditional RMSE plots for each condition. The figures are grouped by the number of cutscores which were used by the conditions for both 40 and 60 item maximum test lengths. Similar to the conditional bias plots, the three conditional RMSE plots are highly similar with only minor differences.

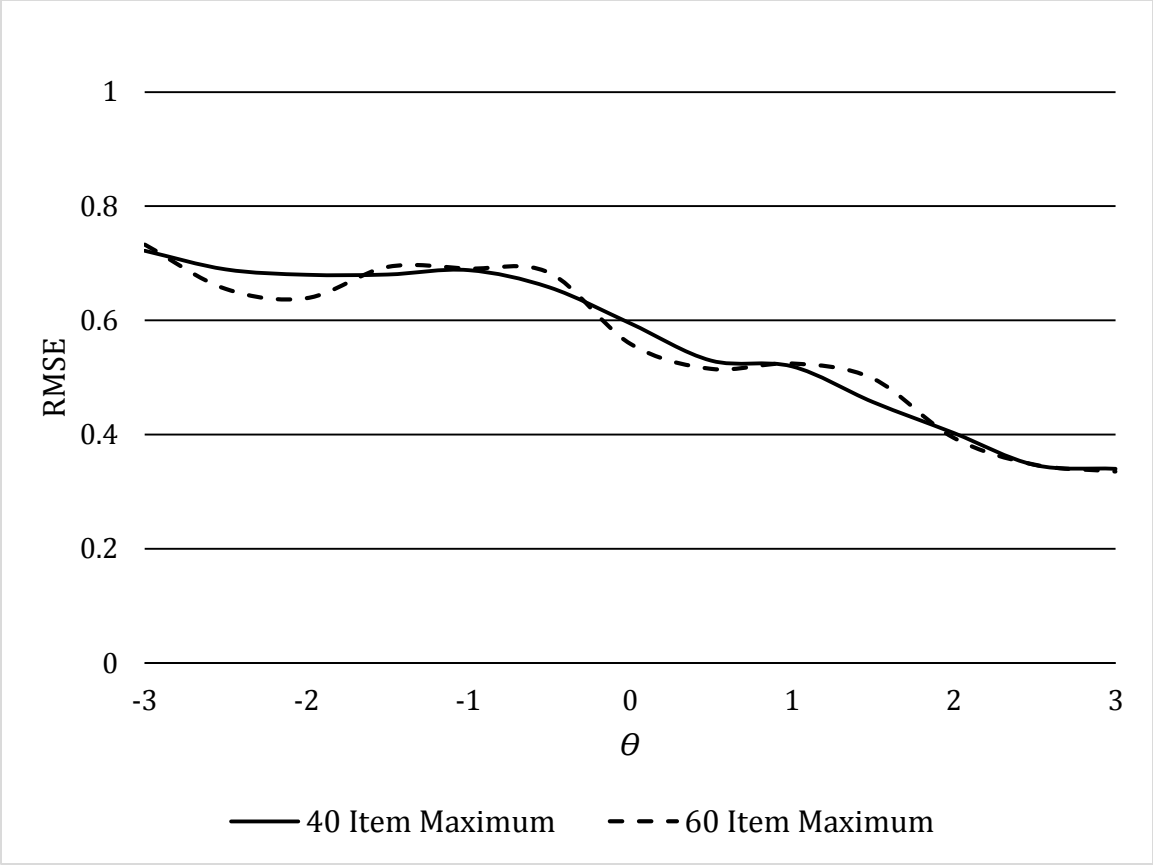


Figure 46: Conditional RMSE for the single-cutscore mGLR procedure using ability-based item selection conditions using 40 and 60 item maximum test lengths.

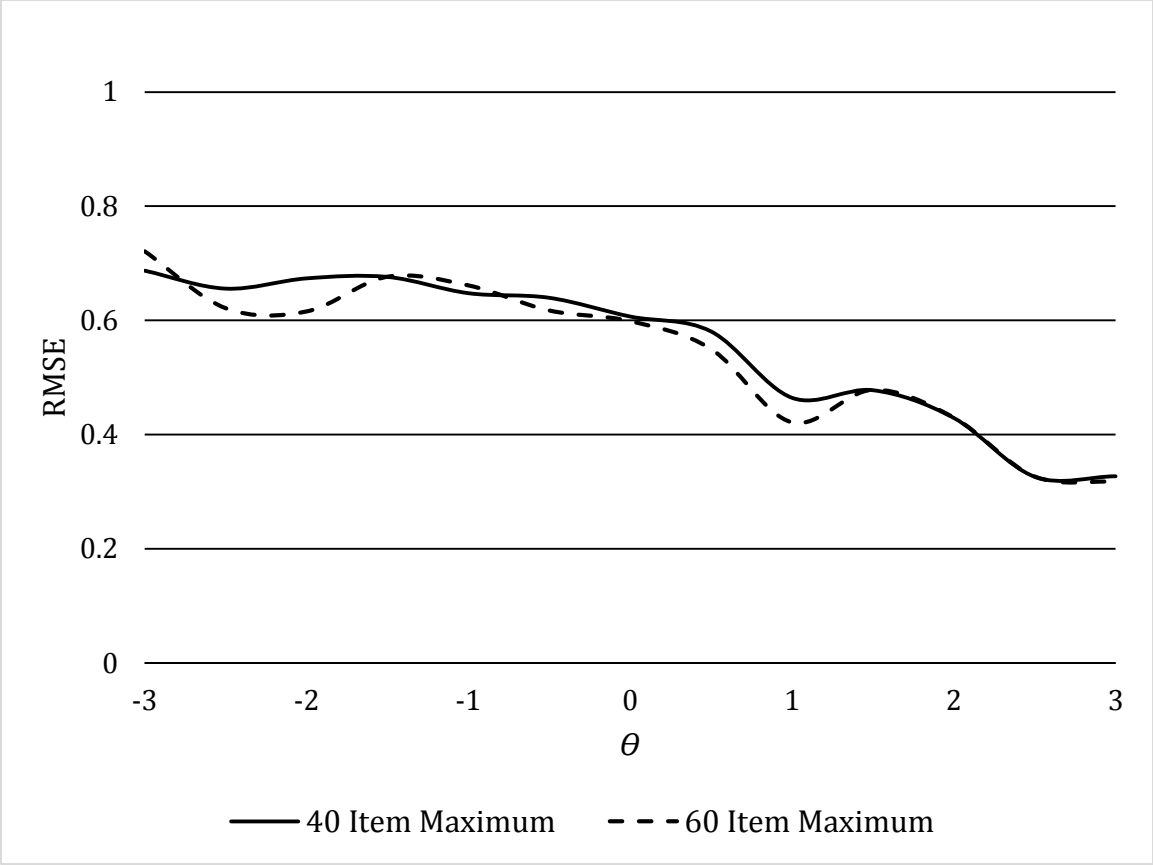


Figure 47: Conditional RMSE for the two-cutscore mGLR procedure using ability-based item selection conditions using 40 and 60 item maximum test lengths.

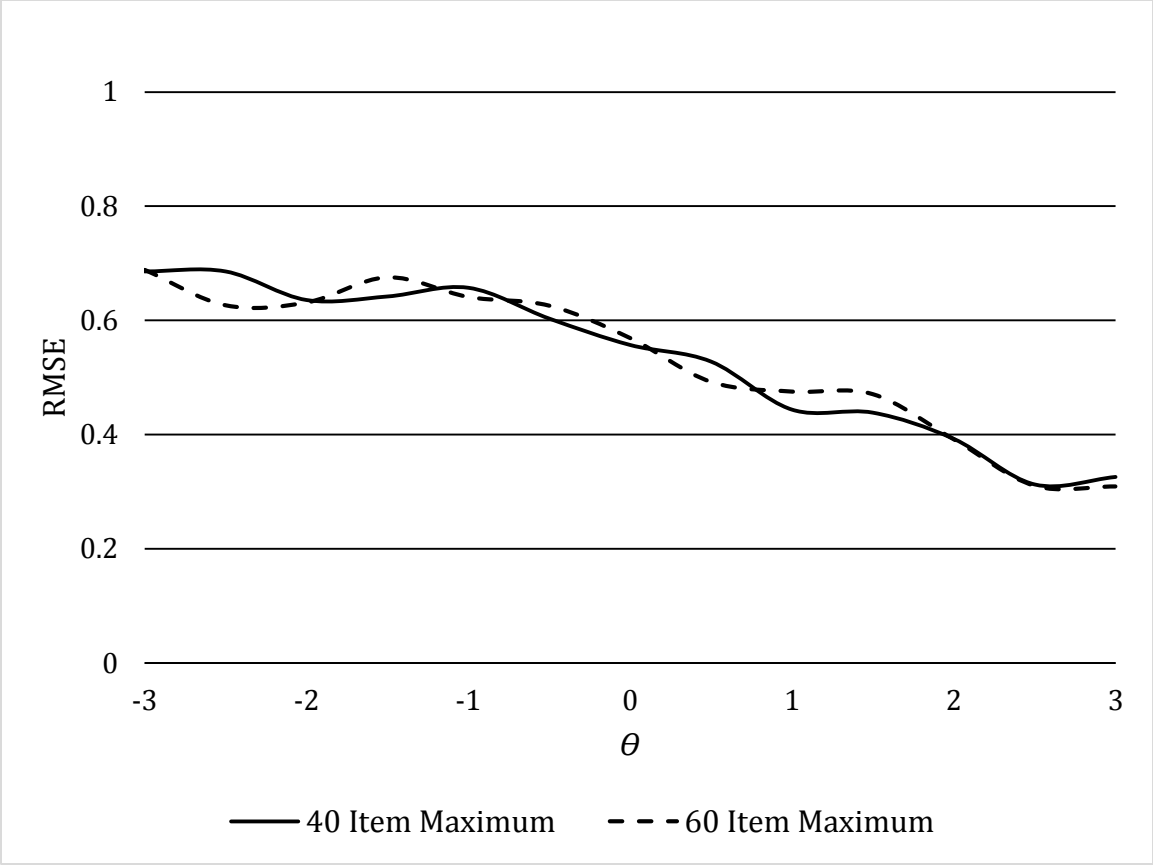


Figure 48: Conditional RMSE for the three-cutscore mGLR procedure using ability-based item selection conditions using 40 and 60 item maximum test lengths.

CHAPTER V: DISCUSSION

This study examined the functionality of three classification procedures under varying conditions. Three cutscore conditions (1, 2, or 3 cutscores) and two maximum test lengths conditions (40 or 60 items) were used to examine the ability of the classification procedures to succinctly and accurately classify examinees. An additional set of conditions were used to investigate the utility of implementing an ability-based item selection method with the mGLR procedure. All conditions were studied using conditional average test length for item efficiency and conditional percent correct classification for precision of classification. Conditional bias and conditional RMSE were calculated in order to evaluate final ability estimates produced by mGLR procedure which used ability-based item selection methodology. Based on real item parameters, a single data set was simulated with 1,000 simulees at each theta value ranging from -3.0 to 3.0 in discrete 0.10 logit increments resulting in 61,000 total simulees.

In the following sections each of the three research questions presented in the second chapter of this dissertation are specifically addressed. Following the discussion of the research questions, consideration is given to the application of this research in a practical setting. Finally, limitations of the current research and recommendations for future research are presented.

Research Questions

How do the three classification testing procedures, TSPRT, GLR, and mGLR, using cutscore-based item selection compare to each other in terms of average test length and percent correct classification in the context of multiple cutscore and test length conditions?

In general, the mGLR procedure using cutscore-based item selection produced conditional ATLs at the cutscores that are equal to or less than the ATL results produced by the TSPRT and GLR procedures. There are only five exceptions when the GLR procedure yielded a smaller ATL than the mGLR. In all of the instances in which the ATL for the GLR was better than mGLR, the modified procedure produced ATLs that were within 1.0 items of the ATL for the GLR. The results of the GLR procedures are consistent with previous research which demonstrated that the GLR was capable of producing shorter tests lengths than the TSPRT procedure (Thompson, 2007, 2009).

Conditional average test lengths at the cutscores for the TSPRT procedures range from 37.34 to 39.80 in the 40 item maximum test length conditions and from 48.86 to 56.65 in the 60 item maximum test length conditions. For the GLR procedure, maximum test lengths at the cutscores range from 36.29 to 38.93 in the 40 item maximum test length and from 47.34 to 54.12 in the 60 item maximum test length conditions. Finally for the mGLR procedure, ATL range from 34.31 to 38.16 for the 40 item maximum test length and from 41.97 to 50.23 in the 60 item maximum test length conditions.

Overall, in terms of PCC the mGLR procedure did not perform as well as the TSPRT and GLR procedures at all of the cutscores. When the mGLR selected items based on the cutscores the overall accuracy of classification for the mGLR was reasonably similar to the other procedures, within 1% to 2%, in most conditions. The greatest discrepancy in PCC between the mGLR and the other procedures was in the three-cutscore condition where the difference was 4%.

When examining the conditional PCC at the cutscores for each of the conditions, there is an unusual pattern of performance among the procedures. For the conditions using a single cutscore, the PCCs at the cutscores with the 40 item maximum test length conditions are slightly better than the results from the conditions with the 60 item maximum test length. In both instances the mGLR yielded lower results.

For the conditions using two cutscores, overall, the procedures performed reasonable similar. The PCC results for the lower cutscore indicate that the mGLR was most accurate, approximately 60% for both test lengths, while the GLR had the poorest results, 41.9% and 43.9%, for the two test lengths. At the upper cutscore, the mGLR performed poorer than the GLR. The mGLR procedure resulted in a PCC as low as 25.4% in the 60-item test length condition. For the theta values between the cutscores, the mGLR produced the highest PCC results, 92.5% and 96.5% for the 40- and 60-item test lengths, while the GLR produced the lowest PCC.

For the conditions using three cutscores, the overall PCC for each of the procedures range from 82.9% to 88%. The mGLR produced the poorest overall results, 82.9% and 84.0%, but outperformed the TSPRT and GLR procedures at the lower and middle cutscores for both test lengths. At the highest cutscore the GLR outperformed both the TSPRT and mGLR procedures with 53.7% and 49.8% for the 40- and 60-item maximum test length conditions. This unusual pattern of results, where there are sizable differences in classification accuracy across the cutscores but within a classification procedure, may be due to the major difference in items selected for administration.

Because each method has a unique scoring method and items are selected to maximize the information at the cutscore that is deemed to be closest to the examinee's ability, the procedures routed simulees through rather different item sets in the simulations. In the three cutscore conditions for example, where the TSPRT procedure tended to select multiple items from a cutscore before switching to a different cutscore for item selection, the GLR procedure would switch to the other cutscore having administered fewer items due to the more aggressive scoring method. Additionally, the mGLR procedure would switch between cutscores for item selection even more rapidly than the GLR procedure because of the nature of the mGLR scoring method. In other words, the more aggressive the scoring procedure, the more often item selection would switch between cutscores.

Additional Analyses

As some of the PCC results were unexpectedly poor, a set of brief additional analyses were performed to examine the issue. The additional analyses were performed for all classification procedures in the conditions where three cutscores were used and the maximum number of items was set to 60. Tables 16 through 19 display the results of the additional analyses. Each table shows the conditional classification results at each of the cutscores. Classification accuracy is typically lowest at the cutscores where it is expected that approximately 50% of simulees would be classified above and below the cutscore.

Tables 16 through 19 are included to display the proportion of examinee classification into a four-category classification system. Hence each row in the tables will sum to 100%. This analysis is helpful in understanding where misclassification was occurring in the simulations—whether the simulees were classified above or below the cutscore. Given this method of examining classification accuracy, all of the classification procedures appear to have achieved reasonable accuracy rates at each cutscore.

Table 16. Conditional classification percentages at cutscores for the TSPRT condition with 60 item maximum test length with three cutscores.

	TSPRT 60 Item Maximum Test Length using Three Cutscores			
	Classification Percentages at Cutscores			
Theta	Below Basic	Basic	Proficient	Advanced
0.5	45.0%	49.1%	5.9%	0.0%
1.0	3.3%	49.5%	45.8%	1.4%
1.5	0.0%	3.8%	60.9%	35.3%

Table 17. Conditional classification percentages at cutscores for the GLR condition with 60 item maximum test length with three cutscores.

	GLR 60 Item Maximum Test Length using Three Cutscores			
	Classification Percentages at Cutscores			
Theta	Below Basic	Basic	Proficient	Advanced
0.5	58.2%	35.0%	6.6%	0.2%
1.0	7.2%	43.2%	44.2%	5.4%
1.5	0.2%	7.2%	42.8%	49.8%

Table 18. Conditional classification percentages at cutscores for the cutscore-based mGLR condition with 60 item maximum test length with three cutscores.

	mGLR Cutscore-based Item Selection 60 Item Maximum Test Length using Three Cutscores			
	Classification Percentages at Cutscores			
Theta	Below Basic	Basic	Proficient	Advanced
0.5	32.2%	58.5%	9.3%	0.0%
1.0	2.5%	50.4%	46.5%	0.6%
1.5	0.0%	8.5%	70.1%	21.4%

Table 19. Conditional classification percentages at cutscores for the ability-based mGLR condition with 60 item maximum test length with three cutscores.

	mGLR using Ability-based Item Selection 60 Item Maximum Test Length using Three Cutscores			
	Classification Percentages at Cutscores			
Theta	Below Basic	Basic	Proficient	Advanced
0.5	50.8%	38.0%	10.2%	1.0%
1.0	10.5%	40.7%	44.9%	3.9%
1.5	2.6%	9.8%	48.8%	38.8%

How does the implementation of an ability-based item selection method with the mGLR procedure compare with the cutscore-based item selection mGLR procedure in terms of average test length and percent correct classification?

When comparing the mGLR procedure using cutscore-based item selection to the mGLR using ability-based item selection, the ability-based procedure produced shorter tests at all of the cutscores. More specifically, the smallest difference in ATL between the two mGLR procedures was in the two- and three-cutscore conditions where the difference in ATL was 4 items. The largest difference between the two mGLR procedures was in the single-cutscore condition where the maximum test length was 60 items with a difference in ATL of 16 items.

Conditional average test lengths at the cutscores for the mGLR procedures using cutscore-based item selection range from 34.31 to 38.16 in the 40 item maximum test length conditions and from 44.91 to 50.27 in the 60 item maximum test length conditions. For the mGLR procedures using ability-based item selection, maximum test lengths at the cutscores range from 28.45 to 34.37 in the 40 item maximum test length and from 34.38 to 41.68 in the 60 item maximum test length conditions. The mGLR using the ability-based item selection method resulted in lower ATLs at all cutscores in all conditions.

The results from comparing the two mGLR procedures indicate that the overall accuracy of classification does not always improve by implementing an ability-based item selection method. In the single-cutscore conditions the overall accuracy for both

procedures was above 90% for both test lengths. The conditional PCCs at the cutscore show that the cutscore-based item selection was superior to the ability-based method.

For the conditions using two cutscores, again, the overall PCCs for both procedures and both test lengths are highly similar. In fact, the ability-based method was 1% better than the cutscore-based method. The conditional PCCs at the lower cutscores indicate that the cutscore-based methods are more accurate with approximately 60% accuracy while the ability-based method produced approximately 50% accuracy. At the upper cutscores the ability-based methods are more accurate, 41.1% and 35.3%, compared to the cutscore-based methods with PCCs, 30.0% and 25.4%.

For the conditions using three cutscores, the two mGLR procedures yielded the same degree of accuracy overall. At the lowest cutscore the cutscore-based method was more accurate. The conditional PCC at the cutscore for the ability-based method resulted in approximately 38% correctly classified for both test length conditions while the cutscore-based method resulted in 53.5% and 58.5% for the 40 and 60 item maximum test length conditions. The PCCs for the middle cutscore were similar, but again, the cutscore-based method produced slightly improved results. The cutscore-based procedure PCCs are 45.9% and 46.5% whereas the PCCs for the ability-based method are 44.1% and 44.9%. Finally, the PCCs for the highest cutscores show the ability-based method yielded the best PCC results. The cutscore-based method produced PCCs of 28.4% and 21.4% while the ability-based method produced PCCs of 39.8% and 38.8%.

Again, the differences in PCC between the procedures may be due to the differences in the set of items which were selected for the simulees. In this case, the scoring for each item would have been the same which suggests that the item selection methods produced rather different sets of items which were administered to the simulees.

How well can ability levels be recovered as assessed using bias and root mean square error when an ability-based item selection method is implemented with the mGLR procedure?

Recall that the conditional bias plots, Figures 43-45, are virtually identical across test length and the multiple cutscore conditions. For the theta values ranging from approximately -1.5 to 2.0, the conditional bias values are very close to 0.0 for all conditions. The conditional bias results seem promising as the theta range that corresponds with the low bias values spans the theta values where the cutscores and accompanying indifference regions were placed. Similarly, the conditional RMSE plots, Figures 46-48, are exceptionally similar across test length and the multiple cutscore conditions. The conditional RMSE values for the theta values that span the region where the cutscore were placed are some of the lowest levels of RMSE produced.

Implications and Future Research

It has become commonplace for many high-stakes assessment programs to be delivered, or at least have an option to be delivered, through a computer-based platform. As previously mentioned, the computer-based delivery of an assessment provides some

greater degrees of flexibility for examinees in terms of location, testing windows, and personalized adaptability. As the movement to deliver assessments through computers is becoming a more widespread reality, such as with the initiatives of the PARCC and SBAC consortiums, there is also now an increased need to understand the capabilities and limitations of computer-based methodologies.

Results from this study expand the psychometric knowledge regarding the capacity of these methods for use in classification testing settings. Additionally this research provides a basis for future improvements and explorations of likelihood ratio based classification methods as the fundamental purpose of this study is to examine a newly proposed scoring procedure. The features in this study, such as multiple cutscores, test lengths, item selection methodologies, and ability estimation, are all variables and judgments which stakeholders and test designers would have to examine when developing an assessment.

Most importantly, this study demonstrated that the GLR and mGLR procedures were both capable of producing shorter tests than the TSPRT method with adequately similar classification accuracy in most of the single-cutscore and two-cutscore conditions. This was a key element of the study as the explicit purpose of the development of both procedures was to improve upon the original TSPRT method. While some of the conditional PCCs for some of the conditions were inappropriately low, this was the initial attempt to study the newly proposed mGLR procedure. It should be noted that all

procedures displayed poorer accuracy results as the number of cutscores increased. Future studies should give adequate consideration to item selection methodologies to improve the classification accuracy. Additionally, other dependent variables, such as adjacent classification and cutscore selection evaluations, could also be used to assess the feasibility of implementing one of classification procedures in an assessment program. Future studies should also give ample consideration to the number of and placement of the cutscores as well as the amount of item information that is available at each of the selected cutscores to ensure that classification decisions can be achieved. Hence, the item pool and cutscore selections need to be balanced through simulation studies to reach test expectations for efficiency and accuracy.

Next, this study demonstrated that by giving flexibility to the indifference region boundaries, the likelihood ratio based methods are able to use ability-based item selection methods. The flexibility of the indifference region boundaries enable the procedures to use each item more efficiently thereby reducing the number of items required to make a classification decision. By allowing items to be selected based on interim ability estimates, a final ability estimate can be established for tracking improvement over time or inform examinees of their ability relative to cutscores used to classify their performance on the assessment.

Two general limitations to this study are the item pool and thus the ability of these results to generalize to other item pools developed for classification purposes. The peak

of the test information function for the item pool used in this study corresponded with a theta value of 1.0. Cutscores were selected based on the available information to ensure that classification decisions were attainable before reaching the maximum test length. As other item pools may be used in future research or in applied settings, consideration should be given to how well the results from this study can generalize given the characteristics of the item pool.

In addition to the aforementioned future research options, researchers may consider using a more aggressive classification method such as the stochastic curtailment procedures suggested by Finkelman (2008, 2009). Research should also study the effects of varying parameters such as the indifference region widths and the allowable error rates, α and β. Finally, as the idea for the development of the mGLR procedure was conceived while researching polytomous IRT CAT methods, future research could investigate the efficacy of the mGLR using one the polytomous models.

Though the newly proposed mGLR procedure did not always achieve the expected similar levels of accuracy that the TSPRT and GLR procedures achieved, the mGLR procedure was able to reduce test length compared to the other two procedures. The mGLR yielded even shorter results when items were selected based on the simulee's interim ability estimates, but again, the accuracy rates were lower than anticipated. However, this study has provided an opportunity to examine how the use of more flexible testing parameters may improve the likelihood ratio-based classification method.

REFERENCES

American College Testing Program. (1993). COMPASS User's Guide. Iowa, IA: ACT.

Bartoff, J., Finkelman, M., & Lai, T. L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika, 73,* 473-486.

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

Bock, R., & Aitkin, M. (1981). Marginal maximum likelihood likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika,* 46, 443-459.

Bock, R., & Mislevy, R. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement,* 6, 431-444.

Chang, S. W. (1998). *A comparative study of item exposure control methods in a computerized setting.* Unpublished PhD Thesis, The University of Iowa, Iowa City.

Davis, L. L. (2002). *Strategies for controlling item exposure in computerized adaptive testing with polytomously scored items.* Unpublished doctoral dissertation, University of Texas, Austin.

De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York, NY: Guilford Press.

DeMars, C. (2010). *Item Response Theory.* New York: Oxford.

Dodd, B. G. (1990). The Effect of Item Selection Procedures and Stepsize on Computerized Adaptive Attitude Measurement Using the Rating Scale Model. *Applied Psychological Measurement,* 14: 355-366.

Eggen, T. J., & Straetmans, G. J. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60,* 713-734.

Eggen, T. J. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*, 249-261.

Eggen, T., & Straetmans, G. (1996). *Computerized adaptive testing for classifying examinees into three categories.* Arnhem: Cito.

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

Epstein, K. L., & Knerr, C. S. (1977). Applications of sequential testing procedures to performance testing. In D. J. (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference.* Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Ferguson, R. L. (1969). *Computer-assisted criterion-referenced measurement (Working Paper No. 49).* Pittsburgh, PA: University of Pittsburgh, Learning and Research Development Center. (ERIC No. ED 037 089).

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Education and Behavioral Statistics, 33 (4),* 442-463.

Finkelman, M. D. (2009). Variations on stochastic curtailment in sequential mastery testing. *Applied Psychological Measurement, 34 (1),* 27-45.

Geeorgiadou, E., Triantanfillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing develop from 1983 to 2005. *The Journal of Technology, Learning, and Assessment, 5,* 1-38.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and Applications.* Boston, MA: Kluwer-Nijhoff Publishing.

Huang, W. (2004). Stepwise likelihood ratio statistics in sequential studies. *Journal of the Royal Statistical Society, 66,* 401-409.

Kingsbury, G. G. & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), New horizons in testing (pp. 257-283). New York: Academic Press.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2 (4),* 359-375.

Koch, W. R., & Dodd, B. G. (1989). Procedures for selecting items from computerized adaptive tests. *Applied Measurement in Education, 2 (4),* 335-357.

Lau, C., & Wang, T. (2000, April). A new item selection procedure for mixed item type in computerized classification testing. *Paper presented at the 2000 AERA Annual Meeting in New Orleans, Louisiana*.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, 14, 367-386.

Lin, C. (2010). Item selection criteria with practical constraints for computerized classification testing. *Educational and Psychological Measurement*, 20-36.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph*, No. 7.

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

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, 23, 157-162.

Owen, R. J. (1969). *A Bayesian approach to tailored testing.* Princeton, NJ: Educational Testing Service.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York, NY: Springer.

Rasch, G. (1960). *Probablistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-254). New York, NY: Academic Press.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice, 8 (3),* 11-15.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35,* 311-327.

Sheehan, K., & Lewis, C. (1992). Computerized testing with nonequivalent testlets. *Applied Psychological Measurement,* 16, 65-76.

Spray, J. A. (1993). *Multiple-category classification using sequential probability ratio test.* ACT Research Report Series, 93-7.

Spray, J. A., & Reckase, M. D. (1987). *The effect of item parameter estimation error on decisions made using the sequential probability ratio test* (ACT Research Report Series No. 87-17). Iowa City, IA: American College Testing.

Spray, J. A., & Reckase, M. D. (1994). *The selection of test items for decision making with a computerized adaptive test.* Paper presented at the Annual Meeting of the National Council for Measurement in Education (New Orleans, LA, April 5-7, 1994).

Spray, J., & Reckase, M. (1996, Winter). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics*, Vol. 21, No. 4, pp. 405-414.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools.* Educational Testing Services (ETS).

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *In Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Centre.

Thissen, D., & Mislevy, R. J. (2000). Testing Algorithms. In H. Wainer, *Computerized adaptive testing: A primer (2nd ed.)* (pp. 101-133). Mahwah, NH: Lawrence Erlbaum Associates.

Thompson, N. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement*, 778-793.

Thompson, N. (2010, June 7-9). Nominal error rates in computerized classification testing. *Paper presented at the First Annual Conference of the International Association for Computerized Adaptive Testing*. Arnhem, Netherlands.

Thompson, N. (2011). Likelihood Ratio Based Computerized Classification Testing. *Paper presented at the 2011 conference of the National Council for Measurement in Education*. New Orleans, LA.

Thompson, N. A. (2007). *A comparison of two methods of polytomous computerized classification testing for multiple cutscores.* Unpublished doctoral dissertation, University of Minnesota, Twin Cities.

Thompson, N. A. (2009). Utilizing the generalized likelihood ratio as a termination criterion. *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.*

van der Linden, W. J., & Glas, C. A. W. (Eds.) (2010). Elements of adaptive testing. New York: Springer.

Vos, H. J. (1999). *Applications of Bayesian decision theory to sequential mastery testing.* Journal of Educational and Behavioral Statistics, 24, 271-292.

Wainer, H. (Ed.) (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27,* 1-14.

Wainer, H., Bradlow, E., & Wang, X. (2007). *Testle Response Theory and Its Applications.* New York: Cambridge University Press.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized Adaptive Testing.* Hillsdale, NJ: Erlbaum.

Wald, A. (1947). *Sequential analysis.* New York: John Wiley.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educaitonal Measuremnt: Issues and Practice, 17 (4),* 17-27.

Way, W. D. (2006). *Practical questions in introducing computerized adaptive testing to K-12 assessments.* Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/EC965AB8-EE70-46E5-B1A5-036BE41AB899/0/RR_05_03.pdf?WT.mc_id=TMRS_Practical_Questions_in_Introducing_Computerized

Weiss, D. J. (Ed.) (1983). New horizons in testing. New York: Academic Press.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361-375.

Whittaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement, 27*, 299-300.

Wouda, J. T., & Eggen, T. J. (2009). Computerized classification testing in more than two categories by using stochastic curtailment. *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.*

Vita

Samuel Heard Haring was born in Grenada, Mississippi the son of Peggye Jean Heard and grandson of Eugene and Donna Heard. He was raised in Texas completing all of his early education in the Wink Loving Independent School District culminating in graduation from Wink High School in 2000. Samuel served a two-year mission for the Church of Jesus Christ of Latter-day Saints in Anaheim, California from 2001 to 2003. In 2003, Samuel entered his freshman year at West Texas A&M with his younger brother, Bear, and graduated cum laude in May 2008. In August 2009, he entered the Graduate School at the University of Texas at Austin. He earned a Master of Education in educational psychology while seeking a Doctor of Philosophy in educational psychology. Samuel began working as an intern for Pearson in Austin, Texas in January of 2012. He began working full time for Pearson as an associate research scientist in March, 2014.

Email: Samuel.Haring@utexas.edu

This dissertation was typed by the author.