2-2012

An Empirical Examination of the Impact of Item Parameters on IRT Information Functions in Mixed Format Tests

Wai Yan Wendy Lam
University of Massachusetts Amherst, wylamw@gmail.com

AN EMPIRICAL EXAMINATION OF THE IMPACT OF ITEM PARAMETERS ON IRT INFORMATION FUNCTIONS IN MIXED FORMAT TESTS

A Dissertation Presented

by

WAI YAN WENDY LAM

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

February 2012

School of Education

AN EMPIRICAL EXAMINATION OF THE IMPACT OF ITEM PARAMETERS ON IRT INFORMATION FUNCTIONS IN MIXED FORMAT TESTS

A Dissertation Presented

By

WAI YAN WENDY LAM

Approved as to style and content by:

_____
Ronald K. Hambleton, Chair

_____
Craig S. Wells, Member

_____
Aline G. Sayer, Member

_____
Christine B. McCormick, Dean
School of Education

DEDICATION

To my husband and my little doggies.

ACKNOWLEDGMENTS

First, I would like to thank my committee members: Professors Ronald Hambleton, Craig Wells and Aline Sayer, for their support, patience and comments which guided me through my dissertation process. I would never have been able to finish my dissertation without the guidance from my committee members and support from my family and friends.

Ron, I still remember my first conversation with you during one of the technical meetings before joining REMP, thank you for the inspirational talk and giving me the opportunity to come join this excellent program! After I decided to come to UMass, you have shown me around town during a weekend, thank you very much for your kindness! I will be forever grateful for your patience and encouragement throughout my graduate studies. You always made yourself available to provide practical and technical support, no matter that it was late in the evening, on weekends or on holidays! Your passion and energy for psychometrics have inspired me to work harder each day. Craig, you are a great mentor and researcher. I really appreciate your advice, encouragement, friendship and expertise throughout my graduate studies and my dissertation work. Aline, I enjoyed very much being in your class, you are a great teacher and always fun to work with! I would also like to thank you for your willingness to spend some of your valuable time serving on my committee.

I would like to extend my gratitude to Professors Steve Sireci and Lisa Keller for their professional guidance and advice. Steve, I would like to thank you for your time, commitment and guidance on my first research paper presented at the NCME meeting, it was really an incredible experience! I am also forever grateful for all the help I received

from Peg Louraine throughout my years in REMP. I am very thankful for my friends in REMP who have helped me and encouraged me in various ways. Special thanks to Yue Zhao and Tie Liang for their friendship, support and encouragement; and to Hanwook Yoo for his friendship and tremendous help.

I would also like to thank my former colleague at Harcourt, Allen Lau, he was my first mentor in the psychometric field. Without his encouragement and support, I would not have had the courage to pursuit my dream.

Last but not least, I would like to thank my dearest husband, Albert, for his unconditional love and support, patience, encouragement and understanding. Whenever I felt down and wanted to give up, he was always there to cheer me up and make me feel I was invincible. I thank my parents and my grandmother who has passed away last year, for their love and support. My little dogs, Brooklyn and CiCi, also deserve big hugs from me, I thank them for staying up with me all the time when I was tired but needed to finish projects or assignments late at night. I dedicate my accomplishments to all of them.

ABSTRACT

AN EMPIRICAL EXAMINATION OF THE IMPACT OF ITEM PARAMETERS ON IRT INFORMATION FUNCTIONS IN MIXED FORMAT TESTS

FEBRUARY 2012

WAI YAN WENDY LAM, B.S., UNIVRSITY OF CALGARY, CANADA

Ed. D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Ronald K. Hambleton

IRT, also referred as "modern test theory", offers many advantages over CTT-based methods in test development. Specifically, an IRT information function has the capability to build a test that has the desired precision of measurement for any defined proficiency scale when a sufficient number of test items are available. This feature is extremely useful when the information is used for decision making, for instance, whether an examinee attain certain mastery level. Computerized adaptive testing (CAT) is one of the many examples using IRT information functions in test construction.

The purposes of this study were as follows: (1) to examine the consequences of improving the test quality through the addition of more discriminating items with different item formats; (2) to examine the effect of having a test where its difficulty does not align with the ability level of the intended population; (3) to investigate the change in decision consistency and decision accuracy; and (4) to understand changes in expected information when test quality is either improved or degraded, using both empirical and simulated data.

Main findings from the study were as follows: (1) increasing the discriminating power of any types of items generally increased the level of information; however,

sometimes it could bring adverse effect to the extreme ends of the ability continuum; (2) it was important to have more items that were targeted at the population of interest, otherwise, no matter how good the quality of the items may be, they were of less value in test development when they were not targeted to the distribution of candidate ability or at the cutscores; (3) decision consistency (DC), *Kappa* statistic, and decision accuracy (DA) increased with better quality items; (4) DC and *Kappa* were negatively affected when difficulty of the test did not match with the ability of the intended population; however, the effect was less severe if the test was easier than needed; (5) tests with more better quality items lowered false positive (FP) and false negative (FN) rate at the cutscores; (6) when test difficulty did not match with the ability of the target examinees, in general, both FP and FN rates increased; (7) polytomous items tended to yield more information than dichotomously scored items, regardless of the discriminating parameter and difficulty of the item; and (8) the more score categories an item had, the more information it could provide.

Findings from this thesis should help testing agencies and practitioners to have better understanding of the item parameters on item and test information functions. This understanding is crucial for the improvement of the item bank quality and ultimately on how to build better tests that could provide more accurate proficiency classifications. However, at the same time, item writers should be conscientious about the fact that the item information function is merely a statistical tool for building a good test, other criteria should also be considered, for example, content balancing and content validity.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Background

The idea of Item Response Theory (IRT) has been around for over half a century (Lord, 1952); however, only in the past thirty years did it achieve widespread popularity. The main reason for the delay is that IRT techniques require a lot more computational power than the classical test theory (CTT) method in test construction and scoring. In addition, there were no readily available and efficient computer software and affordable hardware for IRT analyses. Only recently, years after IRT computer software became available, for example, BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), MULTILOG (Thissen, 1991), PARSCALE (Muraki & Bock, 1997) and the dramatic improvements in computer performance, did full utilization of the IRT technology became possible, for instance, the Computer Adaptive Testing (CAT).

IRT is a statistical theory that links candidate abilities and candidate responses to the test items. Links between ability and item responses are made through non-linear models that are based upon assumptions that can always be checked. These assumptions are: test dimensionality and local independence, which will be discussed in more detail in the next chapter. The theory has been widely applied in different measurement fields (Hambleton, Swaminathan, & Rogers, 1991), to name a few, the GRE (Graduate Record Examination); the WISC (Wechsler Intelligence Scale for Children); NAEP (National Assessment of Educational Progress); and many credentialing exams.

IRT, also referred to as “modern test theory”, offers many advantages over CTT-based methods in test development and they are all well-documented in the literature (see

for example, Lord 1980; Hambleton & Swaminathan, 1985; Hambleton et al., 1991): item parameters are invariant over sample of examinees from the examinee population of interest; ability parameters are also invariant over samples of test items from the population of items measuring the ability of interest; an estimate of the amount of error in each ability estimate is available; probabilities of successful item performance for examinees located along the ability scale are available; and both items and abilities are referenced to the same reporting scale. Therefore, IRT provides a more useful and convenient framework for solving measurement problems, especially on test development, score equating, differential item functioning analysis, and ability estimation for individuals and estimation of their measurement error.

1.2 Statement of Problem

CAT has become the mainstream in the measurement field since the emergence of IRT and the immense improvement of computer power in the last few decades. Some examples of CAT include the GRE, GMAT (Graduate Management Admission Test) and the Microsoft Certified Professional exams. The most distinctive advantage of CAT compared to regular paper and pencil (P&P) test is that the difficulty of a CAT test is tailored to the examinee's ability so that test accuracy and reliability can be substantially improved. In addition, since the test is built to provide maximum information about the examinee's ability, test length and testing time can be reduced; but at the same time, the measurement precision is at least as good as the regular P&P test or sometimes even higher especially for those examinees at the extreme ends of the ability continuum (Lord, 1980). Therefore, in order to achieve full advantage of CAT, it is critical to understand the impact of item parameters on the item information functions (IIF) that maximize test

information functions (TIF) for each examinee, while also satisfying other test construction requirements, such as content balancing. In fact, selecting items with the most information to include in a test in order to maximize the test information at a particular level of ability estimate is the most widely used, and probably the oldest item selection algorithm in CAT. Of course, understanding the relationship between item parameters and item and test information functions is not only beneficial to the development of CAT, it also benefits regular P&P tests as tests constructed based on the IRT framework provide higher measurement precision without adding extra items.

Test construction under the IRT framework uses IIF and TIF to either build or evaluate tests for any desired set of test specifications. The procedure was outlined by Lord (1977). Steps for building a new test are as follows. The following procedure operates on a pool of items that have already been calibrated with an appropriate IRT model(s) so that item information curves are available for each item:

(1) Decide on the shape of the desired TIF. This was termed as the "target information function" by Lord (1977).

(2) Select items from the calibrated item pool with item information functions that will fill up the hard-to-fill areas under the target information function.

(3) Calculate the cumulative item information provided by those items that are already selected to include in the test (which is the interim test information function).

(4) Continue to select items until the test information function approximates the target information function to a satisfactory level.

Lord (1977) also provided a slightly different approach if one is interested in redesigning tests. Depending on the purpose of the test revision, the modification could come from eliminating some difficult items and replacing them with easier items if the goal is to increase the measurement precision for those at the lower ability continuum. If the goal is to increase the measurement precision at a certain cutscore, items that have a difficulty value that is close to the cutscore should be chosen to include in the test. After creating the modified tests, Lord suggested to compute test information functions for various hypothetical forms and compare their information with the original test. The ratio of the two curves (revised to the baseline or original) is called relative efficiency, which varies as a function of the ability level. The process continues until the TIF for the modified test becomes close enough to the result desired.

Regardless of the mode of testing, the effectiveness of a test depends on a number of important factors, for example: size of the item pool, quality of the items in the pool, content specifications of the test, content coverage of items in the pool, composition of different item formats in the test, item exposure controls, ability distribution of examinees, location of cutscores, choices of the IRT model(s) used to calibrate items in the pool, and the precision of the IRT item statistics. A number of questions related to item and test information function can then be generated from the implications of these variables and their interactions. For instance, instead of increasing the number of items in a test, suppose the discriminating powers of test items could be increased, how would the shape of TIF change in relation to the cutscores and what would be the impact on the TIFs if more discriminating items were substituted, and on the measurement precision of scores? In addition, the use of more discriminating items increases the effective lengths

of tests. What are the effective lengths of these tests if more discriminating items are included in the test? Increasing score precision (through increasing TIFs) has implications on the validity of performance category assignments. How much improvement as a function of test information gain can be expected? The idea here is that by replacing existing items with those with better quality, decision consistency and decision accuracy could be improved or test could be shortened but still have the same level of measurement precision.

1.3 Purpose of the Study and Educational Importance

The purposes of this study are: (1) to review what is already known about the relationship between item parameters and the item and test information functions; (2) to examine the consequences of improving the test quality through the addition of more discriminating items with different item format; (3) to examine the effect of having a test where its difficulty does not align with the ability level of the intended population; (4) to investigate the change in decision consistency and decision accuracy; and (5) to understand changes in expected information when test quality is either improved or degraded.

Based on the literature review, which will be presented in the next chapter, it is shown that the relationship between the item parameters and the item information functions are well-studied for dichotomously scored items over the years; however, the effects of item parameters on the level of information provided by polytomous items or on the overall test information based on mixed item format tests are not as obvious. Improving test quality by means of adding more discriminating items or systematically improving the discriminating power of items are one of many possible ways to influence

test information functions. In some other cases, a shift in test information function is made to increase the precision of scores in a particular region of the proficiency continuum or to improve the accuracy of classifications around one or more of the cutscores. For example, composition of the test items that were selected could be changed over time, such as replacing easier items with middle difficulty or harder items in order to produce tests that would provide better measurement precision for more capable examinees. One of the consequences of the above conditions is changes in the rate of decision consistency and accuracy in classifying examinees into different proficiency categories. Other consequences would be changes in the expected information when test quality is either improved or degraded.

Findings from this thesis would help testing agencies and practitioners to have better understanding of the item parameters on item and test information functions. This understanding is crucial for the improvement of the item bank quality and ultimately on how to build better tests that could provide more accurate proficiency classifications. For instance, when growth is taking place over several years, it is highly likely that TIFs that were needed earlier in the testing program are not centered at where they need to be now. Unfortunately, and perhaps surprisingly, many testing agencies are not very familiar with the relationship between item information, test information, and their impact on tests and the utility of tests for placing students in performance categories. In addition, results from this thesis could also provide useful information to test developers as different composition of item types could affect the location where the TIF would peak. For example, item parameters have different effects on item information for various item

formats, TIF might not necessarily peak in a region where difficulty of the test matches the target population when various item format are included in the test.

With the increasing demand for diagnostic information, test users are not satisfied with only a total score from a test, they would also like to know how they performed in a specific domain of a test. For example, a student and also the teacher would be interested to know how the student performed in the algebra section of the Mathematics test so that the teacher could plan for an appropriate remedial action. The reliability of this diagnostic information could be judged based on the examination of the test information function obtained from all the algebraic items in the test.

1.4 Outline of the Study

This thesis consists of five chapters. Background of the study, the importance of understanding the relationship between the item parameters and item and test information functions on the quality of tests, purposes of the study and educational importance have been described in this chapter already. Chapter 2 begins with an overview of IRT assumptions, a review of the common IRT models, item information functions and test information functions based on different IRT models, estimation of the measurement precision, and relative efficiency. A literature review of studies that are related to item and test information functions and also test characteristics that could affect test information which would in turn contribute to the classification accuracy are also included in this chapter. Chapter 3 describes the methods and procedures to conduct the study and also evaluation criteria of the results. Results will be summarized in Chapter 4. Finally in Chapter 5, conclusions and suggestions for future research are offered.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

This chapter begins with an introduction to the assumptions of the most commonly used Item Response Theory (IRT) model, then, description of some of the popular IRT models will be provided next. A brief discussion on how item information is obtained based on different models will follow. Review of the literature on how item parameters affect the amount of item information based on different IRT models will be described next. Finally, decision accuracy (DA) and decision consistency (DC) as a function of the test information function will be addressed.

2.2 Assumptions of the IRT Model

To properly implement an IRT model, several model assumptions should be met. One of the most fundamental assumptions is unidimensionality, which means only one ability trait is being measured by the items in the test. This assumption cannot be met strictly in reality because other factors also contribute to the test performance, for example, level of motivation or test anxiety; therefore, if the data exhibit a dominant component or factor, it is considered that the unidimensionality assumption is being met adequately (Hambleton et al., 1991). The second assumption is local independence. This assumption states that examinees' responses to any pair of items in the test are independent when examinees' abilities are held constant.

2.3 IRT Models

The following subsections provide a brief description of the dichotomous and polytomous IRT models

2.3.1 IRT Models for Dichotomous Response Data

The three most popular unidimensional IRT models for dichotomously scored response data are the one-, two-, and three-parameter logistic IRT models. Examinees' responses to this type of items are discretely scored so that they will receive a score of 1 when their answer is correct and a score of 0 when they provide an incorrect answer. Naming of these models is based on the number of item parameters incorporated in each model and the mathematical expressions of the three models are similar. The probability for a randomly chosen examinee with ability θ who answers item i correctly under the three-parameter logistic IRT model (3-PLM) is expressed in the following equation:

$$P(U_i = 1 \mid a_i, b_i, c_i, \theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad (1)$$

where U_i is the examinee's response to item i, and θ is the examinee's latent ability ranges from (-∞, +∞). a_i is the item discrimination parameter. Items with higher as are more effective in differentiating examinees into different ability levels than are items with lower as. Theoretically, a can range from (-∞, +∞), but negative discriminating items are usually discarded because items with negative a imply the probability of answering the item correctly decreases as examinee's ability increases. In addition, it is unusual to obtain item discrimination parameters higher than 2. Therefore, the usual range for a_i is (0, 2). b_i indicates the level of difficulty of an item. This parameter is also referred as the location parameter indicating the position of the item characteristic curve (ICC) in relation to the ability scale. Theoretical range of b also ranges from (-∞, +∞), but typically item difficulties are between (-4, +4). More difficult items are those with larger positive parameter values and easier items are those with negative parameter values. c_i is the pseudo-chance-level parameter. This parameter is the lower asymptote of

the ICC which represents the probability of those examinees with low ability level who answer the item correctly by chance. c_i is typically assumed to be smaller than the value that the examinees guessed randomly on an item. D is a scaling factor and by setting it to 1.7 will make the logistic function very similar to the normal ogive function. In fact, the difference between the logistic function and the normal ogive function will be less than .001 for all values of θ when $D = 1.7$.

The two-parameter logistic IRT model (2-PLM) is a constrained 3-PLM model in which c_i is assumed to be zero for all items. The mathematical expression for the two-parameter logistic IRT model (2-PLM) is expressed as:

$$P(U_i = 1 \mid a_i, b_i, \theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \tag{2}$$

The one-parameter logistic IRT model (1-PLM), often called the Rasch model, is the most restricted form of the 3-PLM in which items are assumed to be equally discriminating (i.e., all $a_i = 1$) and low ability examinees have zero probability of answering an item correctly (i.e., all $c_i = 0$). In this model, it is assumed that item difficulty (b_i) is the only factor that will have an impact on examinees' performance. Hence, the mathematical form is as follows:

$$P(U_i = 1 \mid b_i, \theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}} \tag{3}$$

2.3.2 IRT Models for Polytomous Response Data

With the increasing popularity of performance assessment, polytomous response items are widely used. These are items that can be scored in multiple score categories. Some examples of polytomous scoring items are constructed response questions and essay writings. The probability of an examinee reaching a specific score category can be

described in one of the following polytomous IRT models. These models are generalized from the dichotomous IRT models and reduced to the dichotomous IRT models when only two response categories exist.

The graded response model (Samejima, 1969) is an extension of Thurstone's method of ordinal intervals (Ostini & Nering, 2006). The model is built on the 2-PLM because this dichotomous model is used as the function to obtain the cumulative category response function (CCRF) and is denoted by the following equation:

$$P_{ix}^{*}(U_i \geq x \mid a_i, b_{ix}, \theta) = \frac{e^{Da_i(\theta - b_{ix})}}{1 + e^{Da_i(\theta - b_{ix})}} \tag{4}$$

where $P_{ix}^{*}(\theta)$ is the conditional probability that an examinee with ability level θ will obtain a score point of x or higher on item i. x is the possible item score point for a polytomous item i; therefore, x can range from (0, 1, …, m_i), and m_i is the highest possible score for item i. b_{ix} is the location parameter for score x, which is the point on the ability scale where $P_{ix}^{*}(\theta = b_{ix}) = .50$. The equation to obtain the score category response function (SCRF) for a specific score point is as follows:

$$P_{ix}(\theta) = P_{ix}^{*}(\theta) - P_{i(x+1)}^{*}(\theta) \tag{5}$$

Clearly, the probability of an examinee obtaining a score point greater than or equal to zero is one (i.e., $P_{i0}^{*}(U_i \geq 0 \mid a_i, b_{ix}, \theta) = 1$), and the probability of an examinee obtaining a score higher than the maximum score of the item is zero (i.e., $P_{i(m_i+1)}^{*}(U_i \geq (m_i + 1) \mid a_i, b_{ix}, \theta) = 0$).

Other polytomous response IRT models are also available. For example, the generalized partial credit model (Muraki, 1992) and the partial credit model (Masters, 1982). These models assume that each of the two adjacent categories (x and x-1) in a

polytomously scored item can be viewed as dichotomous case. The mathematical expression of the two models is given as:

$$P_{ix}(U_i = x \mid a_i, b_{ix}, \theta) = \frac{e^{Da_i \sum_{k=0}^{x} (\theta - b_{ik})}}{\sum_{h=0}^{m_i} e^{\sum_{k=0}^{h} Da_i (\theta - b_{ik})}} \quad (6)$$

where m_i is the number of score categories minus one; b_{ik} is the difficulty parameter associated with score category x; and a_i is the item discrimination parameter. The only difference between the two models is that the generalized partial credit model allows item discrimination parameters (i.e., a_i) to be different across items whereas the partial credit model assumes constant discrimination power across items.

2.4 Item and Test Information Functions

Both item and test information functions have an important role in test development and item evaluation. One of the attractive features of the item or test information function in IRT is that it allows test developers to better understand the contribution of each test item to the total test information and the consequences of selecting a particular item independently from other items in the test.

The mathematical form of the item information functions (IIF) for 3-PLM is expressed in the following equation:

$$I_i(\theta) = \frac{2.89 a_i^2 (1 - c_i)}{[c_i + e^{1.7 a_i (\theta - b_i)}][1 + e^{-1.7 a_i (\theta - b)}]^2} \quad (7)$$

Birnbaum (1968) showed that an item provides its maximum information at:

$$\theta_{\max} = b_i + \frac{1}{Da_i} \ln[0.5(1 + \sqrt{1 + 8c_i})] \quad (8)$$

In general, when $c_i > 0$, maximum information of an item occurs when θ is slightly bigger than b_i. However, when guessing is minimal, the item will give maximum information at b_i.

Items in a typical item bank can be easy or hard, and high or low in discrimination. An example of item and test information functions for some 3-PLM items is shown in Figure 2.1.

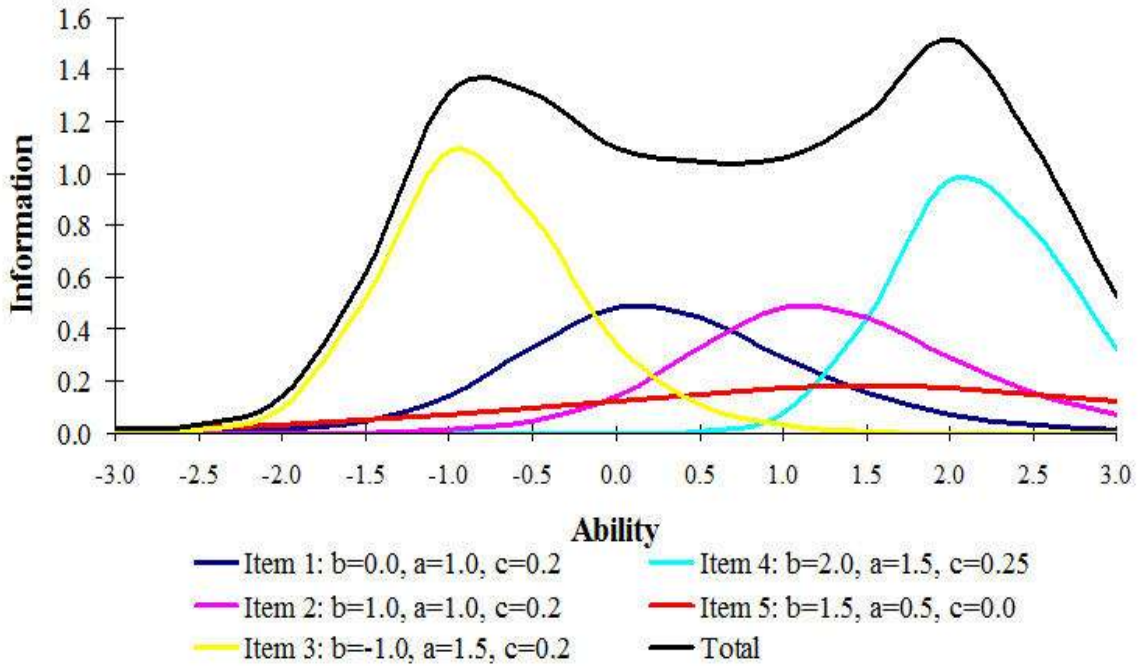


Figure 2.1 An example of item information functions for 3-parameter logistic IRT model (3-PLM) items (Adapted from Hambleton, 2006).

Information for a specific score categories based on the graded response polytomous IRT model is defined in the following fashion (Muraki & Bock, 1997):

$$I_{ix}(\theta) = D^2 a_i^2 \frac{[P^*_{ix}(\theta)[1-P^*_{ix}(\theta)] - P^*_{ix+1}(\theta)[1-P^*_{ix+1}(\theta)]]^2}{P_{ix}(\theta)} \qquad (9)$$

where $P^*_{ix}(\theta)$ is the CCRF as defined in Equation (4) and $P_{ix}(\theta)$ is the SCRF as defined in Equation (5). And the item information is obtained by:

$$I_i(\theta) = \sum_{x=0}^{m_i} I_{ix}(\theta) \tag{10}$$

For generalized partial credit and partial credit model, the item information function is defined as follows:

$$I_i(\theta) = a_i^2 D^2 [\sum_{x=0}^{m_i} x^2 P_{ix}(\theta) - [\sum_{x=0}^{m_i} x P_{ix}(\theta)]^2] \tag{11}$$

And here, $P_{ix}(\theta)$ is defined as in Equation (6).

Test information function (TIF) is a simple sum of the information functions for all items in a test (i.e., $I(\theta) = \sum_{i=1}^{n} I_i(\theta)$), which provides an overall impression of how much information a test is providing across the reporting scale. TIF is directly influenced by the statistics of the test items that are selected for the tests, and it provides an indication of the level of scores precision along the proficiency continuum (see, for example, Ackerman, 1989; Hambleton et al., 1991; Veerkamp & Berger, 1999). It should be noted that tests with more items are always going to have higher information than shorter tests. The more information a test provides at a score point on a reporting scale, the smaller the measurement error will be. In fact, the standard error of measurement at a score point on a reporting scale (called "conditional standard error of measurement" or simply "conditional standard error") is inversely related to the square root of the test information at that score point:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{12}$$

This result is true when *n*, the number of items in the test, is large (Samejima, 1977; Hambleton et al., 1991).

In classical measurement, a standard error of measurement (SEM) that corresponds to test score reliability of .90 is about .32 of the standard deviation of the test scores (since $SEM = \sqrt{1-r}$ (when the SD is set to a value of 1), where *r* is the reliability of the test score). In IRT analyses, with proficiency scores scaled to a reference group with a mean of 0.0 and a standard deviation of 1.0, .32 of the standard deviation of the proficiency scores would be .32 which corresponds to test information of about 10. This value of 10 is sometimes chosen as a target in test development, and a criterion for evaluating tests from an IRT perspective.

In order to evaluate the effectiveness of tests that were built on different items, test information functions could be compared between tests. This concept is called relative efficiency (Lord, 1977) and the formulation is given in the following:

$$RE(\theta) = \frac{I_A(\theta)}{I_B(\theta)} \tag{13}$$

where $I_A(\theta)$ and $I_B(\theta)$ are the information functions for Test *A* and Test *B*, respectively. For instance, if $I_A(\theta) = 11$ and $I_B(\theta) = 10$ and both are 50-item tests, then $RE(\theta) = 1.1$. It means that Test *A* is functioning 10% better than Test *B*. There are two different ways to proceed in order to make the two tests to achieve the same level of precision of θ estimates: Test *A* could be shortened by 10% and would still be able to produce θ estimates with the same amount of precision as those from Test *B*. Alternatively, Test *B* could be lengthened by 10% with items of similar properties that were already in the test so that the test will function as well as Test *A*. The same calculations can be repeated at the cutscores or proficiency score points of interest.

2.5 Studies on Item and Test Information

Effective use of test information functions requires a diverse and high quality item pool, IRT model that fits the data and item statistics estimated with good precision. When an IRT model fits the test data, the IRT model is said to provide invariant item and ability parameters. The invariance property implies sample-free item parameters and test-free ability parameters. This property of sample-free item estimates in IRT has a major benefit in test construction. For example, in many practical situations, a group of new items, often called "field test" items, are embedded in the test and administered to different groups of examinees for the development of item bank. When there are large number of new items need to be tested, it is not possible for all examinees in the target population to try out all new items. Usual practice is to build multiple forms with different field test items and administered to various groups of examinees. Since these try-out items are administered to different groups of examinees, using classical item indices obtained from the experimental items for test construction might not be appropriate for the intended population. Another drawback of test construction using classical theory is that tests cannot be built with a fixed measurement precision. Under classical theory, both characteristics of the item itself and also the relationship of an item with other items in the test contribute to the reliability of the test; therefore, it is not possible to isolate the relationship between the contribution of an item and the test reliability. Most importantly, IRT also put item difficulties and examinees' ability on the same scale so that it is possible to select items that are most useful in certain regions of the ability scale, for example, at a cut-off score between pass and fail region (Dodd & Koch, 1987; Hambleton et al., 1991).

It is well-known that in the context of the 3-PLM IRT model, when *a*-parameter (i.e., item discrimination) increases, it will generally lead to an increase in information (Hambleton & Jones, 1993; Veerkamp & Berger, 1999), hence more measurement precision results. The IIF would be more peaked and concentrated in a smaller range of the proficiency scores scale when *a*-parameter increases (see for example, Green, 1983; Hambleton et al., 1991; Wiberg, 2003). However, when *a* is not estimated with good precision, information will be lower than expected (Hambleton, Jones & Rogers, 1993; Hambleton & Jones, 1994). The *b*-parameter reflects the place on the proficiency scale where the item can provide the most information (Green, Yen & Burket, 1989; Hambleton et al., 1991; Veerkamp & Berger, 1999; Wiberg, 2003). When the variance of *b*-parameters in a test is high, it will tend to spread out the test information and hence lower the measurement precision (Lord, 1977; Luecht, 2006). The *a*-parameter and the difference between the *b*-parameter and the ability score also have an interaction effect on the information. As Hambleton and Jones (1994) and Veerkamp and Berger (1999) pointed out, in the logistic IRT model, items with highest *a*-parameter do not necessarily give maximum information. In fact, if b_i is not close to θ, extreme increase in a_i can lead to a decrease in item information. This effect is called the attenuation paradox (Loevinger, 1954) in IRT by Lord and Novick (1968, p. 368) and Birnbaum (1968, p. 465). Finally, the *c*-parameter reduces the discrimination power of an item and would also make an item appear to be slightly easier than the *b*-parameter might suggest (Hambleton & Cook, 1977; Samejima, 1984, as cited by Veerkamp and Berger, 1999; Wiberg, 2003). Therefore, a non-zero *c* parameter would always lower the measurement precision (Hambleton & Traub, 1971). Since the *c*-parameter tends to lower the

information functions, some researchers might incline to fit a one-parameter or two-parameter IRT model to the data instead. Although item information functions obtained from these models are higher, if these models do not fit the data, item information obtained from these models will generate misleading results (de Gruijter, 1986).

How individual item parameters (i.e., the discriminating parameter (a_i) and the difficulty parameter (b_i)), and more specifically, the step difficulties which is the difficulty parameter associated with score category x as mentioned in Equation (4) and (6) above would affect the item or test information function for the polytomous IRT models or mixed IRT models are not as obvious. Only a handful of research has been done to examine the relationship between the trait level and item category parameters for the polytomously scored items and its effect on item information function. Samejima (1976, 1977) claimed that items that are fitted with graded response model (GRM) produce higher information than dichotomous items. In her studies, Samejima demonstrated that polytomous scoring yielded considerably more IRT information than the optimal dichotomization of the same items. In addition, the problem of attenuation paradox is also improved (Samejima, 1969).

Based on data collected from the abbreviated version of the World Health Organization Quality of Life Survey (WHOQOL-BREF), Lin (2007) studied the information of these polytomous items by fitting the GRM to the data. She found that deleting items with low discriminating values have a greater impact on the information at the mid-range of the proficiency scale. However, information at both ends of the scale remained about the same.

Luo, Ouyang, Qi, Dai, and Ding (2008) examined the relationship between the test information function and the item discrimination and step difficulties for GRM based on simulated response data. They constructed tests with five 4-point items, with four different levels of *a*-parameters (0.5, 1.0, 1.5, 2.0) and step difficulties were categorized into five categories: "1" ($-3.0 \le b_{ik} < -1.8$), "2" ($-1.8 \le b_{ik} < -0.6$), "3" ($-0.6 \le b_{ik} < 0.6$), "4" ($0.6 \le b_{ik} < 1.8$), and "5" ($1.8 \le b_{ik} \le 3.0$). In their study, they were interested in the location of ability level where maximum information occurred for different response patterns when a_i is fixed. In addition, they also examined the location of ability level where maximum information occurred with different levels of a_i for various combinations of response patterns. Their results indicated that, similar to the dichotomous IRT model, increase in item discrimination power would lead to an increase in information. Their results also showed that no specific pattern could be observed from the item difficulty and the ability level in GRM when *a* was held constant. In other words, maximum information might not occur at a θ level that was close to the item difficulty. However, based on the examination of the response patterns from the simulated data, they concluded that: (1) TIF would peak in the ability region where majority of the scores came from, regardless of the a_i. For example, if the response pattern is 11113, TIF would peak in the ability region corresponds to category 1, which is between -3.0 to -1.8; (2) if the occurrences of different response categories are equal, location of the maximum information will depend on the magnitude between the remaining categories and the other two groups. In this scenario, maximum information will occur in the ability region closer to the remaining category. For instance, maximum information for response pattern of 11344 will occur in the ability region corresponds to

category 4, which is between 0.6 to 1.8; (3) no conclusion can be generated about the location of maximum information if the occurrences of different response categories are equal and same distances between the remaining response category with the other two (e.g.: 22344) or far away from the other two (e.g.: 15122).

Thissen (1976) compared the IRT information function between dichotomously scored items and IRT information obtained from Bock's nominal model (1972) and found that the nominal response model yielded substantially more information than the dichotomously scored items, particularly at the lower levels of the θ scale. In addition, incorrect responses contained useful information as well.

Masters (1988a) and Bejar (1977) noted that polytomously scored response data can provide more information about the examinees' ability scores; hence, more precise ability estimates can be obtained. In addition, Master (1988b) also pointed out that more detailed diagnostic information about the examinees and the items can be obtained from polytomously scored response items.

Donoghue (1994) studied the IRT information for three grade levels (Grade 4, 8 and 12) of NAEP Reading field tests, where items in each of the test were calibrated simultaneously using the 3-PLM IRT model on the multiple choice items, 2-PLM on short response items, and generalized partial credit model (GPCM) on 3-point extended response items. His results indicated that polytomous scoring items provided much more information than those from the short response items and multiple choice items. Specifically, item information obtained from the GPCM extended response item provided about 2.3 to 3.7 times more information than a typical multiple choice item; and this type of item also provided about 1.8 to 2.6 times more information than short response item.

Dodd and Koch (1985) studied the information functions for the partial credit model (PCM). They found that step difficulties for the PCM model have a major effect on item information functions: small distance between the first step and the last step of difficulties produced the most information within a narrow range of the θ continuum. On the other hand, when the distance between the first and last step of difficulties was big, information function would be more spread out and less peaked.

In another study, Dodd and Koch (1987) examined the effects of variations in item step values on item and test information in PCM based on three-step and four-step items. For the four-step items, the orderings of the step difficulty values of -1.0, -0.5, 0.5, and 1.0 were systematically varied to yield 24 items that only differed in terms of the ordering of step difficulty values. And for the three-step items, 6 items were built by varying the orderings of the step difficulty values of -1.0, 0.0 and 1.0. Simulated response data was used to evaluate the effects of adding or deleting items with specific step characteristic as a mean of test revision. Their results showed that the PCM item produced maximum information when: (1) the first step difficulty parameter was close to the last step difficulty parameter; (2) more step difficulties were out of sequential order and displaced at higher step levels, holding the distance between the first and last step difficulties constant; and (3) the magnitude of the distance between the steps that were out of sequential order. They also found that items with more score categories yielded more total information across the entire θ scale than fewer score categories.

In Cohen's (1983) study, he demonstrated that if reducing the polytomously scored response categories to the dichotomous level, it will lead to a systemic loss of information. In a similar context, Yamamoto and Kulick (1992) examined the amount of

information changed when scoring items that were not intended to be scored polytomously. They found that these polytomous items contained slightly less information on average than if they were scored as dichotomous items.

2.6 Decision Accuracy and Decision Consistency

For criterion-referenced tests, the most important interpretations of students' test performance are based on proficiency classifications. Therefore, it is very important to obtain a reliable proficiency estimate and consider the rate of accurate classifications resulting from the use of a test and associated cutscores (Luecht, 2006). The concept of decision consistency (DC) was introduced by Hambleton and Novick (1973) and is defined as the consistency of examinee decisions resulting from either two administrations of the same examination or from parallel forms of an examination. This concept is akin to the index of reliability that reflects the consistency of classifications across repeated testing. Swaminathan, Hambleton, and Algina (1974) suggested the use of the *Kappa* statistic (Cohen, 1960) that would take into account the chance agreement in decision consistency. When two forms are strictly parallel, *Kappa* has a maximum value of 1.0. Livingston and Lewis (1995) defined decision accuracy (DA) as the "extend to which the actual classifications of the test takers agree with those that would be made on the basis of their true score, if their true scores could somehow be known" (p. 180). It is measured by the proportion of correct classifications, false-positive and false-negative rates. Since these concepts were introduced, they have seen wide application and evidence of DA and DC are typically provided in state-mandated testing and in credentialing test programs. Graphical representation for DC and DA are presented in Table 2.2 and 2.3 below, respectively.

Table 2.1 Level of Decision Agreement between Two Parallel Test Forms (i.e., Decision Consistency).

		Classification Decisions based on Test 1	
		Master (✓)	Non-master (×)
Classification Decisions based on Test 2	Master (✓)	Consistent (✓, ✓)	Inconsistent (×, ×)
	Non-master (×)	Inconsistent (×, ×)	Consistent (✓, ✓)

Table 2.2 Level of Consistent Decision Classifications across True Score and Observed Score (i.e., Decision Accuracy).

		Observed Score	
		Master (✓)	Non-master (×)
True Score	Master (✓)	True Positive (✓, ✓)	False Negative (✓, ×)
	Non-master (×)	False Positive (×,✓)	True Negative (✓, ✓)

IRT test information can be very helpful in this sense as test developers have full control in lowering the standard error of a test at any desired set of score levels (with the constraints imposed by the need for content validity); thus, more measurement precision could be achieved. Since a convenient analytical method in predicting the changes in DA as a function of TIF is not readily available, Luecht (2006) examined various test characteristics that could affect test information which would in turn contribute to the classification accuracy in a 3-PLM IRT model: (a) test length; (b) mean of *a*-parameters of the test; (c) standard deviation of the *a*-parameter of the test; (d) mean of *b*-parameters of the test; (e) standard deviation of the *b*-parameter of the test; and (f) location of cutscores. In general, his results showed that increasing test length will always increase DA regardless of the difficulty of items and location of the cuts. When the items are easy and the cutscore is located in the lower ability continuum, both the mean and the standard

deviation of the *a*-parameter will have a substantial effect on DA; on the other hand, *a*-parameter will have a smaller effect on DA when the test is more difficult and the cutscore is moving up on the ability scale. The rate of decrease in DA would be more severe when a test is more difficult than is needed compared to a test easier than is needed. In addition, as the cutscore is moving up on the ability scale, on average the false-negative error rate would increase but the false-positive error rate would remain relatively constant.

2.7 Summary

Although information functions for the dichotomously response IRT models are well-studied, based on the literature review, it is clear that there is a need to better understand the information functions for the polytomous and especially for the mixed IRT models as tests comprised of different item formats are becoming increasingly popular. Based on the literature review, item and test information will generally increase when the discrimination power increases and also when the difficulty of the test is suitable to the intended population for tests only consisting of dichotomously scored items. However, not much is found from the literature about item and test information for the polytomous or mixed IRT model except that polytomous items generally provide more information than dichotomous items. In addition, in the GRM case, increase in item discrimination power will also increase the amount of information an item provide. Certain patterns could be observed about where and when TIF would peak for tests with only polytomous items that are calibrated using GRM or PCM.

Although the focus of this study is still on item and test information functions, this study approached the problem in a different manner. First, the emphasis was on the

impact on TIF when test quality was improved by means of increasing the discriminating power of items in mixed format tests. This could be achieved by increasing the discriminating power for all items in the test; only for those that are low in discrimination values; or by choice of item types. Second, it is not unusual that difficulties of tests change over time, and at other times items are replaced for security reasons, these conditions definitely have an impact on TIF.

Limited studies on the practical consequences of changes in test characteristics were found. Since the primary purpose of examinations is to classify examinees into different proficiency categories, this study also focused on examining the changes in TIF based on improving (i.e., increase in discrimination power) or degrading the test quality (i.e., difficulty of test does not align with examinees ability) in decision consistency and decision accuracy. In addition, the expected information, which is an indicator of the match between the information function and the examinee ability distribution, was also examined. Ideas on how to carry out the analysis are described in detail in the next chapter.

CHAPTER 3

METHODOLOGY

3.1 Introduction

In this chapter, the methodology for the study is described. Two state-wide assessments data were analysed in this study. The purpose of the first study was to investigate the effect of improving the discrimination power of a test and the impact of a less optimal test on test information and measurement precision. The second study was focused on changes in test quality on decision consistency (DC), decision accuracy (DA), and expected information based on simulated response data.

3.2 Design for Study One

This section provides descriptions of the item parameters for the two empirical tests used in this study. The characteristics of the examinee samples are also described.

3.2.1 Item Parameters

Empirical item parameters from two large-scale statewide assessments were used in the analysis of this study: middle school Mathematics assessment and high school English language Arts (ELA). For the middle school Mathematics test, there are 39 items in the test, in which 29 items were dichotomously scored, 5 are short answer items which were scored from 0 to 1, and the remaining 5 items are constructed response items which were polytomously scored from 0 to 4. Therefore, the total raw test score for this Mathematics assessment is 54 points. Three-parameter logistic IRT (3-PLM) model, two-parameter logistic IRT (2-PLM) model, and the graded response model (GRM) (Samejima, 1969) were applied to the multiple choice items, short answer items and constructed response items, simultaneously. Summary of the item parameter estimates for the middle school Mathematics test is presented in the following table.

Table 3.1 Summary of Item Parameter Estimates for Middle School Mathematics by Item Type.

Item Type[1]	*n*	Parameter	Mean	SD	Min	Max
MC	29	*a*	1.07	.30	.60	1.82
		b	-.30	.59	-2.00	.71
		c	.18	.08	.05	.36
SA	5	*a*	.78	.20	.55	1.08
		b	.16	1.06	-1.36	1.14
CR	5	*a*	1.04	.09	.92	1.13
		b	-.60	.54	-1.20	.25

[1] MC – Multiple choice items, SA – Short answer items, CR – Constructed response items

For the high school ELA test, there are 42 items in the test, in which 36 items were dichotomously scored, 4 are constructed response items which were scored from 0 to 4, and the remaining 2 items are essay writing items which were polytomously scored from 0 to 6 and 0 to 10. Therefore, the total raw test score for this ELA assessment is 68 points. The 3-PLM model, and the GRM (Samejima, 1969) were applied to the multiple choice items, constructed response items and essay items, simultaneously. Summary of the item parameters are presented in Table 3.2 below.

Table 3.2 Summary of Item Parameter Estimates for the High School English Language Arts (ELA) by Item Type.

Item Type[1]	*n*	Parameter	Mean	SD	Min	Max
MC	36	*a*	1.12	.29	.59	1.96
		b	-.04	.49	-.84	.97
		c	.22	.06	.11	.38
CR	4	*a*	1.18	.11	1.09	1.34
		b	.42	.17	.27	.57
EI	2	*a*	1.67	.14	1.57	1.77
		b	-.02	.51	-.38	.34

[1] MC – Multiple choice items, CR – Constructed response items, EI – Essay Items

There are several reasons for choosing the above tests to include in this study: first, high school tests are the most consequential for students, and so the more that is known about these tests and how they might be improved, the better the results over time

will be. Moreover, since the middle school Mathematics test does not have essay items, test information is almost certainly lower than the high school ELA test; gains are of the most interest and the most consequential to the results when information is moderate to begin with. In addition, in searching for the testing format of statewide assessments, it was found that over 50% of the states are using mixed format tests. Therefore, results obtained from this study should be useful for practitioners and also generalizable to other tests that have a similar testing format.

3.2.2 Examinee Sample

Since the primary focus of this study was on the IRT item statistics and item information functions, proficiency scores were chosen to follow the standard normal distribution (i.e., $N\sim(0,1)$) to approximate the actual proficiency score distribution, which is commonly found in large-scale assessments.

In this study, a sample of 1,000 proficiency scores was drawn at random for use in data simulation and for the calculation of the classical item statistics. A sample of 1,000 would be large enough to obtain a stable estimate of item-test score correlations. Test developers are usually more comfortable in using classical item statistics (for example, *p*- and *r*-values) when building test forms, examining the increase in point-biserial correlations will increase the understandability needed to bring about the increase in the *a*-parameter.

3.3 Procedures and Evaluation Criteria for Study One

Two different criteria were used to examine the effect of changing the test quality on test information function (TIF), and the procedures are laid out in the following sections.

3.3.1 Changes in Item Discrimination Value (*a*-parameter)

Increase in item or test information could come in many different ways, for example: (1) if a test were lengthened, the addition of more test items would increase the information; (2) if the quality of the test items could be improved, perhaps through more attention to the preparation of item writers, or through the production of more items with the expectation that the best items, statistically, in a bigger pool of items would raise the quality of test items selected, and (3) if the composition of the test items that were selected could be changed, such as by replacing easier items with middle difficult or harder item, information could be increased at the desired ability region (Hambleton & Lam, 2009).

Results from the literature review suggested that increase in item discrimination will generally lead to an increase in item and test information. Three levels of increase in the *a*-parameter (i.e., discriminating power) estimates were considered in this study: .05, .10, and .30. The first and second increase represent fairly minor increases that could be possible, albeit with modest effort, or with an increase in test length. The third condition (i.e., increasing *a* by .30) would require a more substantial effort. These increases are similar in a situation where better items (i.e., higher discriminating power with similar difficulty levels) are available in the item pool and they are chosen to include in the test.

The increases were easy to simulate - simply by increasing the *a*-parameter estimates by the desired amounts. Then, the TIFs and conditional standard error curves were recalculated (for ability scale between -3.0 to 3.0, increment of .01). With each of the increases in the *a*-parameter estimates, this study considered (1) the increase in the TIF (the original and the revised TIFs are displayed), (2) the increase in the item-test score correlations needed to bring about the increase in the *a*-parameter estimates, (3) the

impact on the conditional standard errors (again, the original and the revised SE curves are displayed), and (4) the effective test length increases (i.e., relative efficiency). The effect of increasing the *a*-parameter estimates was analysed by item type and also at the overall test level. In addition, the original *a*-parameter estimates of the test were categorized into three levels: low discrimination group ($a_i < .80$), medium discrimination group ($.80 \leq a_i < 1.0$), and high discrimination group ($a_i \geq 1.0$), effects of increasing the *a* on TIF, measurement precision and relative efficiency were also examined at the overall test level if: (1) only *a* were increased for the low discrimination group; (2) *a* were increased for the low and medium discrimination group.

3.3.2 Changes in Item Difficulty Value (*b*-parameter)

Although the two test forms used in this study were built to target the population of interest; it is not unusual that the characteristics of the test become "off-target". This might happen if the test was breached or when teachers did not provide enough instruction to the course. Therefore it was worthwhile to simulate this less optimal situation where items in the test were not targeted to the ability distribution and study the impact. Less optimal items were generated for this purpose with the difficulties of all items in the original test shifted by -1.0, 1.0, and then by 2.0, representing a substantially more difficult test for the population of interest. TIF, measurement precision and relative efficiency were evaluated based on item type and also at the overall test level for these simulated conditions and compared with the original test.

Item information functions (IIF) based on original parameter estimates and different variations of the *a*- and *b*-parameters are also included in the Appendix. The most optimal test for each subject was built based on examinations of these IIFs. TIF,

measurement precision and relative efficiency of the most optimal test were evaluated at the overall test level.

3.4 Design for Study Two

This study was an extension of Study One in order to examine the effects of changing test quality on decision consistency (DC), decision accuracy (DA), and expected information based on simulated response data. Increasing the TIF, generally, is a good thing to do statistically because score precision is increased; however, this is not always possible in practice. Narrowing the confidence bands for scores provide students with more accurate assessments of their true levels of proficiency. For students close to the cutscores, the extra precision may influence their performance classifications.

3.4.1 Item Parameters

The original item parameter estimates and all variations of *a*- and *b*-parameters in Study One were considered in this study.

3.4.2 Examinee Sample

Ten thousands (10,000) true ability, θ, were simulated from a normal distribution with mean of 0 and standard deviation of 1 for the middle school Mathematics test and also for the high school ELA test. This sample size was large enough to minimize sampling errors in the statistics of interest and avoid confounding sampling errors with the interpretations of real differences. And, a normal distribution of scores is not uncommon in practice. Given the true ability, and item parameters of both binary-scored and polytomously-scored items for each test, response data for each examinee were simulated using the computer program WinGen3 (Han, 2006). All items and all examinees were taken into account in calibration, which was conducted in PARSCALE (Muraki & Bock, 1997). Expected *a prior* (EAP) was used for examinee ability estimates

in all the calibration procedures because of its capacity to produce estimates for candidates who scored the highest on all items or the lowest on all items.

The above-described procedures were repeated for all variations of the *a*- and *b*-parameters as described in Study One. For example, if the studied condition is to increase the *a*-parameter by .05, after adding .05 to the *a*-parameter for all items in the test, response data will be simulated based on the true ability and the updated item parameters. Examinees will then be scored based on the EAP method.

3.4.3 Cutscores and Proficiency Categories

In this study, four performance categories and three cutscores were applied. According to the standard setting of the assessments, the cutscores on the ability metric were determined and they are reported in Table 3.3.

Table 3.3 Cutscores (in Ability Scale) for the Middle School Mathematics Test and High School English Language Arts (ELA) Test.

Test	Cutscores		
	Cat 1/Cat 2	Cat 2/Cat 3	Cat 3/Cat 4
Middle School Mathematics	-.510	.232	1.112
High School ELA	-.414	.384	1.430

When true θ scores were simulated from a standard normal distribution, the corresponding percentages of examinees in each proficiency category for the two tests are as follows.

Table 3.4 Percentages of Examinees in Each Proficiency Category for the Middle School Mathematics Test and High School English Language Arts (ELA)

Test	% of Examinees			
	Cat 1	Cat 2	Cat 3	Cat 4
Middle School Mathematics	31%	28%	28%	13%
High School ELA	34%	31%	27%	8%

3.5 Procedures and Evaluation Criteria for Study Two

Three criteria were used to determine the consequences of improving or degrading the item and test quality: decision consistency (DC), decision accuracy (DA), and the expected information.

3.5.1 Decision Consistency and Decision Accuracy

Decision consistency (DC) is the reliability issue concerning the consistency of decisions made over repeated parallel administrations. In other words, it refers to the consistency of decisions resulting from two parallel test forms or two administrations of the same test. When the results obtained from both tests agreed, then the decisions for the examinee is considered as consistent. Both DC and *Kappa* statistic were reported for evaluation.

Decision accuracy (DA) is the proportion of decisions resulting from the test design that are in agreement with the true classifications of the examinees. The simulated ability level was treated as the truth, based on modifications of the test characteristics (i.e., either by increasing the *a*-parameter or varying the difficulty level of the test by changing the *b*-parameter), new response data were generated and new ability estimates were obtained. This study considered the chances of students with true scores one standard error of measurement below a cutscore actually being classified in the higher category (i.e., false positive error) and the chances of students with true scores one standard error of measurement above a cutscore actually being classified in the lower category (i.e., false negative error) in the original test, and in all the improved and degraded tests, and also the most optimal tests. The comparisons of interest were the errors across all tests at a given true score and the comparisons highlighted the relative advantages of the improved tests and tests that matched with the ability of the target

population. Since DA is an indicator of whether a decision made reflects the truth, it can be seen in its essence as a measure of validity.

3.5.2 Expected Information

Donoghue (1994) suggested the concept of expected information as an indicator of the match between the information function and the examinee ability distribution. The expected information is defined as:

$$E_i(I) = \sum_{q=1}^{Q} w_q \bullet I_i(\theta_q) \tag{14}$$

where w_q is the weight of the posterior ability distribution associated with the quadrature point q, and $I_i(\theta_q)$ is the information for item i at the quadrature point q. The expected information indices by item type for the original test, increasing average a for the overall test at three different levels and manipulations of the overall test difficulty were reported. In addition, relative information based on various item formats for the above studied conditions was also presented.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

Simulation studies were carried out according to the research designs described in Chapter 3, and the results of these studies are presented in this chapter. Results are organized by subject area – middle school Mathematics test then high school ELA test. For the first study, within each subject level, the impact of changing the item discrimination value (i.e., the *a*-parameter) by various item format and at the overall test level on test information function (TIF), conditional standard error of measurement (CSEM), and relative efficiency (RE) are reported. In addition, the effect of increasing the *a*-parameter for low discriminating items (i.e., when $a_i < .80$) and for the low and medium discrimination items (i.e., $.80 \leq a_i < 1.0$) on TIF, CSEM and RE are presented. Item point-biserial (*r*-value), mean and standard deviation of the overall test based on simulated response data are also reported. Next, effects of shifting the level of item difficulty on TIF, CSEM and RE are reported by various item formats and also at the overall test level.

Results for the second study are also organized by subject area. Within each subject, three consequences of improving or degrading the item and test quality by means of increasing the item discrimination power or manipulating the level of test difficulty are reported – decision consistency (DC), decision accuracy (DA) and expected information based on all conditions in the first study.

4.2 Study One – Changes in Item Discrimination Value

4.2.1 Middle School Mathematics Test

Based on the summary of the item parameter estimates for the middle school Mathematics test as presented in Table 3.1, the average discrimination power for short answer items (SA) was lower than the multiple choice (MC) items and constructed response (CR) items and MC items had the highest discrimination power over the other two item formats. The spread of the *a*-parameters for the MC and SA items were comparable, but the *a*-parameters for the CR items had a larger spread.

4.2.2 Effects of Increasing Discriminating Power on the Multiple Choice Items

Figures 4.1 and 4.2 display the information functions and the conditional standard errors for the multiple choice (MC) items. Each figure contains the information or conditional standard errors based on the original item parameter estimates and the three levels of increase in discriminating power: $a + .05$; $a + .10$; and $a + .30$.

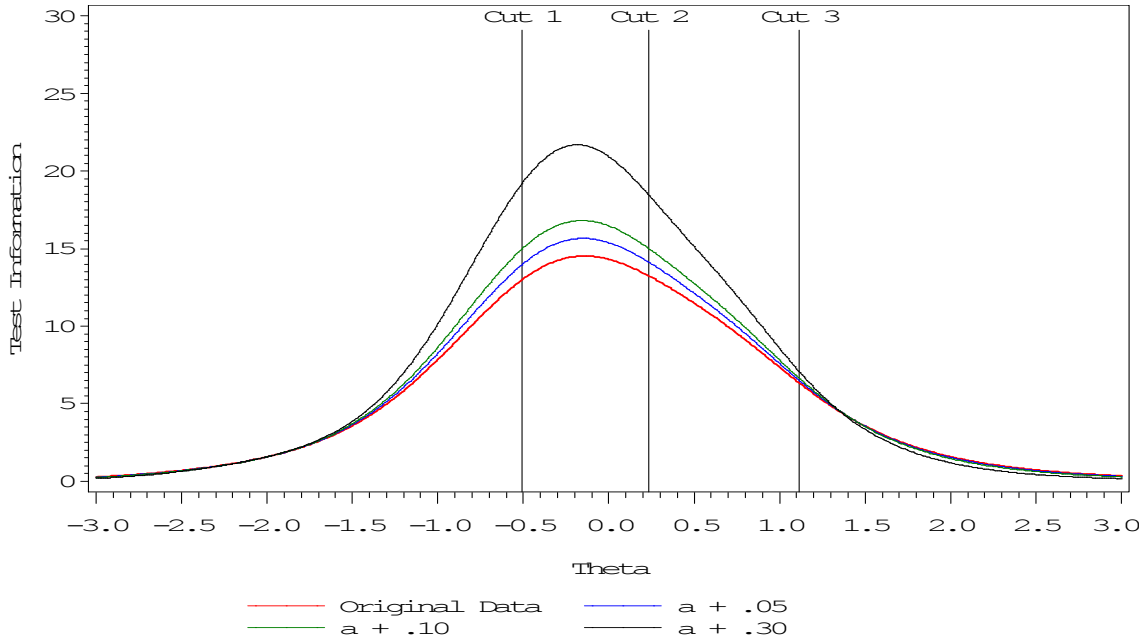


Figure 4.1 Middle school Mathematics test – increasing discriminating power: Test information based on multiple choice items only (29 items, maximum score = 29)

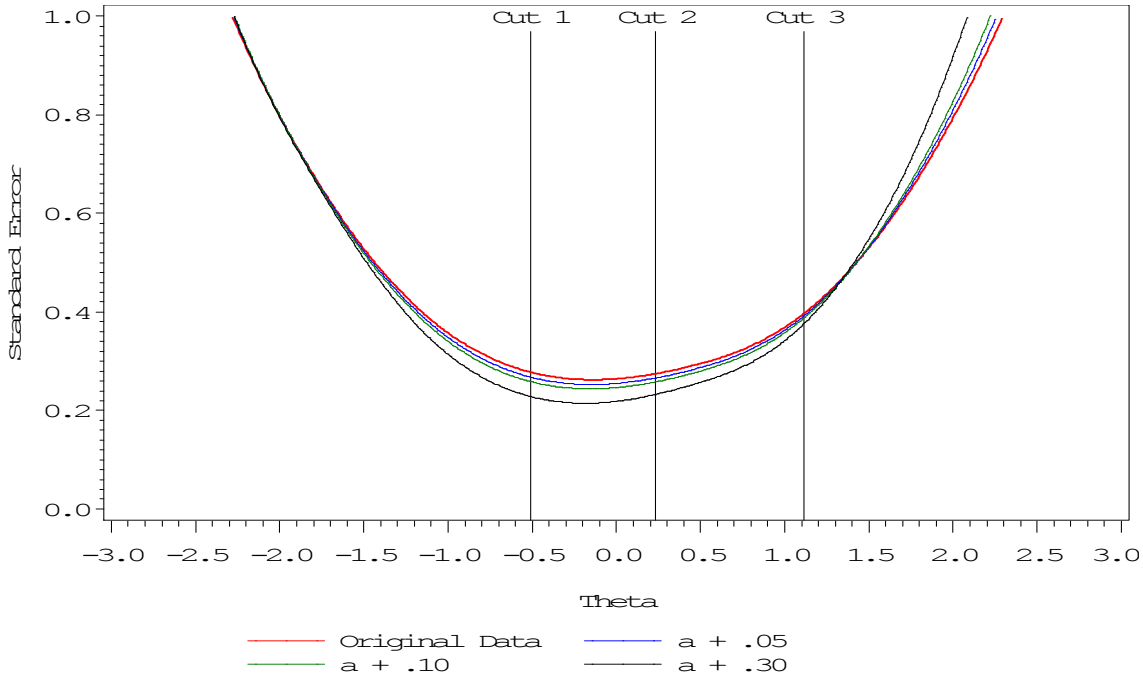


Figure 4.2 Middle school Mathematics test – increasing discriminating power: Conditional standard error of measurement based on multiple choice items only

The level of information at the three cutscores based on the original item parameter estimates of MC items only was approximately 13.0, 13.0 and 6.5, and the corresponding standard errors of measurement were .28, .28, and .39 at Cut 1, Cut 2 and Cut 3, respectively. Increasing the discriminating power of the MC items by .05 increased the information at the three cutscores to about 14.0, 14.0, and 7.0. When the discriminating power of the MC items was increased by .10, the amount of information at the three cutscores became 15.0, 15.0, and 7.0. Finally, increasing the item discrimination by .30 for MC items yielded a substantial increase in information for the first two cutscores: approximately 19.0 at Cut 1, 18.5 at Cut 2, and only increased the information to 8.0 at the third cutscore. Maximum information based on MC items occurred at proficiency score at about -.20, and increasing the discriminating power of test items did not affect where maximum information occurred.

Results from the conditional standard errors of measurement confirmed that increasing the discriminating power generally decreased the standard error of measurement, thus, providing more measurement precision. However, as shown in Figure 4.2, increasing the discriminating power of items did not guarantee lower measurement error. As in the case of the conditional standard error curves for the MC items, when ability parameters were above 1.3, measurement precision was actually the lowest when items had the highest discrimination value (i.e., when *a* was increased by .30).

Figure 4.3 displays the relative efficiency of each of the three improved tests versus the original test.

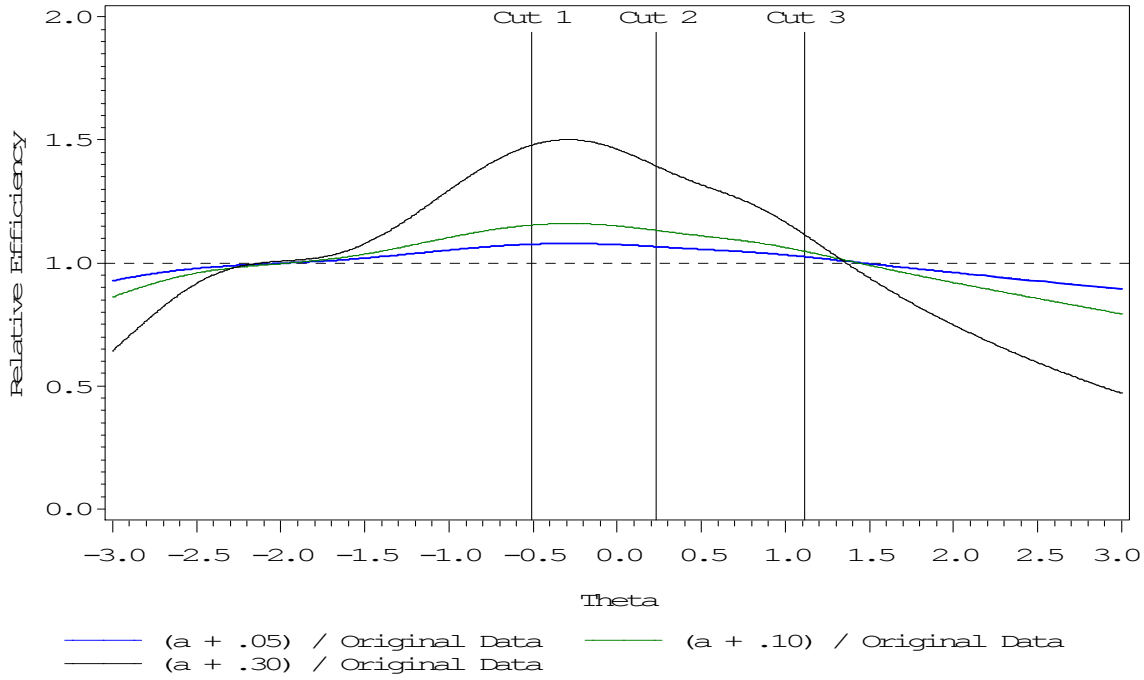


Figure 4.3 Middle school Mathematics test – increasing discriminating power: Relative efficiency based on multiple choice items only

When only considering MC items in the middle school Mathematics test, an increase of .05 in the *a*-parameter estimates increased the effective length of the MC-item test relative to the original MC-item test at the cutscores by about 6%. In other words, if the overall discrimination power of the MC items could be increased by .05, and holding

the difficulties (i.e., the *b*-parameters) and pseudo-guessing parameters (i.e., *c*-parameters) constant, the MC section of the test could be shortened by 6% (from 29 items to 28 items) but this shortened MC-test could still achieve the same level of measurement precision as the longer version of the MC-test. In addition, the effect of increasing the *a*s was about the same at the first two cutscores, but the effect on the third cutscore was not as much.

An increase of .10 in *a* for the MC items increased the effective length relative to the original test by about 12%, meaning that on average the MC section of the test could be shortened by 3 items with the same measurement precision when compared to the original MC test. The same pattern of the effectiveness of increasing the *a*s was observed at the three cutscores as in the previous case: the increase was more noticeable at the first two cutscores than the third cutscore.

An increase of .30 in the average *a*-parameter for the MC items increased the effective length of the test relative to the original MC test by about 14 to 48%, depending on the location of the cutscores. Same as the previous two cases, the increase was more effective in the first two cutscores than the third cutscore. However, when comparing the three improved MC tests with the original MC test, increasing *a*s made the new tests less efficient than the original MC test when proficiency scores were above 1.3 and below -2.3.

4.2.3 Effects of Increasing Item Discriminating Power on the Short Answer Items

Figures 4.4 and 4.5 present the information functions and the conditional standard errors for the short answer (SA) items in the middle school Mathematics test. Each figure contains the information or conditional standard errors based on the original item parameter estimates and the three levels of increase in discriminating power.

The level of information at the three cutscores for SA items based on the original item parameter estimates was approximately 1.1, 1.5 and 1.8 and their corresponding standard errors of measurement were .95, .82, and .75 at Cut 1, Cut 2 and Cut 3, respectively. Low information and high standard errors were expected as the results were only based on 5 dichotomously scored items.

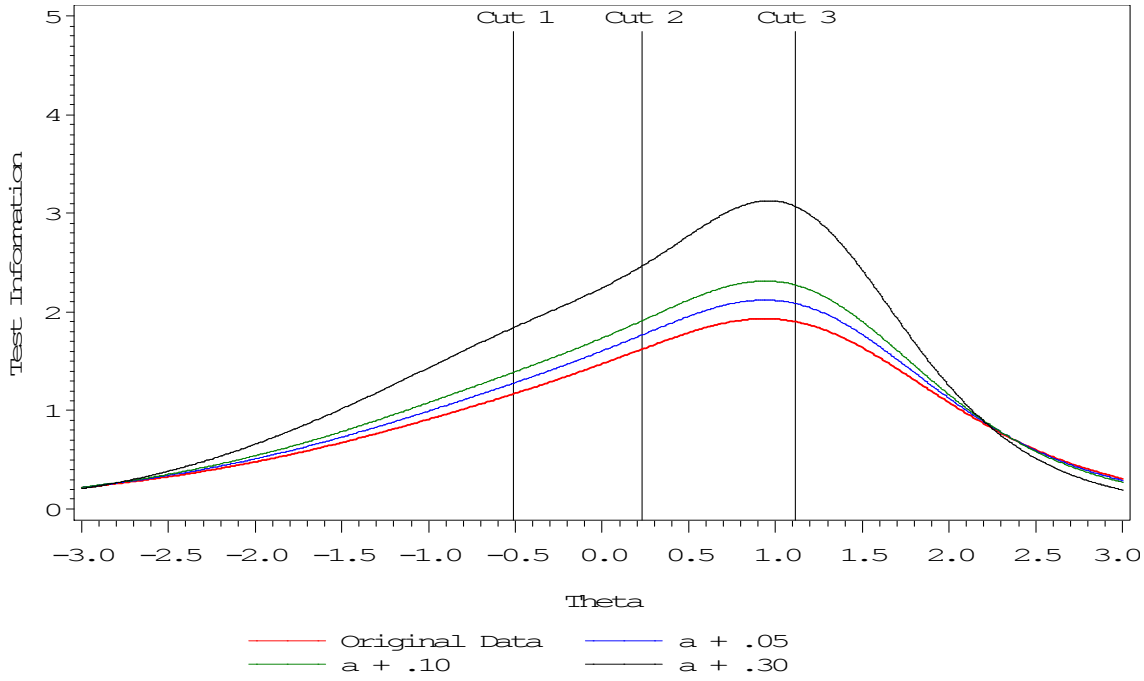


Figure 4.4 Middle school Mathematics test – increasing discriminating power: Test information based on short answer items only (5 items, maximum score = 5)

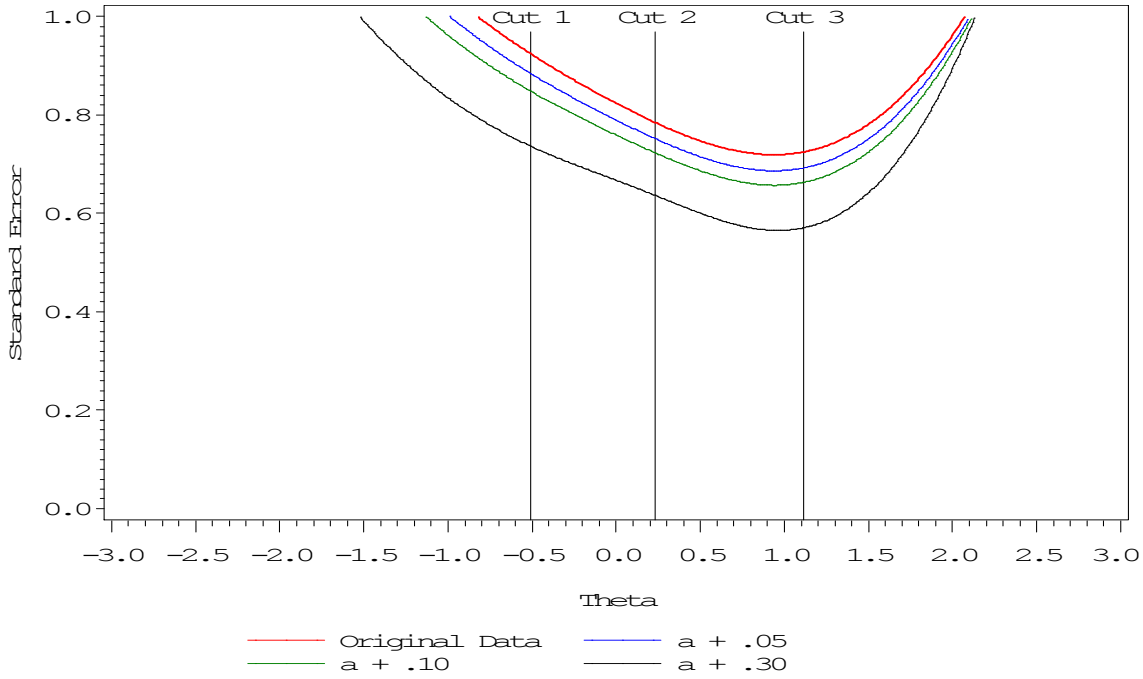


Figure 4.5 Middle school Mathematics test – increasing discriminating power: Conditional standard error of measurement based on short answer items only

Increasing the discriminating power of the SA items by .05 or .10 only had a small effect on the information function. Increasing the item discrimination by .30 for SA items increased the information to approximately 1.8, 2.4, and 3.0 at Cut 1, Cut 2 and Cut 3, respectively. However, when the proficiency scores were above 2.2, increasing the *a*s led to a decrease in information. Standard error of measurement was above .50 for all cutscores even when the average *a* was increased by .30.

Figure 4.6 reports the relative efficiency of the SA items from the original test compared to the three improved SA tests. Increasing the item discrimination power by .05 pushed the new test to have an effective test length roughly 9% more than the original test. An increase of .10 in the average *a*-parameter made the effective length of the new test about 19% longer. Finally, with an increase of .30, the SA test was more effective at Cut 1 and Cut 3: the effective length was about 58% longer than the original SA items at Cut 1 and about 62% longer than the original SA items at Cut 3. At Cut 2, the effective

length was about 52% longer than the original test when the *a* was increased by .30. Notice that increasing the average discriminating power by .30 for the SA items had an adverse effect on measurement precision at the extreme ends of the proficiency continuum.

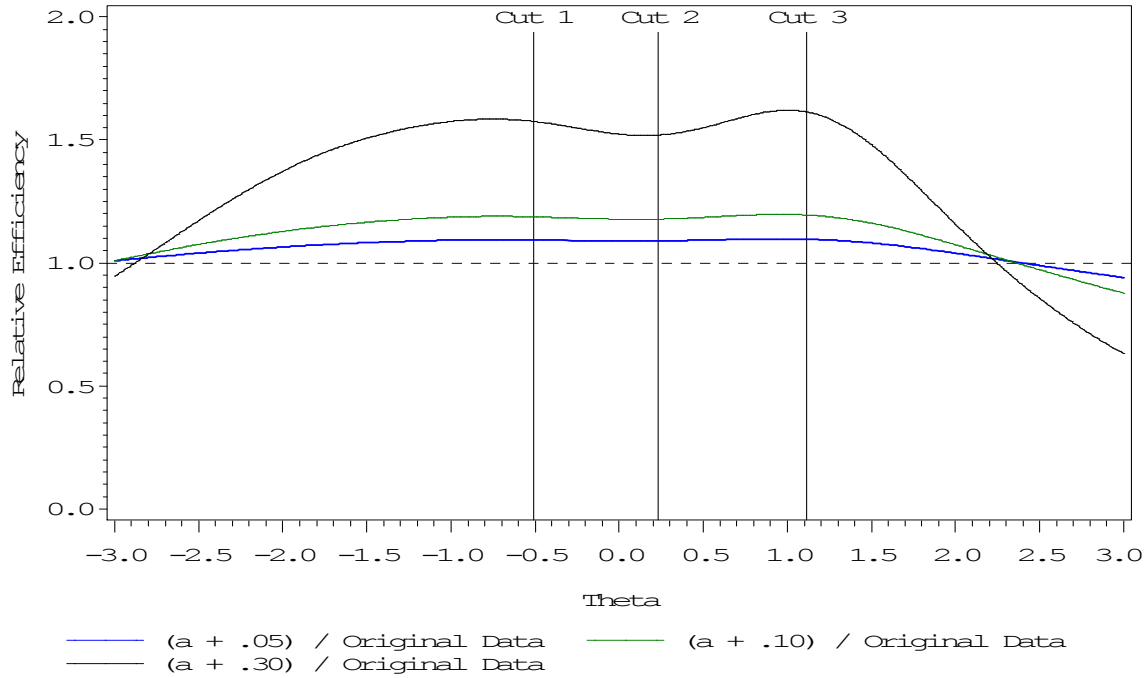


Figure 4.6 Middle school Mathematics test – increasing discriminating power: Relative efficiency based on short answer items only

4.2.3.1 Effects of Increasing Item Discriminating Power on the Constructed Response Items

Figures 4.7 and 4.8 display the information functions and the conditional standard errors of measurement for the constructed response (CR) items from the middle school Mathematics test.

The amount of information provided by the first two cutscores was quite similar within each version of the test, and information at the third cutscore provided the least amount of information in all cases. The level of information at the three cutscores for CR items based on the original item parameter estimates were approximately 4.5, 4.0 and 3.5;

the corresponding standard errors of measurement were .47, .50, and .53 at Cut 1, Cut 2 and Cut 3, respectively. The information function peaked at θ = -1.20 for all cases. Therefore, increasing the item discrimination value increased the amount of information and lowered the measurement error but did not affect the place where the information peaked. In addition, when proficiency scores were above 2.6, increasing the discrimination value of the CR items did not have any effect on the information.

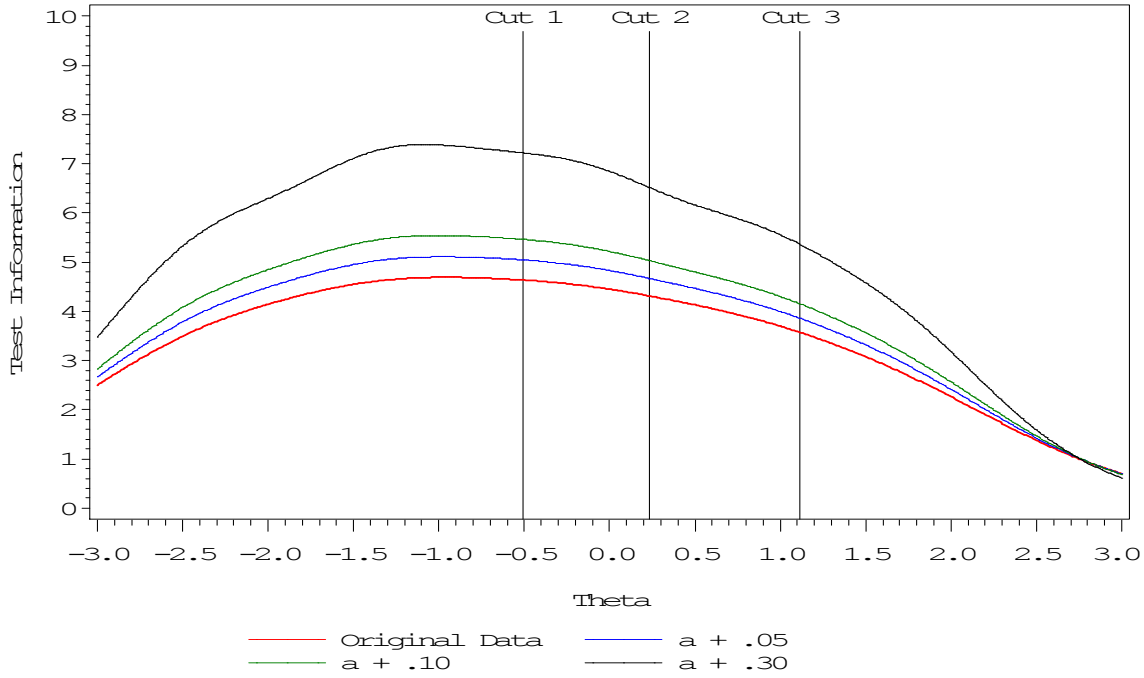


Figure 4.7 Middle school Mathematics test – increasing discriminating power: Test information based on constructed response items only (5 items, maximum score = 20)

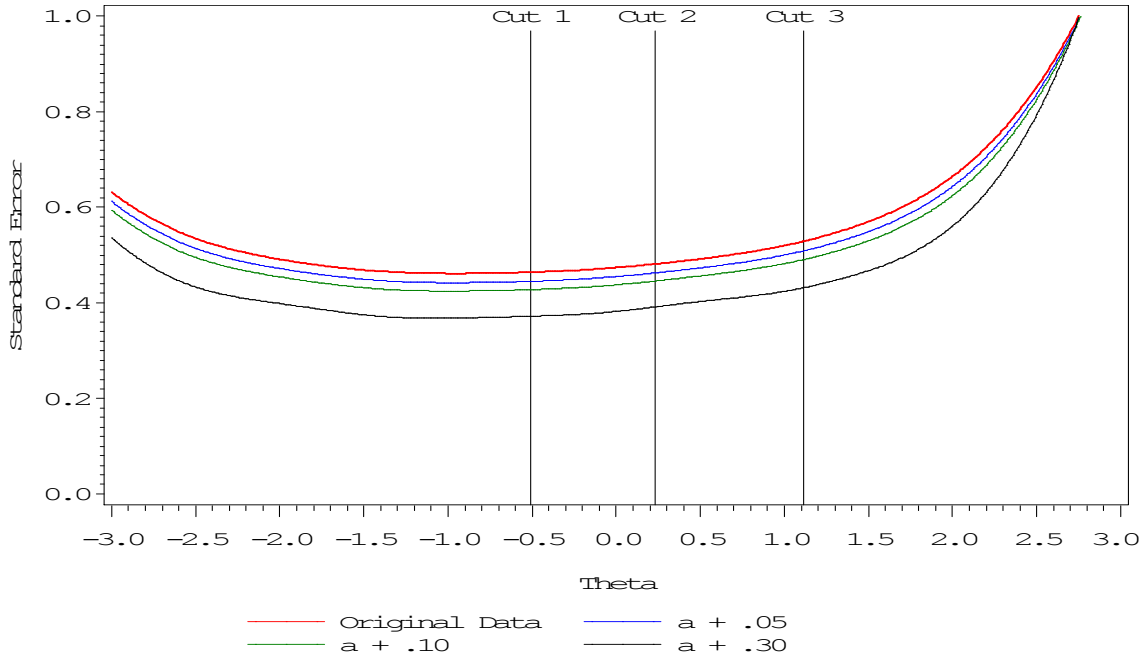


Figure 4.8 Middle school Mathematics test – increasing discriminating power: Conditional standard error of measurement based on constructed response items only

Figure 4.9 reports the relative efficiency of the CR items for the original test compared to the three improved tests.

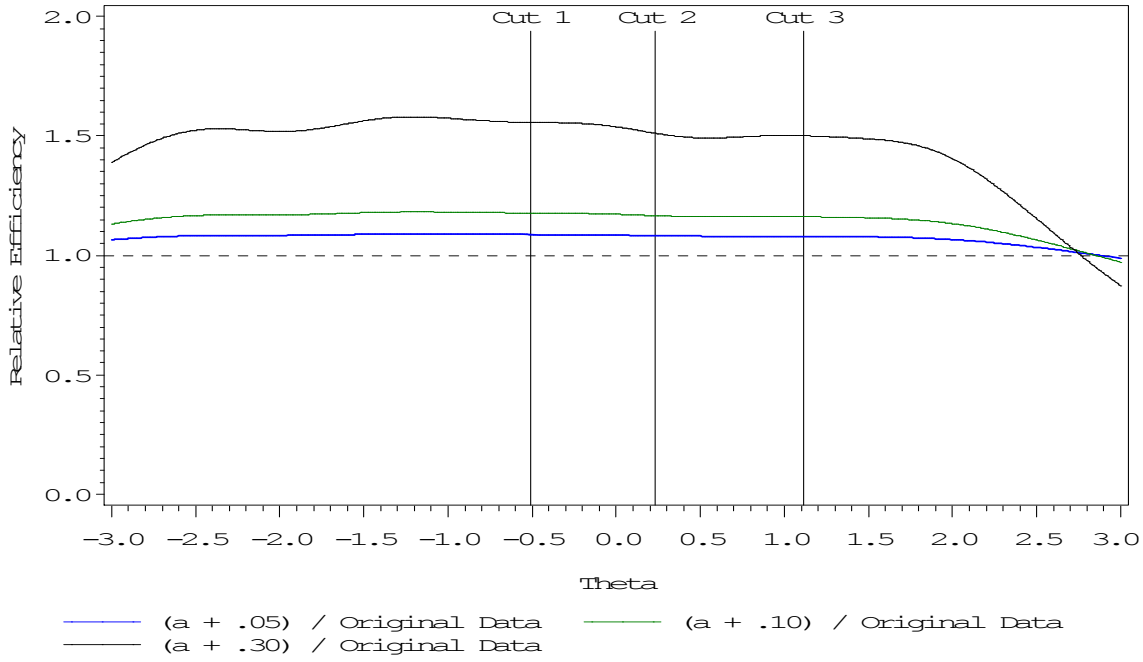


Figure 4.9 Middle school Mathematics test – increasing discriminating power: Relative efficiency based on constructed response items only

An increase of .05 in the *a*-parameter estimates increased the effective length relative to the original CR items at the cutscores by about 8%; an increase of .10 increased the effective length relative to the original items by about 17%; and an increase of .30 in the average *a*-parameter of the CR items made the effective length to be about 53% longer than the original CR items. In the case where the average *a* of the CR items was increased by .30, this new CR test became more effective at the first cutscore (effective length = 1.56) than the second (effective length = 1.52) and third cut (effective length = 1.51) when comparing to the original CR items. In addition, when proficiency scores were above 2.7, increasing the *a*s by .30 actually made the test less effective compared to the other two scenarios where the average of the *a*-parameters was increased by .05 and .10.

4.2.3.2 Effects of Increasing Item Discriminating Power on the Overall Test

Figures 4.10 and 4.11 compare the information functions and the standard errors for the overall test based on original parameter estimates and when the item discriminating parameters were increased by .05, .10 and .30.

Test information for the original Mathematics test was approximately 19.0, 19.5, and 12.0 at Cut 1, Cut 2 and Cut 3, and the corresponding conditional standard errors of measurement for the three cuts were about .23, .23, and .29. Gains from the improved tests can be seen. Using information = 10 as a criterion, which corresponds to a classical reliability estimate of about .90, the improved tests were very helpful in adding to test information but the additions were not essential. The original test information function was not only high enough at the cutscores but also fairly well centered. Higher amount of information at the lower two cutscores was expected as the difficulty level for most of the

items in the test were in that range: the average difficulty of the MC items was -.30 and the average difficulty of the CR items was -.60.

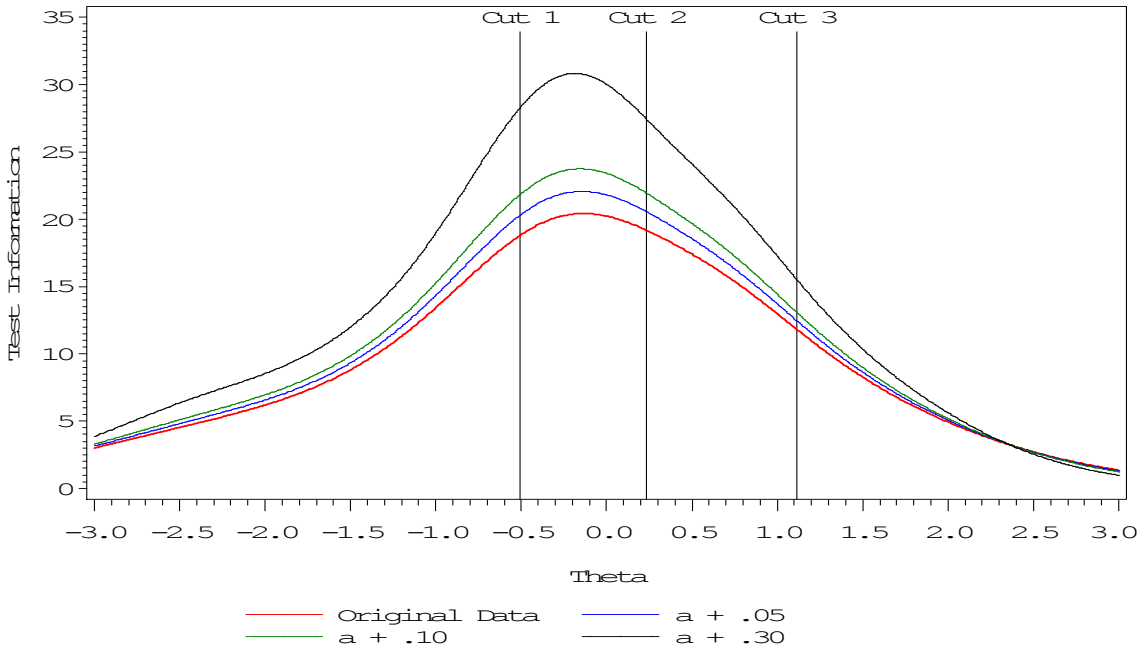


Figure 4.10 Middle school Mathematics test – increasing discriminating power: Test information for the overall test and the improved tests (39 items, maximum score = 54)

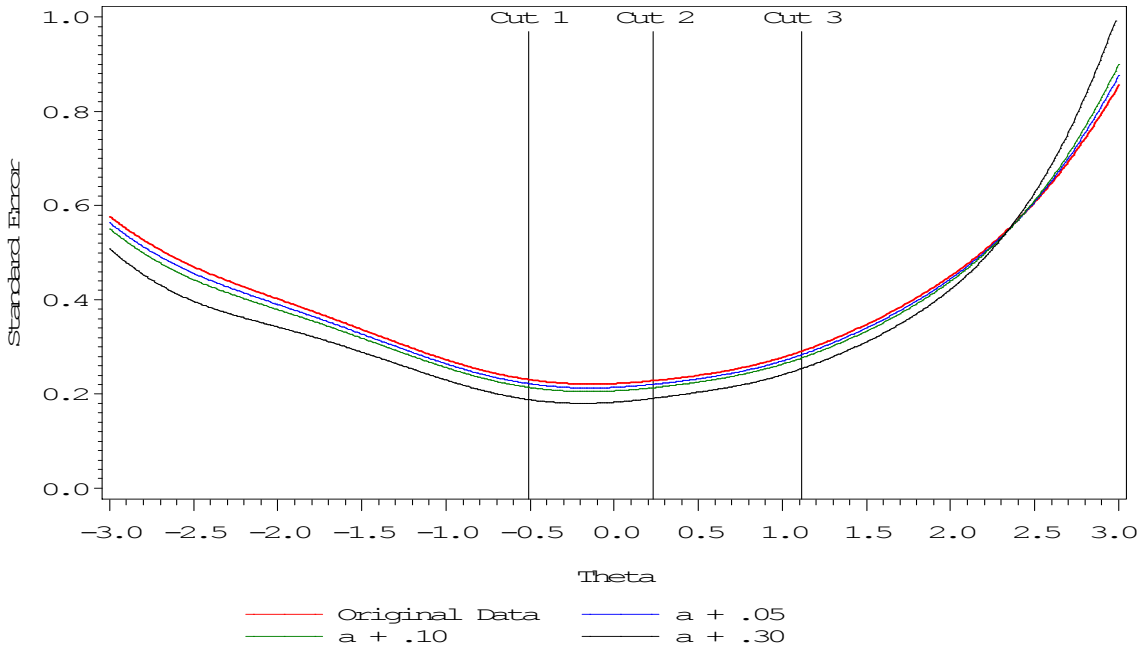


Figure 4.11 Middle school Mathematics test – increasing discriminating power: Conditional standard error of measurement for the overall test and the improved tests

Figure 4.12 below displays the relative efficiency of each of the three improved tests versus the original test.

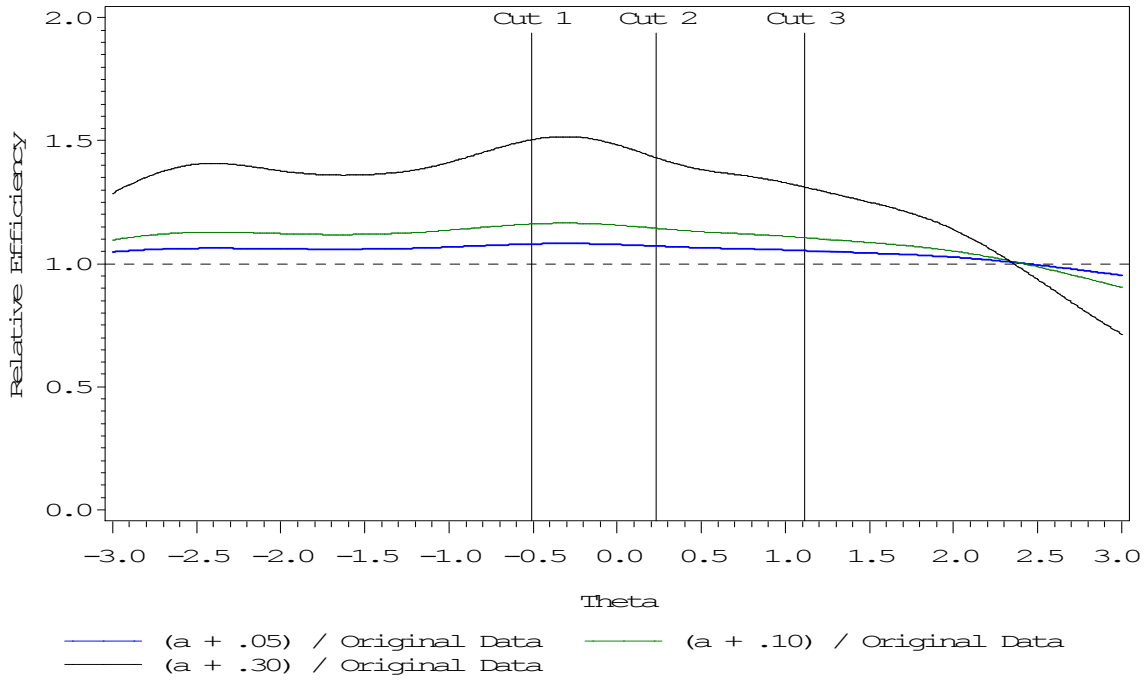


Figure 4.12 Middle school Mathematics test – increasing discriminating power: Relative efficiency for the overall test and the improved tests

An increase of .05 in the *a*-parameter estimates increased the effective length relative to the original test at the cutscores by about 6%; an increase of .10 increased the effective length relative to the original test by about 14%; and an increase of .30 in the average *a*-parameter increased the effective length of the test relative to the original test by about 32 to slightly over 50% except for proficiency scores above 2.3 and here the relative length dropped by 30%.

Table 4.1 provides a summary of the item-test score correlations for MC, SA, CR and all test items based on data simulated from 1,000 proficiency scores which were randomly drawn from standard normal distribution (i.e., *N*~(0,1)). Increasing the IRT discriminating power by .05 would slightly increase the classical item-test score correlation by about .01; increasing the IRT discriminating power of the test by .10 would

lead to an increase in classical item-test score correlation by about .02. Finally, increasing the overall test discrimination power by .30 would lead to an average increase of .06 in item-test score correlations.

Table 4.1 Middle School Mathematics Test: Average Classical Item-Test Score Correlations by Item Type and Total Test (N = 1,000).

Item Type[1]	Original Data	a + .05	a + .10	a + .30
MC	.362	.374	.382	.421
SA	.348	.338	.364	.388
CR	.544	.559	.556	.621
All Items	.383	.393	.402	.443

[1] MC – Multiple choice items, SA – Short answer items, CR – Constructed response items

Table 4.2 highlights that increasing item discrimination power in a test had an impact on score spread but had little impact on the mean test score, and this finding is well known.

Table 4.2 Middle School Mathematics Test: Mean and Standard Deviation of Test Scores (N = 1,000).

Average Score (SD)	Original Data	a + .05	a + .10	a + .30
	33.00 (9.43)	32.77 (9.59)	33.07 (9.72)	33.58 (10.41)

4.2.3.3 Effects of Increasing Item Discriminating Power on the Low Discriminating Items

Sometimes there might not be a lot of good quality items (i.e., items with high discrimination value) for test developers to choose from when building a test, so by replacing some low discriminating items with those that have higher discriminating values might be more probable, provided that the change does not affect the content validity. The focus of this section is to examine the effect of increasing the discriminating power for those items that are low in discrimination in a mixed format test. Distribution of the item discrimination powers is presented in Figure 4.13.

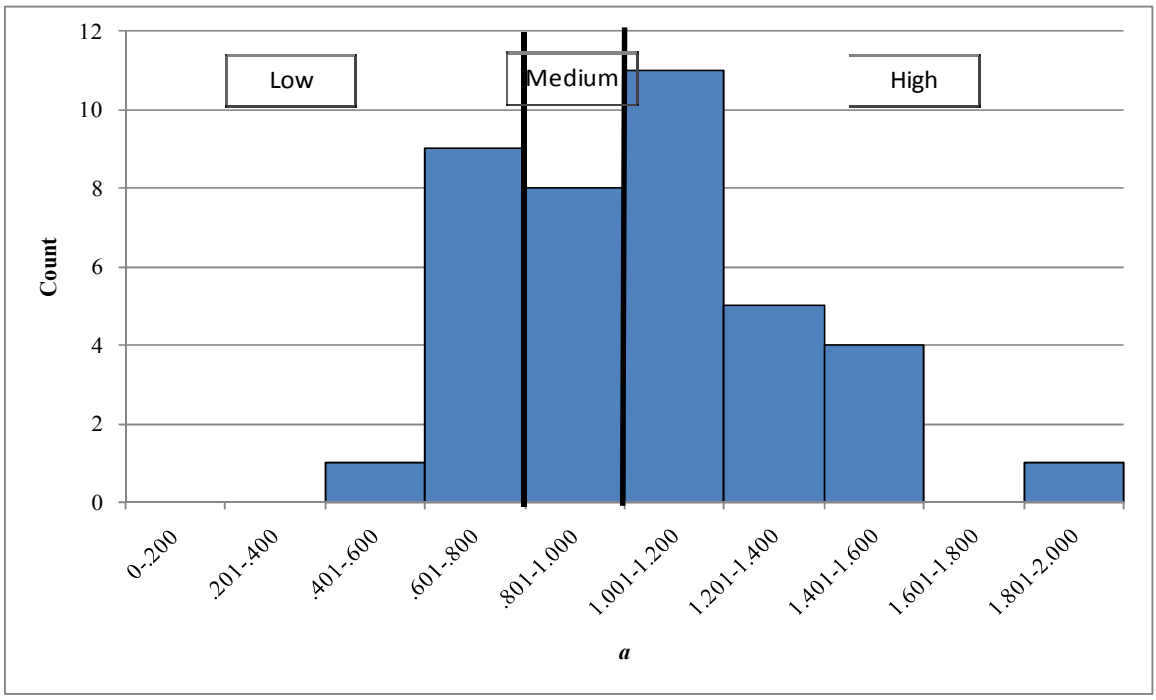


Figure 4.13 Middle school Mathematics test: Distribution of the *a*-parameters

Based on the original item parameter estimates, ten items in this test were categorized in the low discrimination group (when $a_i < .80$), and within these ten items, seven were multiple choice (MC) items and the remaining three were short answer (SA) items.

Figures 4.14 and 4.15 compare the information functions and the standard errors for the overall test based on original parameter estimates and when the item discrimination parameters for items in the low discrimination group were increased by .05, .10 and .30.

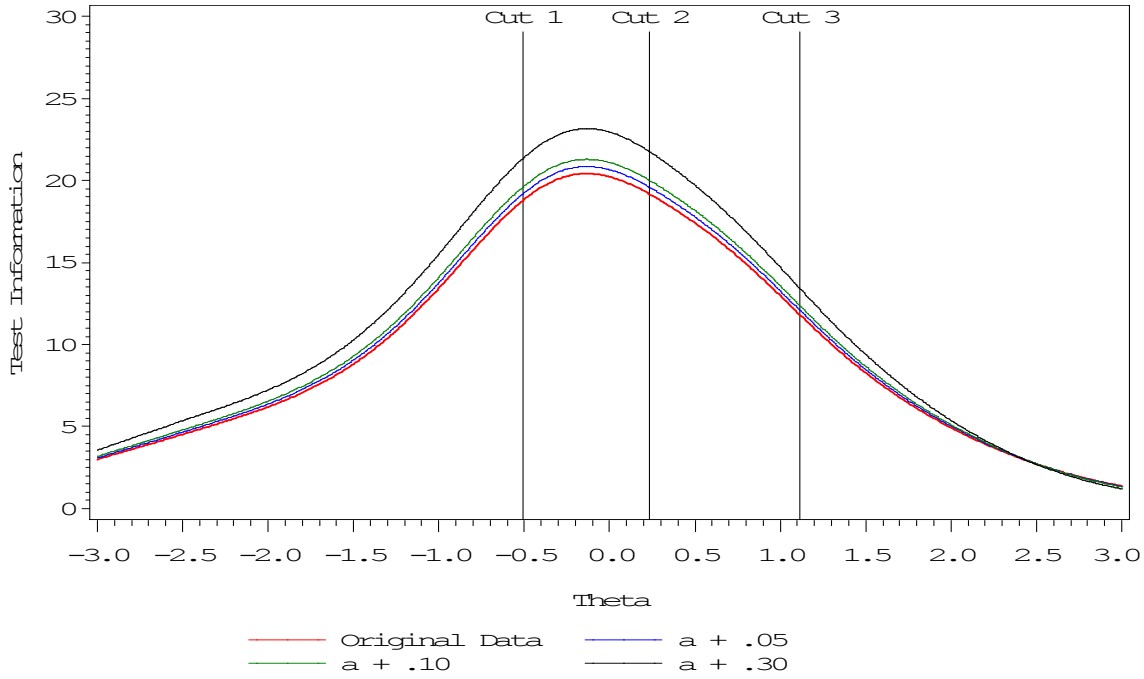


Figure 4.14 Middle school Mathematics test – increasing discriminating power: Test information for the overall test and improved item discrimination for low discrimination group (39 items, maximum score = 54)

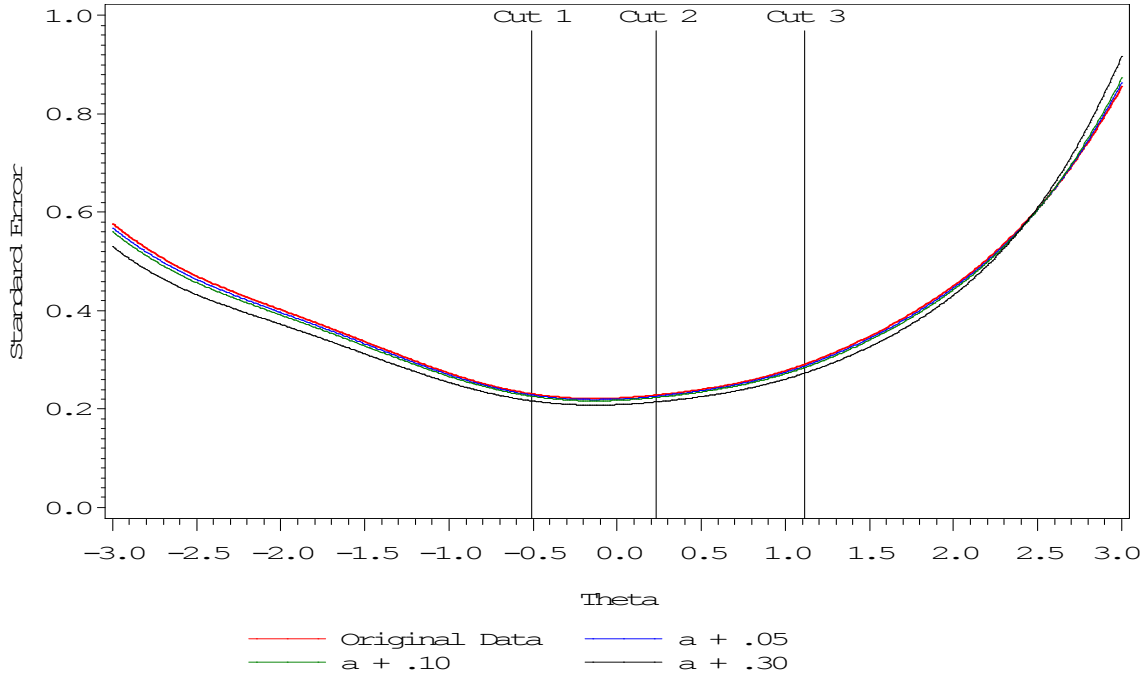


Figure 4.15 Middle school Mathematics test – increasing discriminating power: Conditional standard error of measurement for the overall test and improved item discrimination for low discrimination group

Test information for the original Mathematics test was approximately 19.0, 19.5, and 12.0 at Cut 1, Cut 2 and Cut 3, respectively, and their corresponding conditional standard errors of measurement were about .23, .23, and .29. Slight increases in test information from the improved tests can be seen. For example, increasing the low discriminating items in the test by .10 increased the test information to approximately 19.5, 20.0, and 12.5 for the three cuts and the corresponding conditional standard errors of measurement were about .23, .22, and .28. Figure 4.15 also shows that when proficiency scores were above 2.1, extreme increase in the discriminating power for the low discriminating items (i.e., increasing *a* by .30) would lead to a lower measurement precision test for those at the high end of the proficiency continuum.

Figure 4.16 displays the relative efficiency of each of the three improved tests versus the original test.

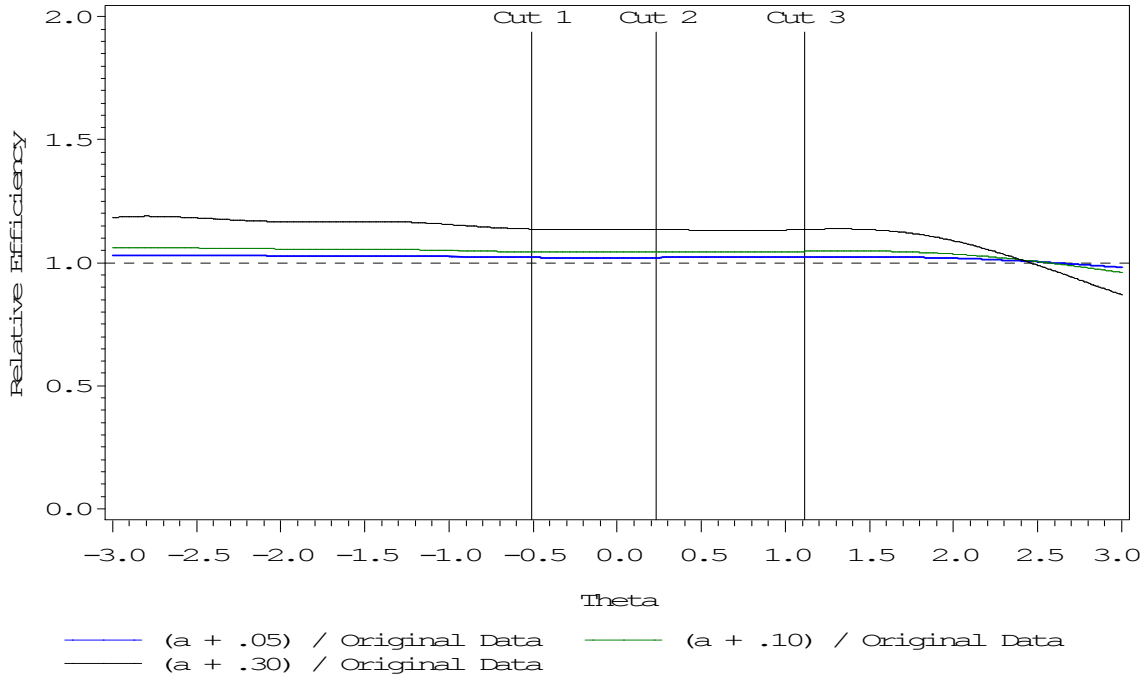


Figure 4.16 Middle school Mathematics test – increasing discriminating power: Relative efficiency for the overall test and improved item discrimination for low discrimination group

Regarding the relative efficiency results, an increase of .05 in the *a*-parameter estimates for the low discriminating items increased the effective length relative to the original test at the cutscores by about 2%; an increase of .10 increased the effective length relative to the original test by about 3%; and an increase of .30 in the averaged *a*-parameter increased the effective length of the test relative to the original test by about 7%. However, the test became less effective than the original test at the high end of the proficiency continuum (i.e., when proficiency scores were above 2.3).

4.2.3.4 Effects of Increasing Item Discriminating Power on the Low and Medium Discriminating Items

Based on the results presented in Figure 4.13, eight items were classified as medium discrimination group (i.e., $.80 \leq a_i < 1.0$), and of the eight items, six of them are multiple choice (MC) items, one is short answer (SA) item and the other one is a constructed response (CR) item. This section presents the results of test information, conditional standard error of measurement and relative efficiency of increasing the discriminating power of the low and medium discriminating items.

As presented in section 4.2.1.4 and 4.2.1.5, test information for the original Mathematics test was approximately 19.0, 19.5, and 12.0 at Cut 1, Cut 2 and Cut 3, and their corresponding conditional standard errors of measurement were about .23, .23, and .29. As expected, increasing the item discrimination power for the low and medium discrimination group made the test more informative and with less measurement error compared to the amount of information provided by only increasing the discriminating power of the low discriminating items as in the previous section. For example, increasing the *a*s by .30 for the low and medium discriminating items increased the test information to approximately 22.5, 22.5 and 14.0 at the three cutscores (see Figure 4.17); comparing

to 20.5, 21.0 and 13.0 when only increased the *a*s by .30 for the low discrimination group. In addition, slightly lower conditional standard errors of measurement at the three cutscores (see Figure 4.18): .21, .21 and .27, when increasing the discriminating power by .30 for the low and medium discrimination group whereas .22, .22 and .28 when only increased the *a*s by .30 for the low discriminating items.

Relative efficiency was higher when increasing discriminating power for both low and medium discriminating items. For example, an increase of .30 in the average *a*-parameter increased the effective length of the test relative to the original test by about 18% (see Figure 4.19). In addition, this test only became less effective than the original test when proficiency scores were above 2.4.

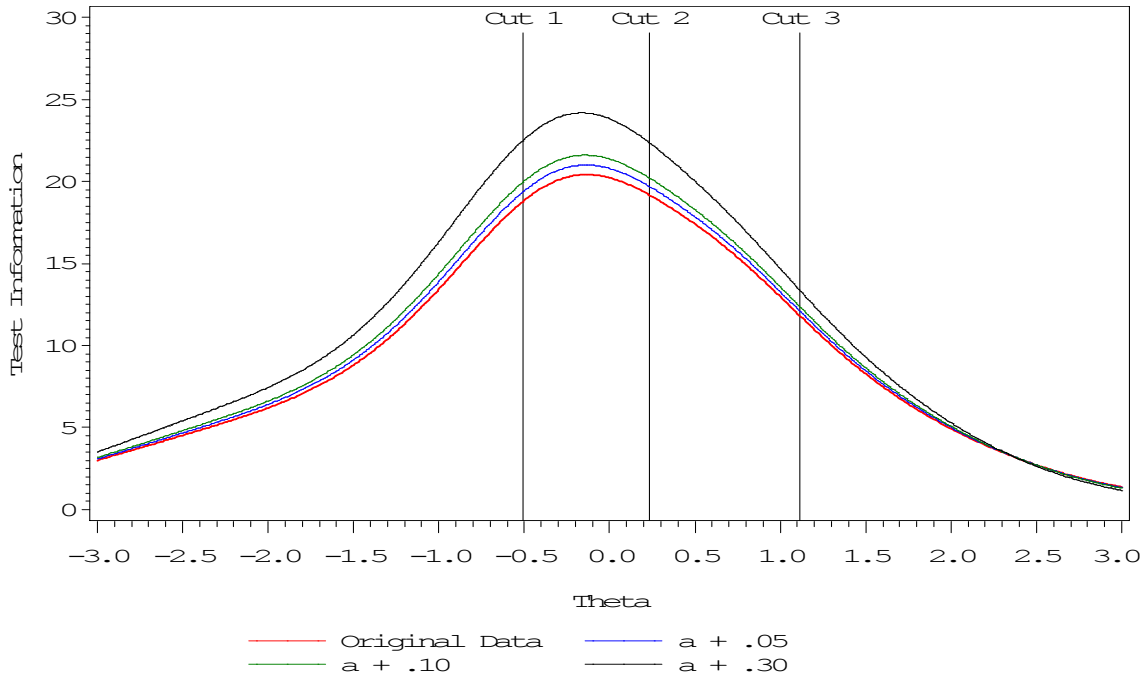


Figure 4.17 Middle school Mathematics test – increasing discriminating power: Test information for the overall test and improved item discrimination for low and medium discrimination group (39 items, maximum score = 54)

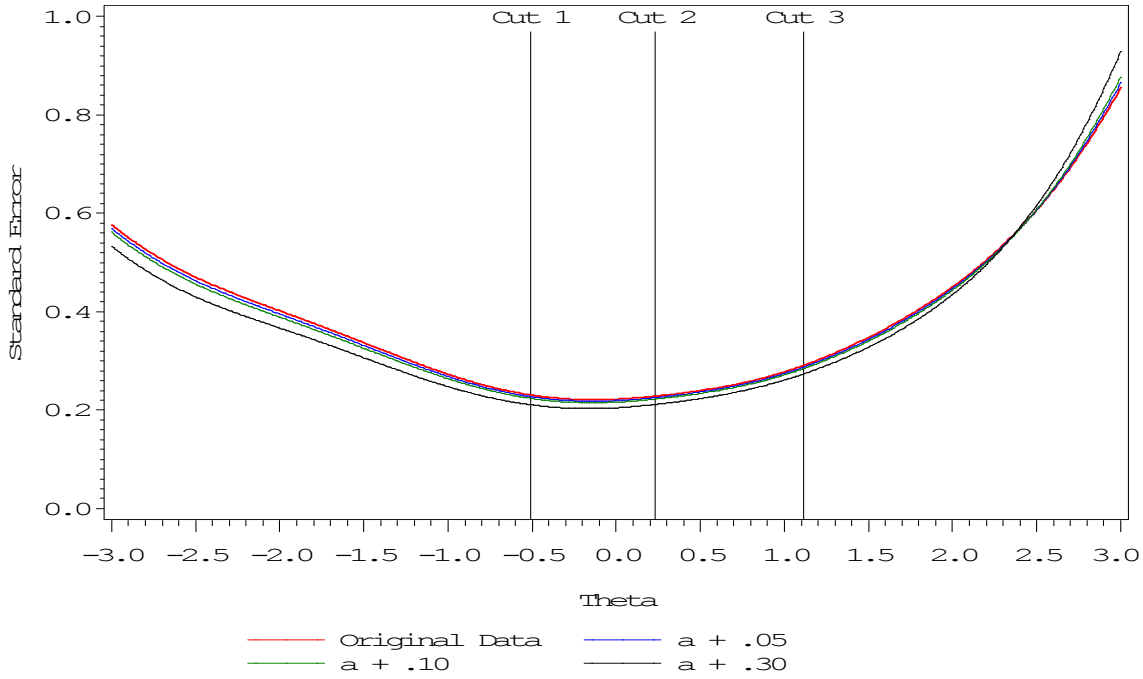


Figure 4.18 Middle school Mathematics test – increasing discriminating power: Conditional standard error of measurement for the overall test and improved item discrimination for low and medium discrimination group

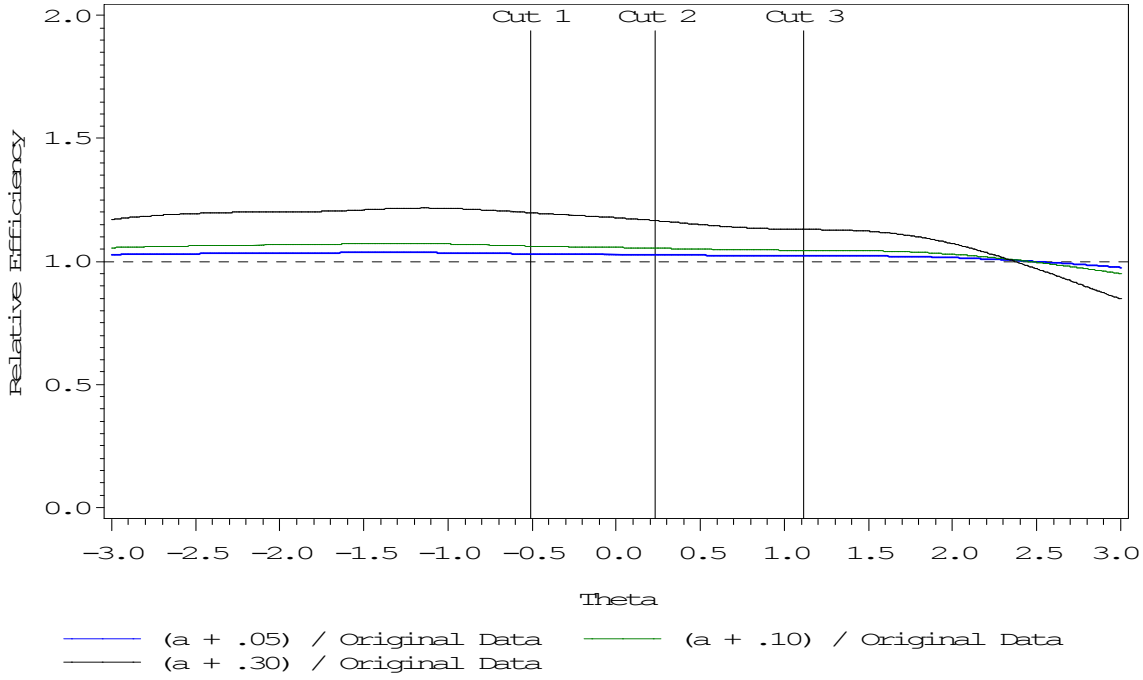


Figure 4.19 Middle school Mathematics test – increasing discriminating power: Relative efficiency for the overall test and improved item discrimination for low and medium discrimination group

4.2.3.5 Summary

In summary, multiple choice (MC) items in the middle school Mathematic test provided modest amount of information at the first two cutscores, and the information function was quite well centered too. Increasing the discriminating power for the MC items helped to increase the information but it was not essential in this test and it did not help pushing the information up at the third cutscore. For short answer (SA) items, information peaked at the third cutscore, increasing discriminating power of the SA items helped to increase the amount of information at all cutscores. On the other hand, information peaked at the lowest cutscore for the constructed response (CR) items, increasing the discriminating power of those items helped to increase the amount of information at the three cutscores. The original test information function for the overall test was rather well centered. Information increased when the discriminating power was increased at the overall test level; however, the addition was not essential. The same conclusion applied to those results obtained from increasing discriminating power for the low discrimination group (i.e., $a_i < .80$) and increasing discriminating power for the low and medium (i.e., $.80 \leq a_i < 1.0$) discrimination group. In all cases, increasing discriminating power decreased the conditional standard error rate and hence, achieving higher measurement precision. In addition, tests were generally more efficient than the original test when discriminating power was increased; however, increasing discriminating power could also lead to less efficient tests especially at the extreme ends of the proficiency continuum. For example, the conditional standard error of measurement was higher for proficiency scores above the third cutscore when average *a* was increased by .30.

4.2.4 High School English Language Arts (ELA) Test

Based on the summary of the item parameter estimates for the high school ELA test as presented in Table 3.2, the average discriminating power for all item types were above 1.0. Essay items (EI) had the highest average discriminating power followed by the constructed response items (CR) then lastly multiple choice items (MC). In addition, MC items had a wider spread of discriminating power than the other two item format.

4.2.4.1 Effects of Increasing Discriminating Power on the Multiple Choice Items

Figures 4.20 and 4.21 display the information functions and the conditional standard errors of the MC items in high school ELA test. Each figure contains the information or conditional standard errors based on the original item parameter estimates and the three levels of increase in discriminating power: *a* + .05; *a* + .10; and *a* + .30.

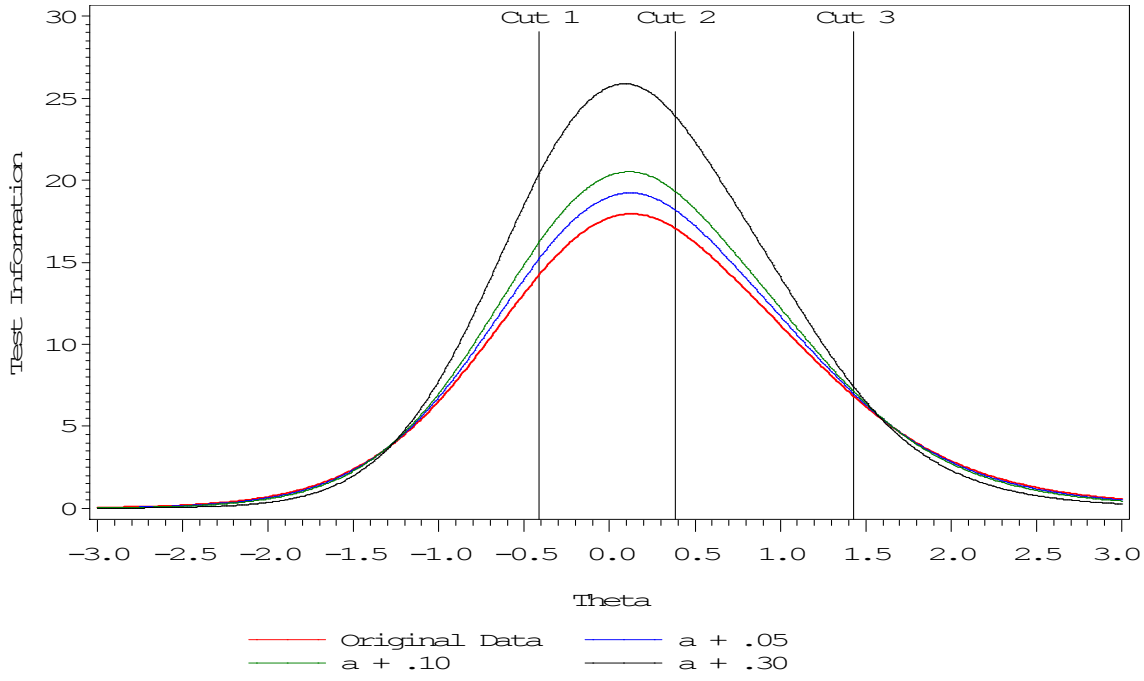


Figure 4.20 High school ELA test – increasing discriminating power: Test information based on multiple choice items only (36 items, maximum score = 36)

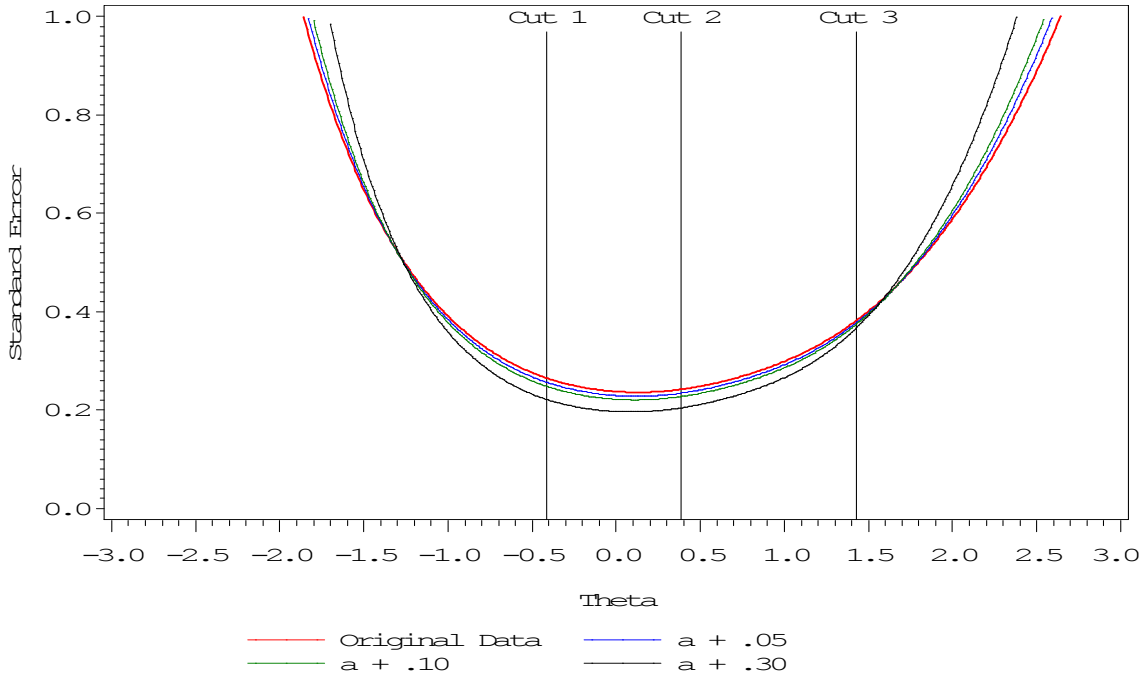


Figure 4.21 High school ELA test – increasing discriminating power: Conditional standard error of measurement based on multiple choice items only

Figure 4.22 below displays the relative efficiency of each of the three improved tests versus the original test.

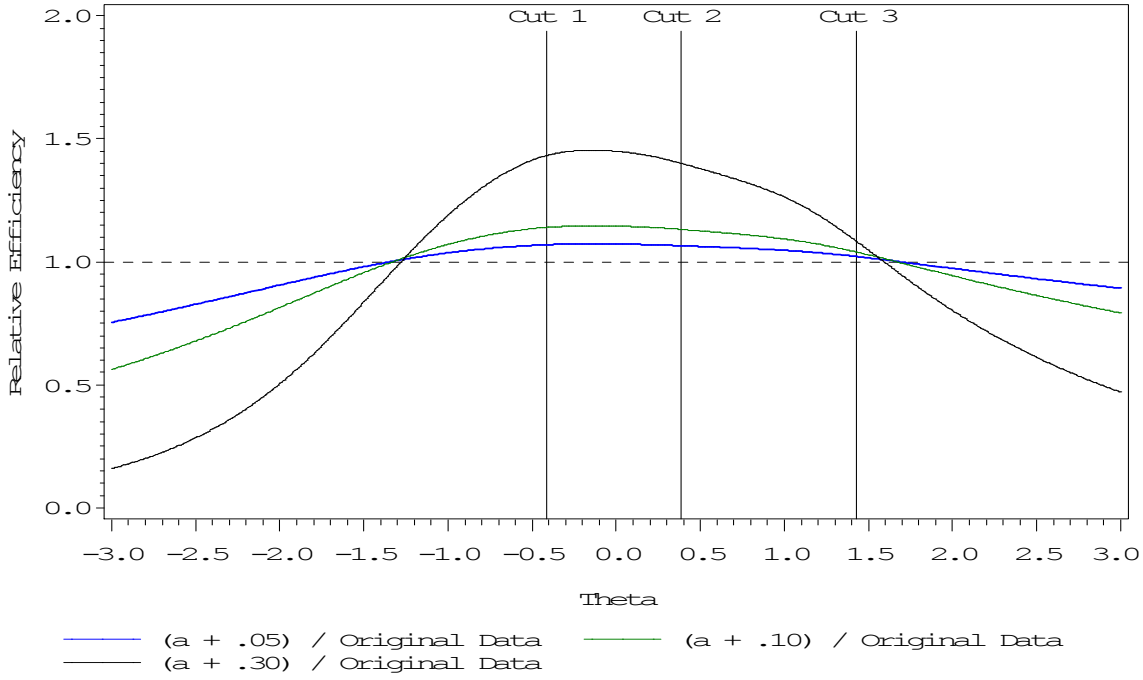


Figure 4.22 High school ELA test – increasing discriminating power: Relative efficiency based on multiple choice items only

The level of information at the three cutscores based on the original item parameter estimates of MC items only was approximately 14.0, 17.0 and 7.0, and the corresponding standard errors of measurement were .27, .24, and .38 at Cut 1, Cut 2 and Cut 3, respectively. Increasing the discriminating power of the MC items by .05 increased the information at the three cutscores to about 15.5, 18.5, and 7.0. When the discriminating power of the MC items was increased by .10, the amount of information at the three cutscores became 16.5, 20.0, and 7.0. Increasing the average item discrimination by .30 for MC items increased the information to 20.5 at Cut 1, 24.0 at Cut 2 and 7.5 at Cut 3. Test information function based on MC items peaked at proficiency score at about .15, and increasing the discriminating power of test items did not affect the location of where the maximum information function occurred.

Results from the conditional standard error of measurement confirmed that increasing the discriminating power generally decreased the standard error of measurement, thus, providing higher measurement precision. However, as shown in Figure 4.21, increasing the discriminating power of items did not guarantee for lower measurement error. When ability parameters were above 1.6 or below -1.3, measurement precision was actually lower when items had higher discrimination value.

Increasing the item discriminating power by .05 on the MC items pushed the new test to have an effective test length of about 5% more than the original test. An increase of .10 in the average *a*-parameter made the effective length of the new test about 11% longer. Finally, with an increase of .30, test with MC items only was more effective at Cut 1 and Cut 2: the effective length was about 44% longer than the original MC items at

Cut 1 and about 40% longer than the original MC items at Cut 2. At Cut 3, the effective length was only about 10% longer than the original test when *a* was increased by .30.

4.2.4.2 Effects of Increasing Discriminating Power on the Constructed Response Items

Figures 4.23 and 4.24 present the information functions and the conditional standard errors for the CR items in high school ELA test. Each of the figure contains the information or conditional standard errors based on the original item parameter estimates and the three levels of increase in discrimination power: *a* + .05; *a* + .10; and *a* + .30. Figure 4.25 reports the relative efficiency of the CR items for the original test compared to the three improved tests.

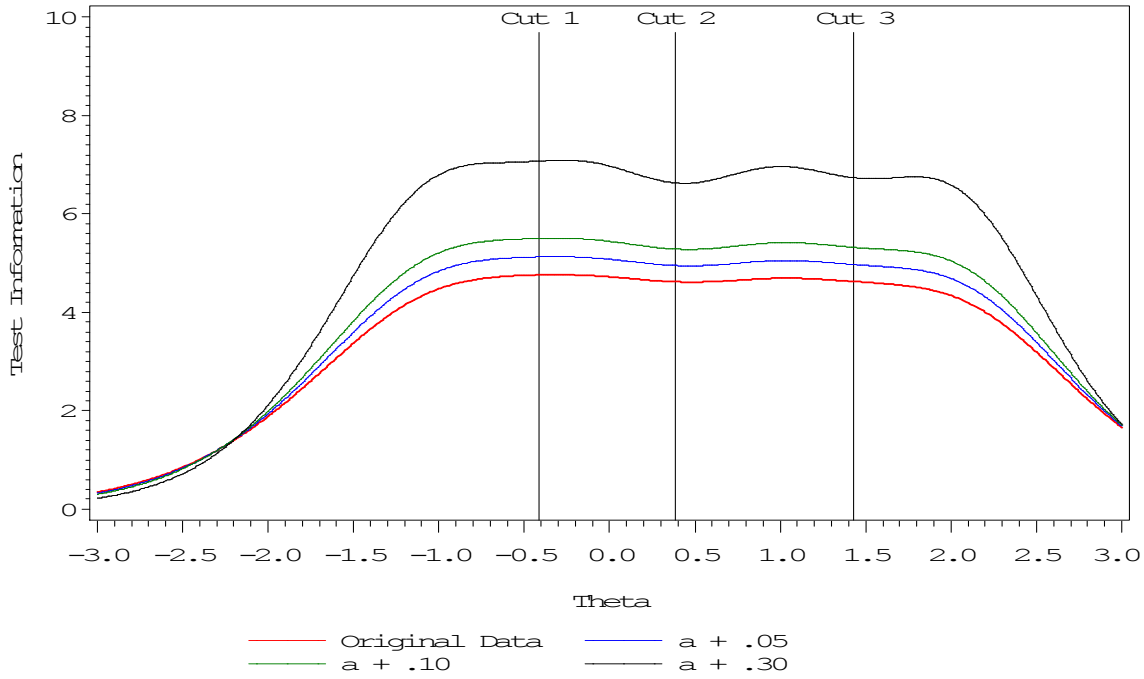


Figure 4.23 High school ELA test – increasing discriminating power: Test information based on constructed response items only (4 items, maximum score = 16)

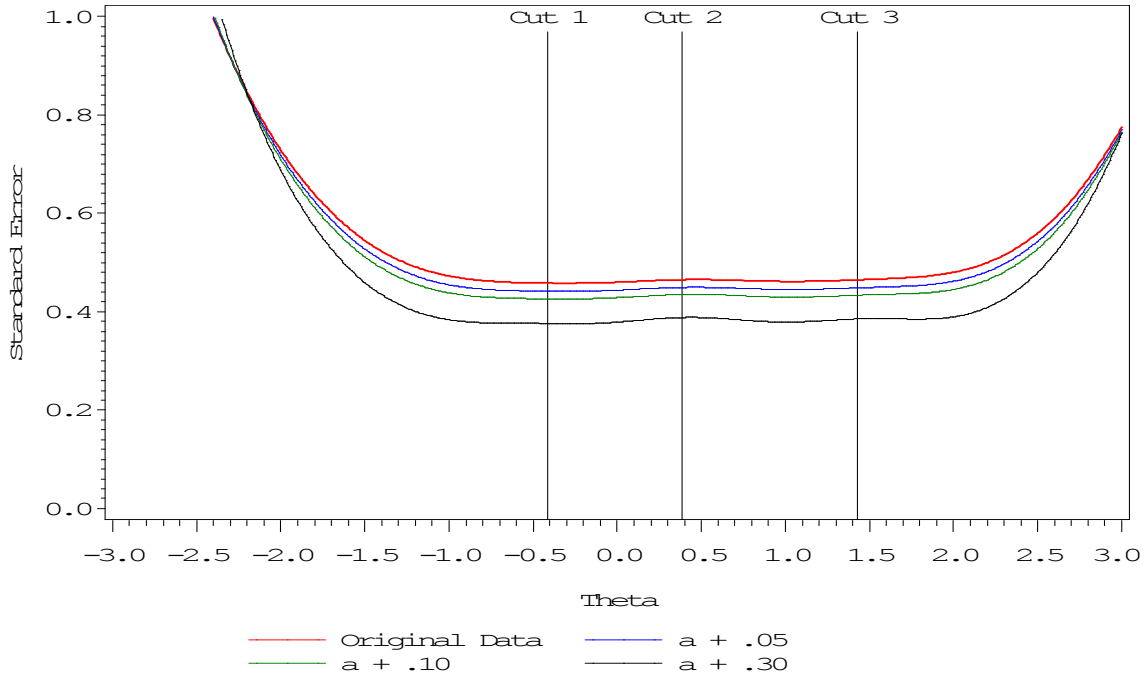


Figure 4.24 High school ELA test – increasing discriminating power: Conditional standard error of measurement based on constructed response items only

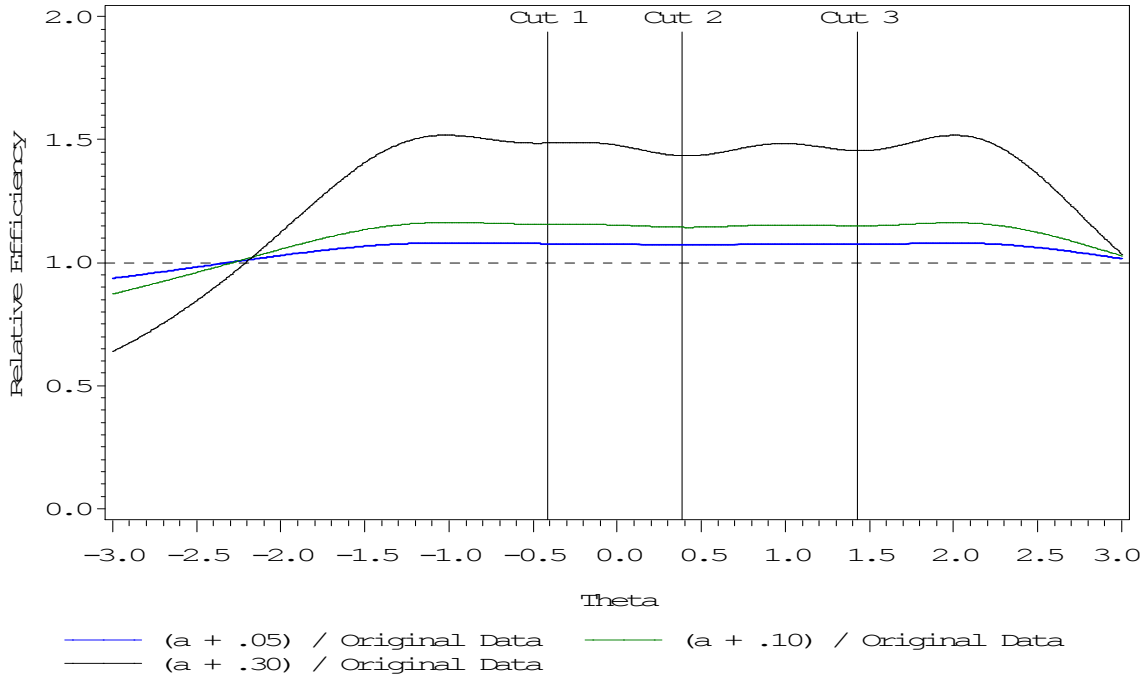


Figure 4.25 High school ELA test – increasing discriminating power: Relative efficiency based on based on constructed response items only

The amount of information provided by the three cutscores was comparable within each version of the test. The level of information at the three cutscores for CR

items based on the original item parameter estimates was approximately 4.75, 4.50 and 4.50 and their corresponding standard errors of measurement were .46, .47, and .47 at Cut 1, Cut 2 and Cut 3, respectively. In general, increasing the item discrimination value increased the amount of information and lowered the measurement error at a constant rate across the three cutscores. However, when average *a* was increased by .30, test information function became multi-modal.

An increase of .05 in the *a*-parameter estimates increased the effective length relative to the original CR items at the cutscores by about 8%; an increase of .10 increased the effective length relative to the original items by about 15%; and an increase of .30 in the average *a*-parameter of the CR items made the effective length to be about 47% longer than the original CR items. In the case where the average *a* of the CR items was increased by .30, this new test was more effective for proficiency scores between -2.2 to 3.0, when proficiency scores were below -2.2, this test became least efficient.

4.2.4.3 Effects of Increasing Discriminating Power on the Essay Items

Figures 4.26 and 4.27 display the information functions and the conditional standard errors of measurement for the essay items (EI) from the high school ELA test.

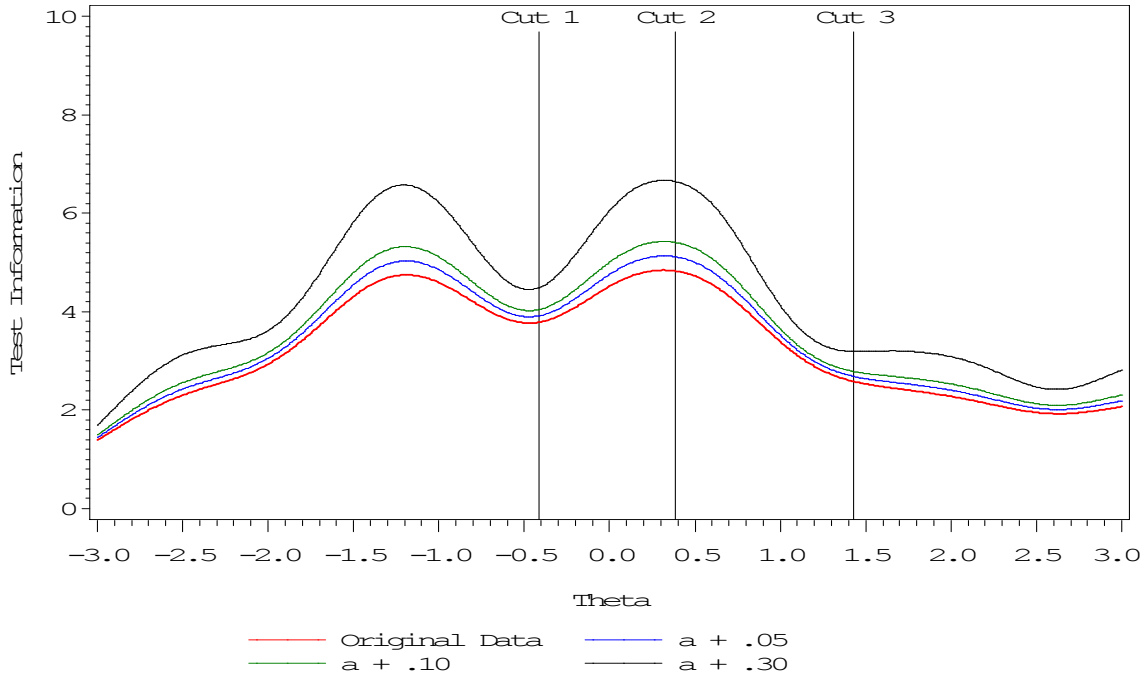


Figure 4.26 High school ELA test – increasing discriminating power: Test information based on essay items only (2 items, maximum score = 16)

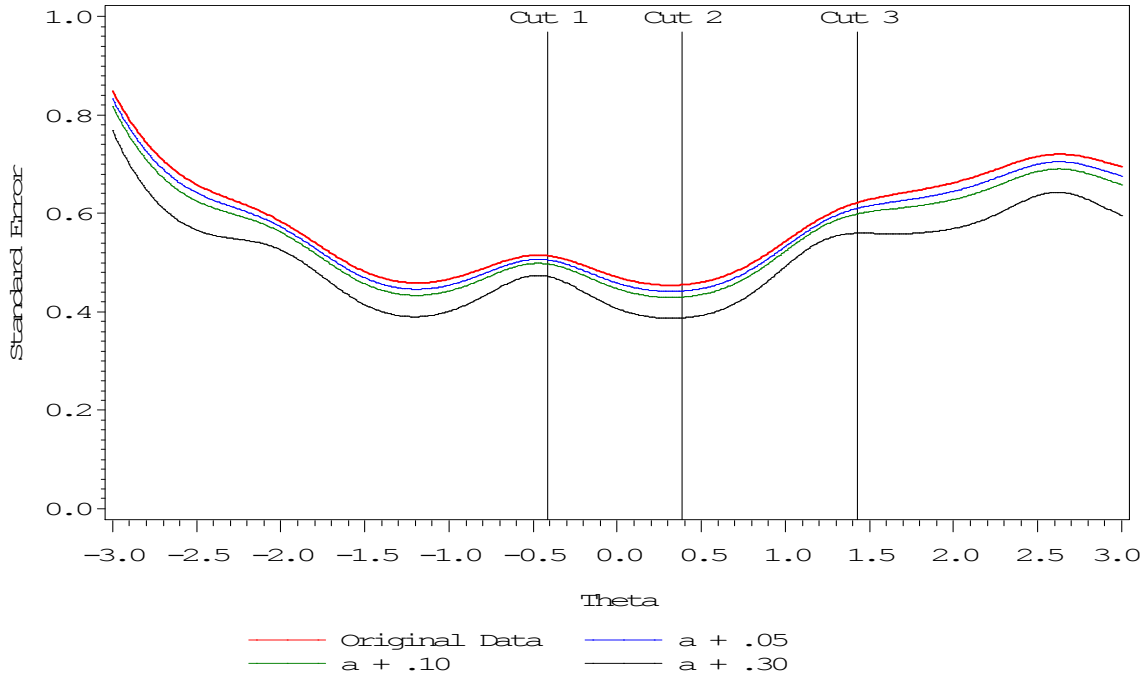


Figure 4.27 High school ELA test – increasing discriminating power: Conditional standard error of measurement based on essay items only

Test information functions based on EI for the high school ELA test were bimodal: it first peaked at around a proficiency score = -1.2 then it dropped and peaked

again at a proficiency score at about .30. The amount of information provided at the first and the third cutscore were lower. The level of information at the three cutscores for EI based on the original item parameter estimates was approximately 3.75, 4.75 and 2.50, their corresponding standard errors of measurement were .52, .46, and .63 at Cut 1, Cut 2 and Cut 3, respectively. Increasing item discrimination value increased the amount of information and also lowered the measurement error but it did not affect the location where the maximum information function occurred.

Figure 4.28 reports the relative efficiency for EI for the original test compared to the three improved tests.

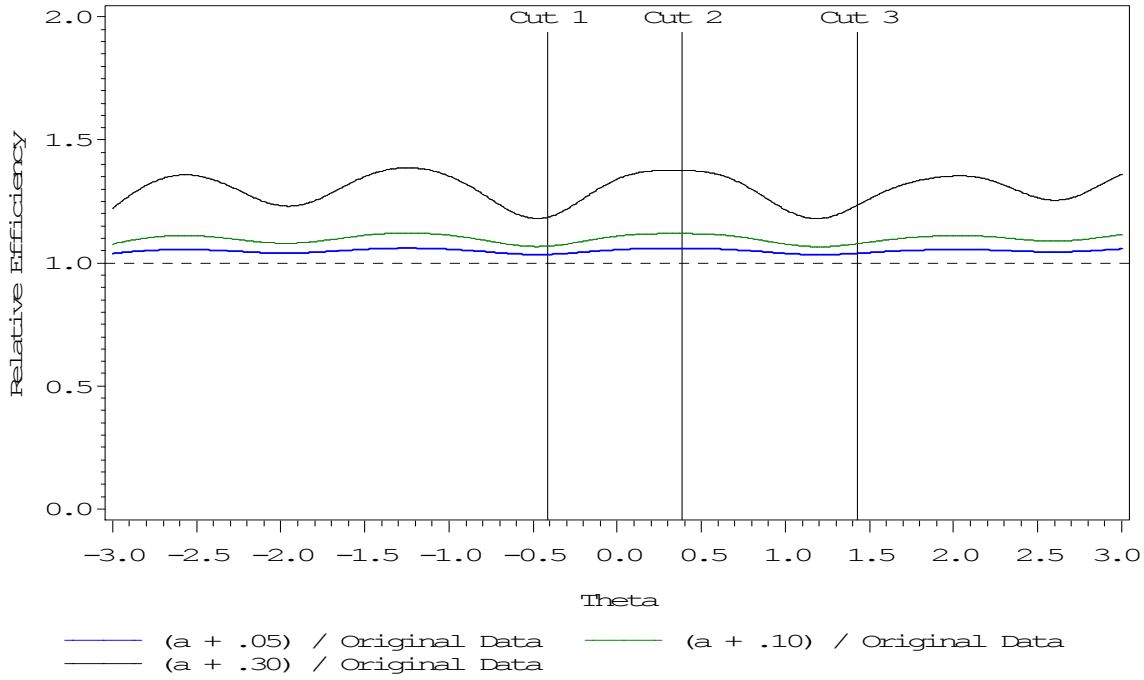


Figure 4.28 High school ELA test – increasing discriminating power: Relative efficiency based on constructed response items only

An increase of .05 in the *a*-parameter estimates increased the effective length relative to the original EI at the cutscores by about 5%; an increase of .10 in *a* made the effective length relative to the original items by about 9%; and an increase of .30 in the

average *a*-parameter of the EI made the effective length to be about 27% longer than the original EI.

4.2.4.4 Effects of Increasing Discriminating Power on the Overall Test

Figures 4.29 and 4.30 compare the information functions and the standard errors for the overall high school ELA test based on original parameter estimates and when the item discriminating parameters were increased by .05, .10 and .30. Figure 4.31 displays the relative efficiency of each of the three improved tests versus the original test.

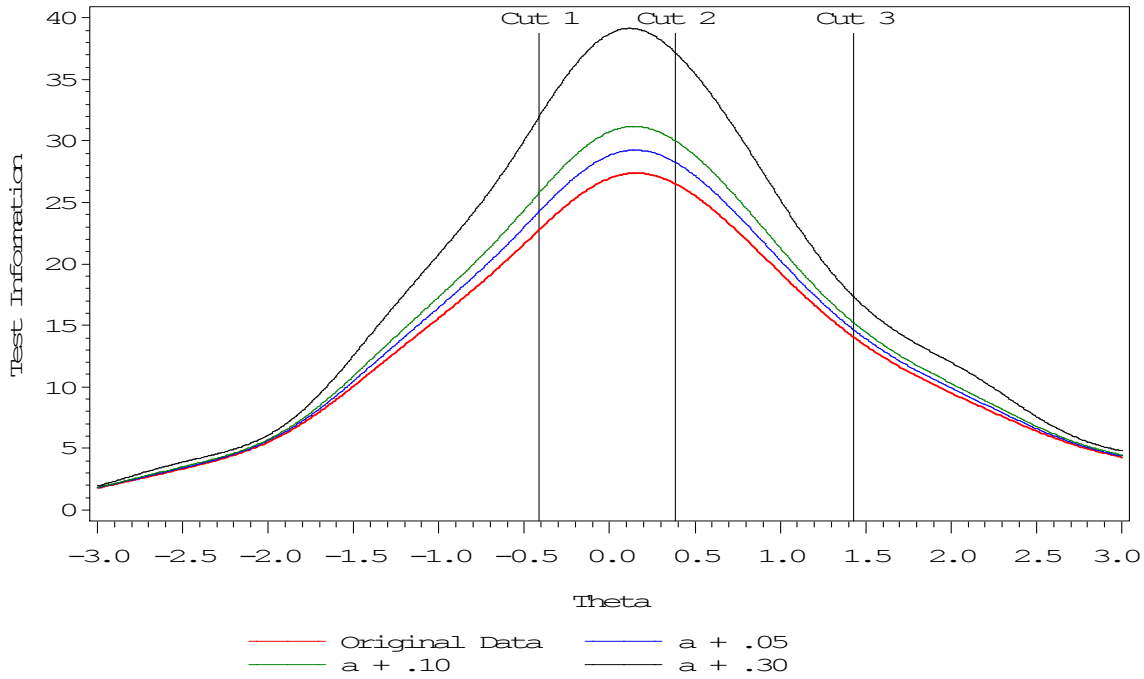


Figure 4.29 High school ELA test – increasing discriminating power: Test information for the overall test and the improved tests (42 items, maximum score = 68)

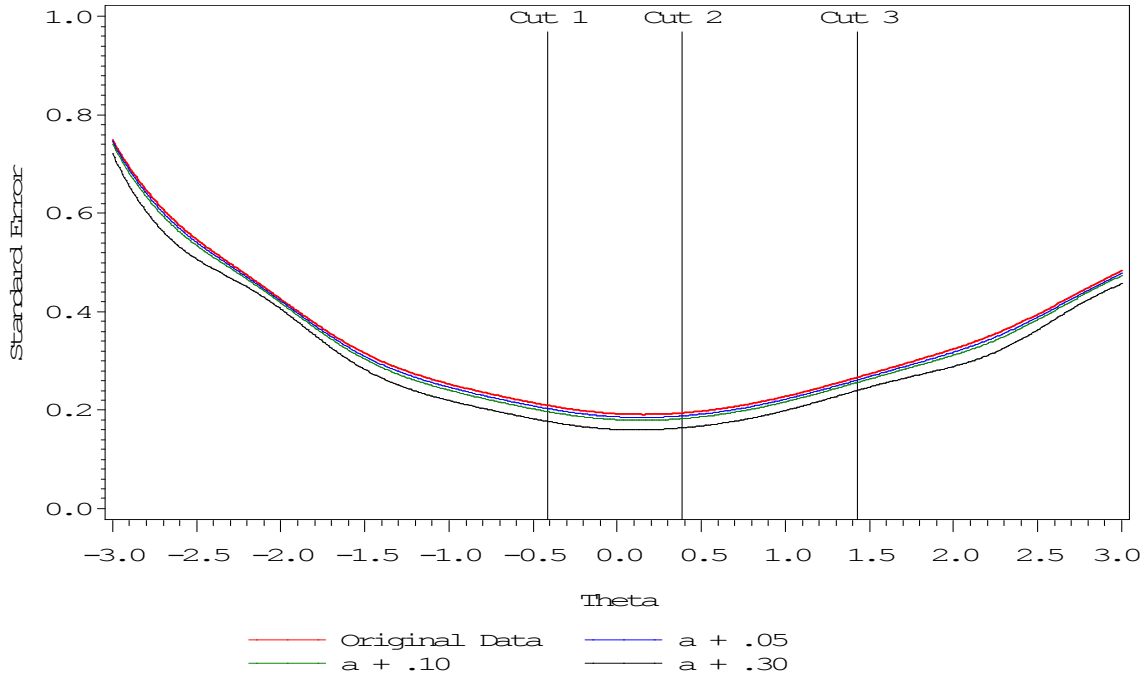


Figure 4.30 High school ELA test – increasing discriminating power: Conditional standard error of measurement for the overall test and the improved tests

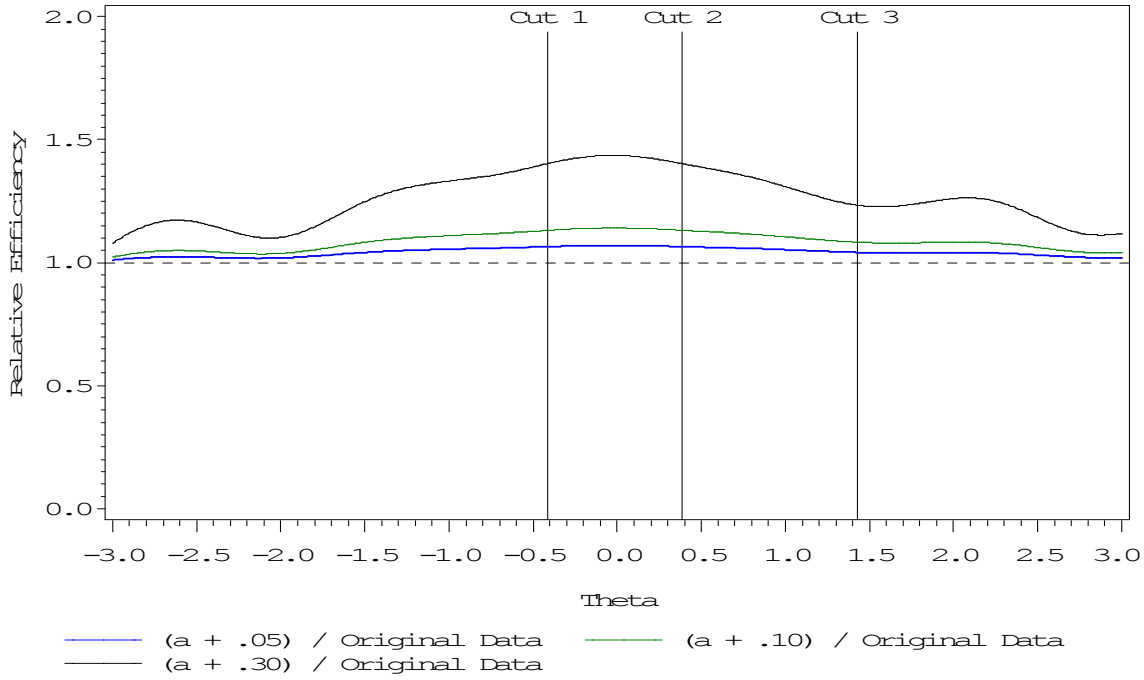


Figure 4.31 High school ELA test – increasing discriminating power: Relative efficiency for the overall test and the improved tests

Again, using information = 10 as a criterion, which corresponds to a classical reliability estimate of about .90, information was substantial at the three cutscores (TIF at

Cut one ≈ 22.5; TIF at Cut 2 ≈ 26.0 and TIF at Cut 3 ≈ 14.0), the corresponding conditional standard errors of measurement were .21, .20 and .27, respectively. The original information function met reasonable expectations at all three cutscores. At the same time, relatively speaking, information was excessive at the first two cutscores and borderline at Cut 3. With an increase of .05 and .10 in the item discrimination indices, the relative efficiency was about 5%, and 13% higher, respectively; with an increase of .30 in the item discrimination indices, relative efficiency was about 1.41 at the two lower cutscores, and about 1.24 at the highest cutscore.

Table 4.3 provides a summary of the item-test score correlations for MC, CR, EI and all test items based on data simulated from 1,000 proficiency scores who were randomly drawn from a standard normal distribution (i.e., N~(0,1)). Table 4.4 highlights that increasing item discrimination impacted on score spread and to a much less extent on the mean test score. This was the expected result.

Increasing the IRT discriminating power by .05 would slightly increase the classical item-test score correlation by about .02; increasing the IRT discriminating power of the test by .10 would lead to an increase in classical item-test score correlation by about .03. Finally, increasing the overall test discriminating power by .30 would lead to an average increase of .06 in item-test score correlations.

Table 4.3 High School ELA Test: Average Classical Item-Test Score Correlations by Item Type and Total Test (N = 1,000).

Item Type[1]	Original Data	a + .05	a + .10	a + .30
MC	.356	.374	.388	.418
CR	.584	.609	.618	.659
EI	.688	.689	.721	.744
All Items	.393	.411	.426	.457

[1] MC – Multiple choice items, CR – Constructed response items, EI – Essay items

Table 4.4 High School ELA Test: Mean and Standard Deviation of Test Scores (N = 1,000).

Average Score (SD)	Original Data	a + .05	a + .10	a + .30
	34.80 (11.57)	34.83 (11.93)	34.80 (12.35)	34.96 (12.95)

4.2.4.5 Effects of Increasing Item Discriminating Power on the Low Discriminating Items

Distribution of the item discrimination powers for the high school ELA test is presented in Figure 4.32.

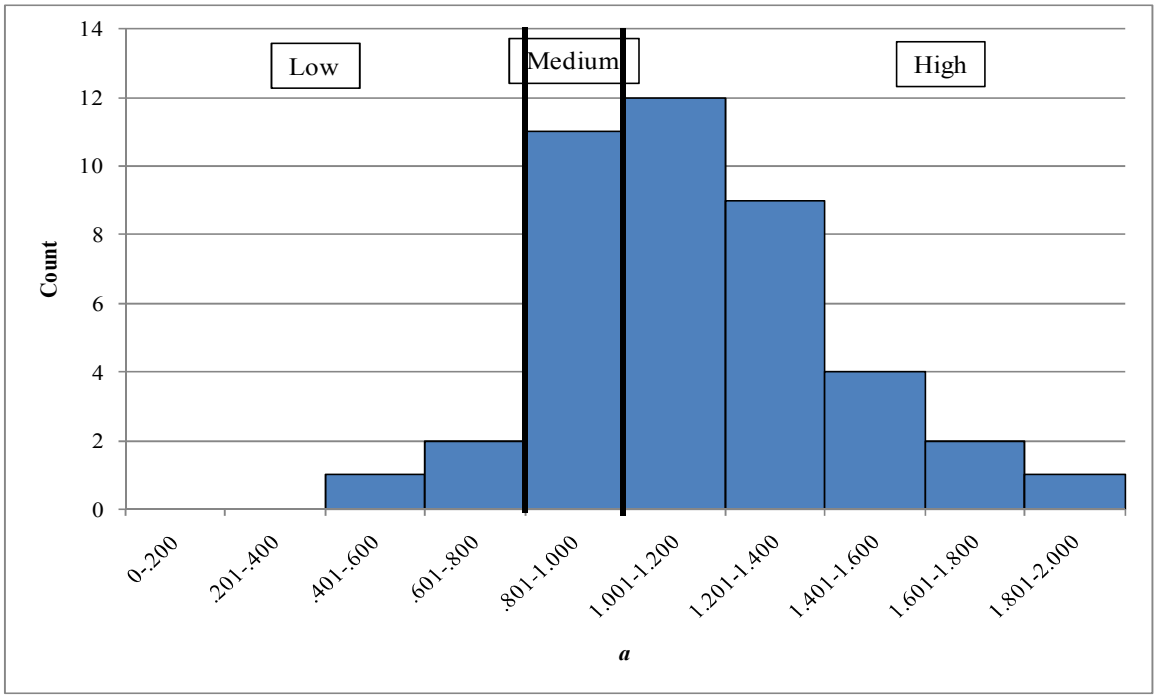


Figure 4.32 High school ELA test: Distribution of the a-parameters

Based on the original item parameter estimates, only three items in this test were categorized in the low discrimination group (i.e., when $a_i < .80$), and they were all multiple choice items.

Figures 4.33 and 4.34 compare the information functions and the standard errors for the overall test based on original parameter estimates and when the item discriminating parameters for items in the low discrimination group were increased by

.05, .10 and .30. Figure 4.35 displays the relative efficiency of each of the three improved tests versus the original test.

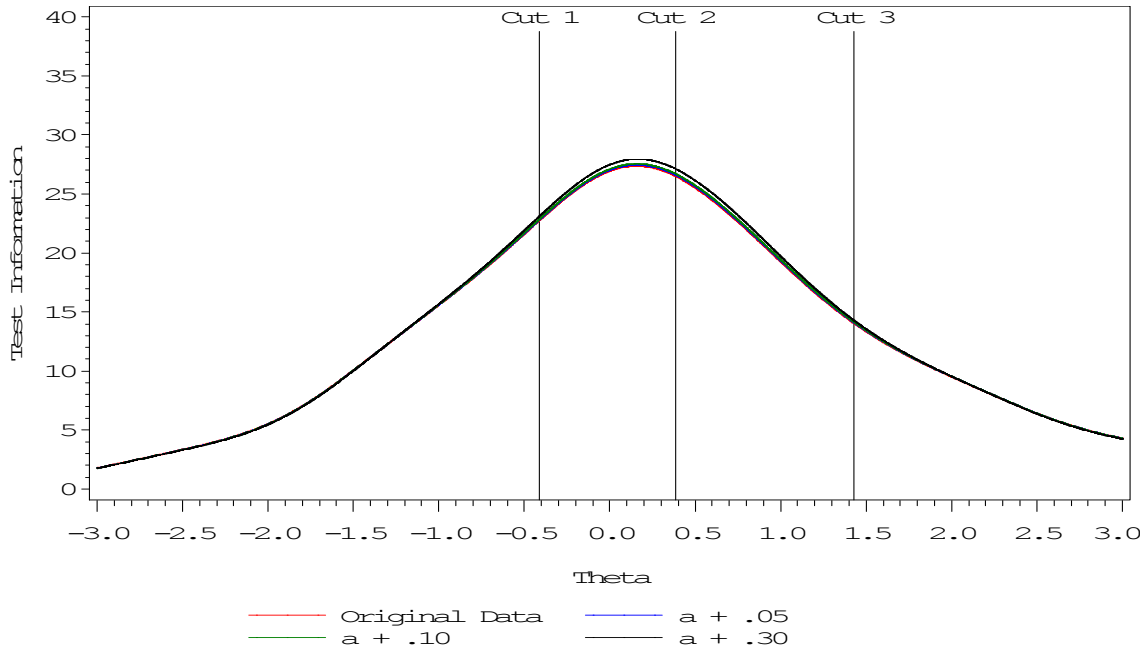


Figure 4.33 High school ELA test – increasing discriminating power: Test information for the overall test and improved item discrimination for low discrimination group (42 items, maximum score = 68)

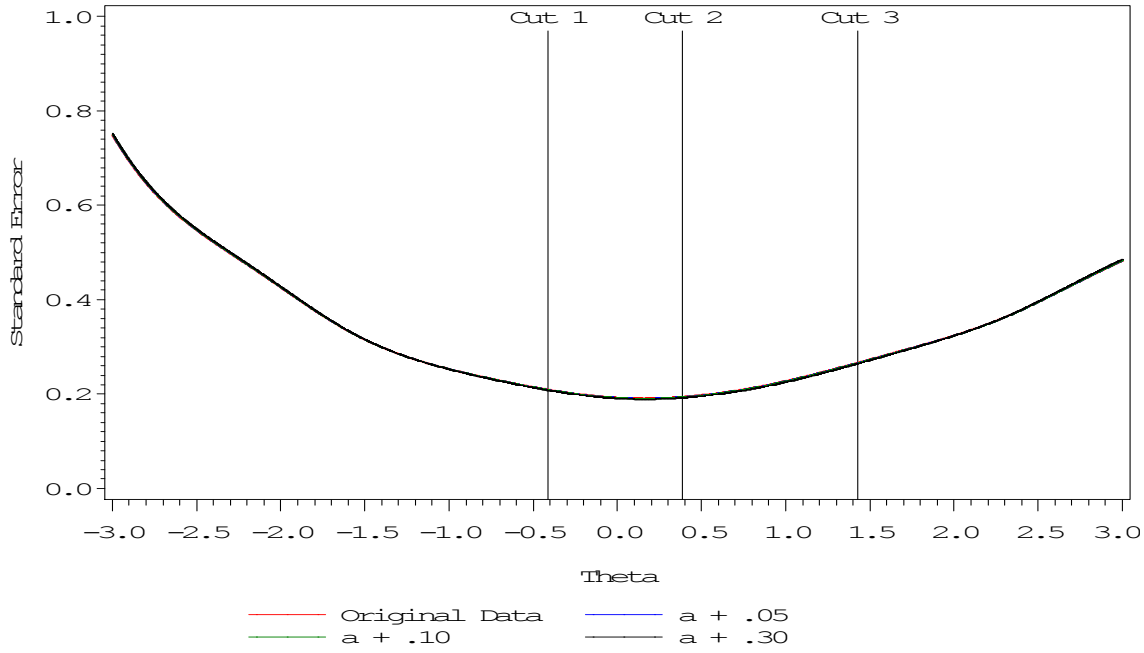


Figure 4.34 High school ELAs test – increasing discriminating power: Conditional standard error of measurement for the overall test and improved item discrimination for low discrimination group

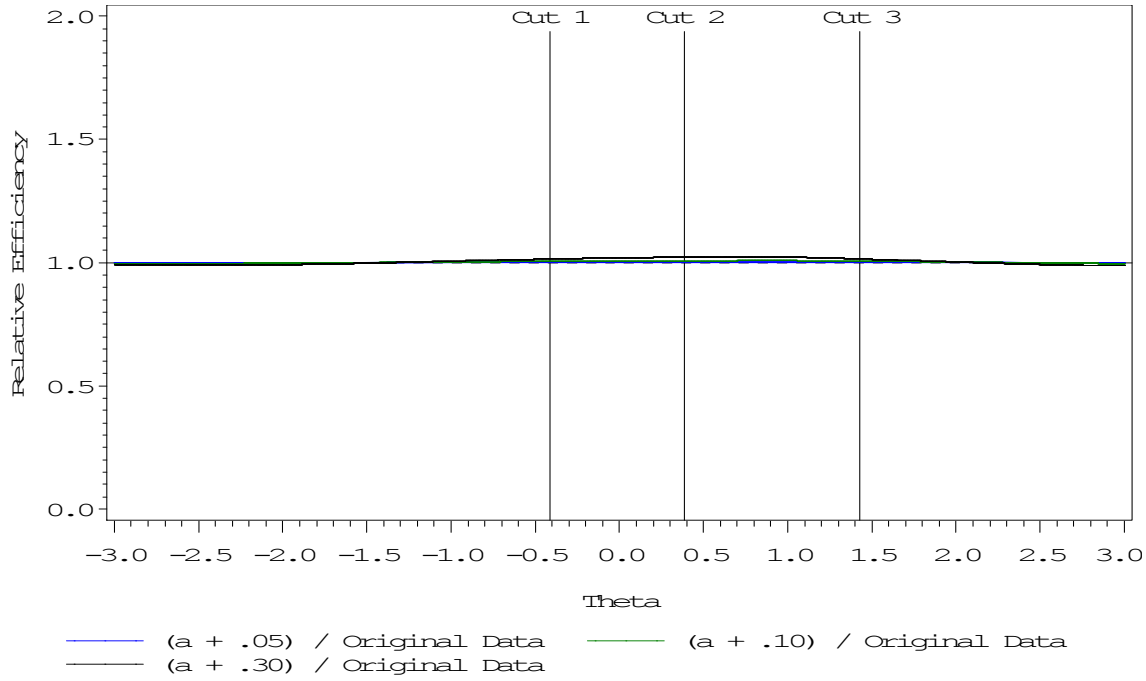


Figure 4.35 High school ELA test – increasing discriminating power: Relative efficiency for the overall test and improved item discrimination for low discrimination group

Test information for the original ELA test was approximately 22.5, 26.0, and 14.0 at Cut 1, Cut 2 and Cut 3, and the corresponding conditional standard errors of measurement for the three cutscores were about .21, .20, and .27. Increasing the item discriminating parameter estimates for items with low *a*-parameters (i.e., the three multiple choice items) by .05 and .10 did not increase the test information nor improved the standard errors of measurement. When the discriminating power was increased by .30 for the three MC items, slight improvement in test information could be observed for proficiency scores between -.40 to approximately 1.40.

Regarding the relative efficiency results, an increase of .05 or .10 in the *a*-parameter estimates for the low discriminating items did not improve test efficiency compared to the original test. However, an increase of .30 in the average *a*-parameter for

the low discriminating items slightly increased the test efficiency compared to the original test for proficiency scores between -.40 to approximately 1.40.

4.2.4.6 Effects of Increasing Item Discriminating Power on the Low and Medium Discrimination Items

Based on the result presented in Figure 4.32, eleven items were classified in the medium discrimination group (i.e., $.80 \leq a_i < 1.0$), and they were all MC items. This section presents the results of test information, conditional standard error of measurement and relative efficiency of increasing the discriminating power of the low and medium discriminating items.

Figure 4.36 to Figure 4.37 compare the information functions, standard errors of the overall test based on original parameter estimates and when the low and medium item discriminating parameters were increased by .05, .10 and .30.

Test information for the original ELA test was approximately 22.5, 26.0, and 14.0 at Cut 1, Cut 2 and Cut 3, and their corresponding conditional standard errors of measurement were .21, .20, and .27. Increasing the low and medium discriminating parameters by .05 or .10 only slightly increased the information at the lower two cutscores. The increase in information was more prominent when the discriminating powers for the low and medium discriminating items were increased by .30; however, the increased were only seen at the lower two cutscores.

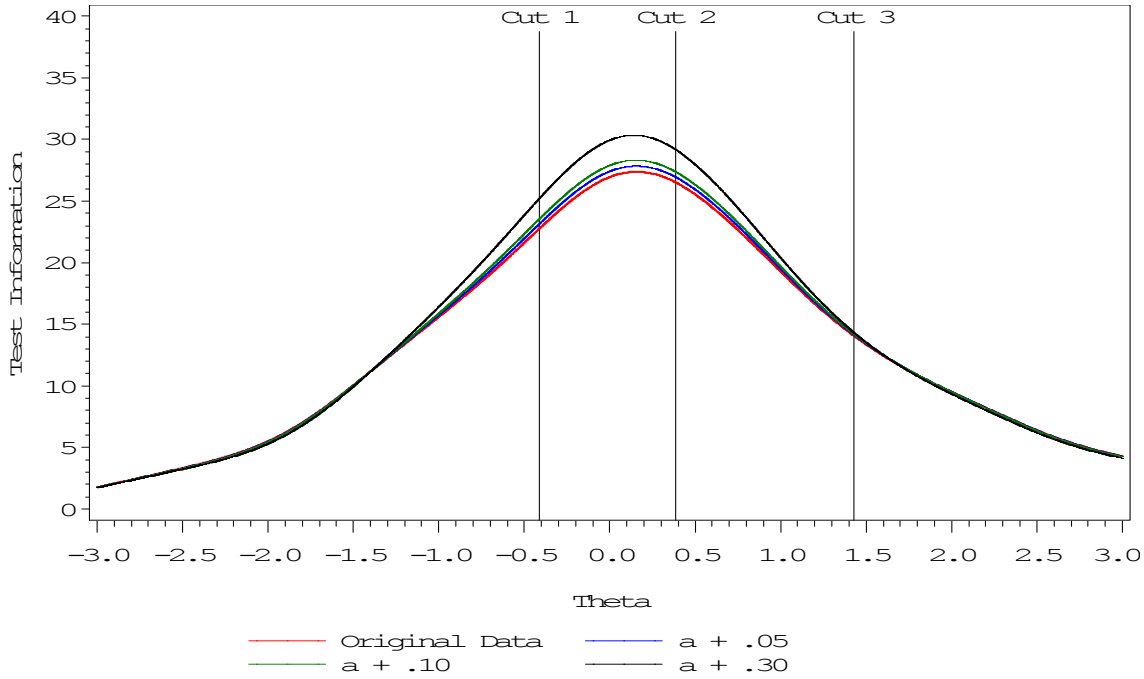


Figure 4.36 High school ELA test – increasing discriminating power: Test information for the overall test and improved item discrimination for low and medium discrimination group (42 items, maximum score = 68)

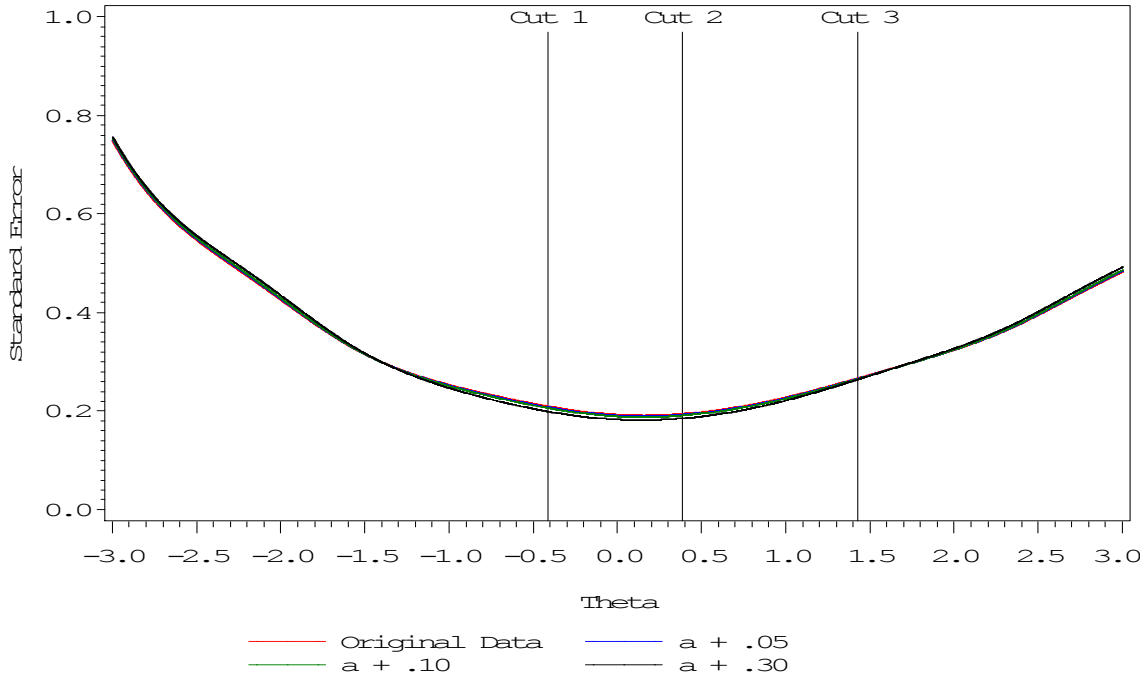


Figure 4.37 High school ELAs test – increasing discriminating power: Conditional standard error of measurement for the overall test and improved item discrimination for low and medium discrimination group

Finally, Figure 4.38 below displays the relative efficiency of each of the three improved tests versus the original test. Efficiency for increasing item discrimination for the low and medium discrimination group by .05 and .10 were almost identical to the original test. However, the new test information obtained from increasing the discriminating power by .30 for the low and medium discriminating items was more effective compared to the original test at certain regions of the proficiency scale. For example, proficiency scores between -1.40 to around 1.50. This new test was actually less efficient compared to the original test at both ends of the proficiency continuum.

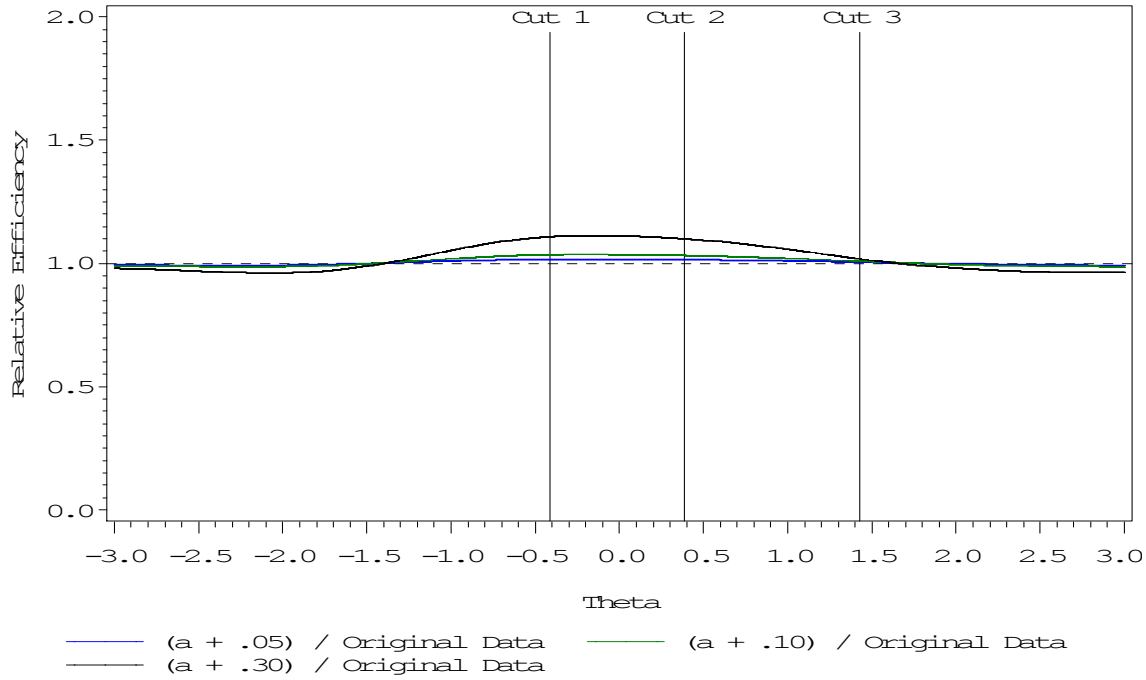


Figure 4.38 High school ELA test – increasing discriminating power: Relative efficiency for the overall test and improved item discrimination for low and medium discrimination group

4.2.4.7 Summary

In summary, MC items in the high school ELA test provided substantial amounts of information at the first two cutscores, but insufficient at the third cutscore. However, the information function was quite well centered. Increasing the discriminating power for

the MC items only help to increase the information at the first two cutscores but it was not essential in this test and it did not help pushing the information up at the third cutscore. For CR items, the information function was rather flat across the proficiency continuum. Increasing discriminating power for the CR items only led to a slight increase in the amount of information when *a* was increased by .05 or .10. Although extreme increases in the discriminating power of the CR items (i.e., *a* + .30) also led to an increase in information across the proficiency continuum, the increment was lower at the second and third cutscore. For the essay items, information functions were bimodal. Regardless of the increase in discriminating power, all information functions peaked at proficiency scores around -1.2 and .30. The original test information for the overall test was quite well centered. When the discriminating power was increased at the overall test level, information increased; however, the addition was not essential. The same conclusion applied to those results obtained from increasing discriminating power for the low discrimination group (i.e., when $a_i < .80$) and increasing discriminating power for the low and medium (i.e., $.80 \leq a_i < 1.0$) discrimination group. In all cases, increasing discriminating power decreased the conditional standard error and hence, higher measurement precision. In addition, improved tests were more efficient than the original test when discriminating power was increased. For instance, increasing the discriminating power by .05 made the effective test length relative to the original test by about 5%, regardless of whether the increase was coming from increasing the discriminating power for the overall test, by item format or only from low or low and medium discriminating items.

4.3 Study One – Changes in Item Difficulty Value

4.3.1 Middle School Mathematics Test

Based on the summary of the item parameter estimates for the middle school Mathematics test as presented in Table 3.1, constructed response (CR) items seem to be the easiest followed by multiple choice (MC) items, and the short answer (SA) items are the most difficult among the three item format. In addition, spread of the difficulty for the SA items are wider compare to MC and CR items.

4.3.1.1 Effects of Changing Difficulty Level on the Multiple Choice Items

Figures 4.39 and 4.40 display the information functions and the conditional standard errors for the MC items in middle school Mathematics test. Each figure contains the information or conditional standard errors based on the original item parameter estimates and the three manipulations of the difficulty level: *b* - 1.0; *b* + 1.0; and *b* + 2.0.

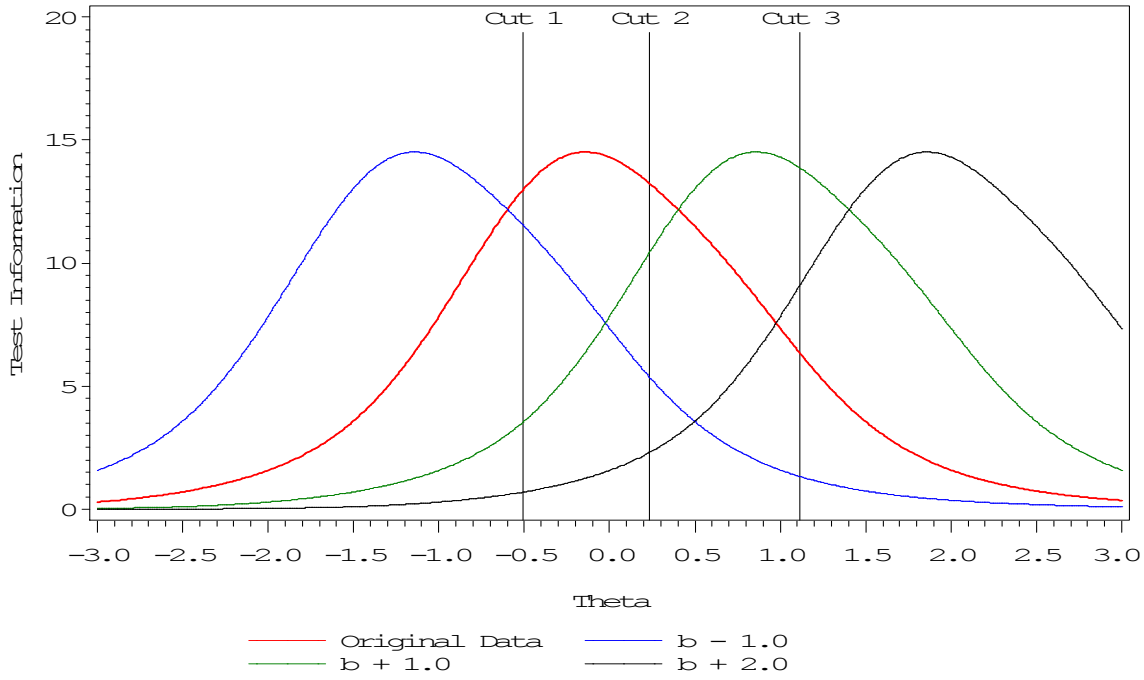


Figure 4.39 Middle school Mathematics test – manipulating difficulty level: Test information based on multiple choice items only (29 items, maximum score = 29)

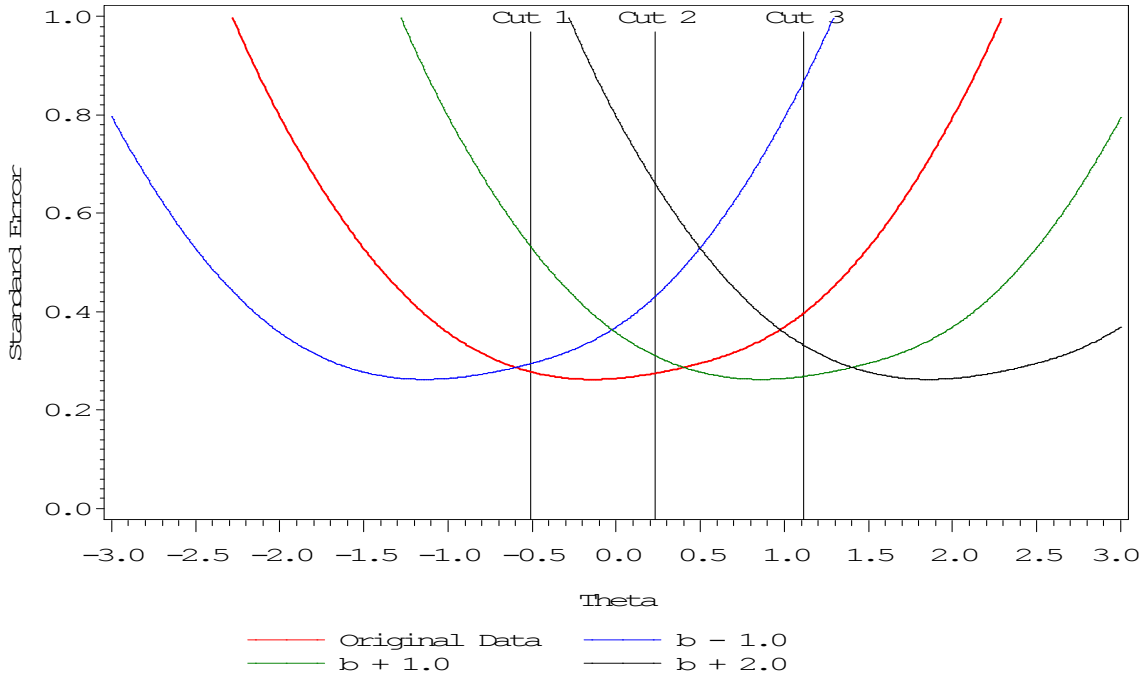


Figure 4.40 Middle school Mathematics test – manipulating difficulty level: Conditional standard error of measurement based on multiple choice items only

The level of information at the three cutscores based on the original item parameter estimates of the MC items only was approximately 13.0, 13.0 and 6.5, and their corresponding standard errors of measurement were .28, .28, and .39, respectively. The information function peaked at proficiency score at about -.20. Decreasing the average difficulty of the MC items by 1.0 shifted the information function to the left of the proficiency continuum by 1 unit, thus the new information function peaked at around -1.2. And the information at the three cutscores became 11.5, 5.5 and 1.5 which corresponded to standard errors of measurement of .29, .43, and .82, respectively. When the average difficulty of the MC items was increased by 1.0, information at the three cutscores became 3.5, 10.0, and 14.0, and their corresponding standard errors of measurement were .53, .32, and .27. The information function was shifted to the right of the proficiency continuum and it peaked at a proficiency score = .80. Extreme increase in the average difficulty of the MC items (i.e., $b + 2.0$) further shifted the information

function to the right, and now the information function peaked at about a proficiency score = 1.8. Information at the three cutscores were .50, 2.0 and 9.0 and their corresponding standard errors of measurement were 1.41, .71, and .33, respectively.

Based on the results of the conditional standard error of measurement, decreasing the average difficulty of MC items by 1.0 only lowered the measurement errors for proficiency scores below -.60; measurement errors were higher when proficiency scores were above -.60. When average difficulty was increased by 1.0, measurement errors were higher when proficiency scores were below .40, and measurement errors became lower than the original test when proficiency scores were above .40. With an extreme increase in the average difficulty of the MC items (i.e., $b + 2.0$), measurement errors were lower than the original test only when proficiency scores were above 1.0.

Figure 4.41 displays the relative efficiency of each of the three variations of tests versus the original test. When considering only MC items in the middle school Mathematics test, decreasing the *b*-parameter estimates by 1.0 made the new test less efficient than the original test for all three cutscores. Specifically, at Cut 1, the new test only performed 90% as well as the original test; at Cut 2, the performance of the new test was about 40% of the original test; and at Cut 3, the new test only functioned about 20% of the original test. However, this new test was more efficient than the original test for those with proficiency scores below -.40.

An increase of 1.0 in *b* for the MC items made the new test to be less efficient compared to the original test when proficiency scores were below .40. Therefore, the original test was more efficient for cutscores 1 and 2. When proficiency scores were above .40, the new test became more efficient, especially at Cut 3. This new test was two

times more efficient than the original test. Notice that when proficiency scores were above 2.2, the relative efficiency started to decline.

Increasing the average difficulty of the MC items by 2.0 only made the new test more efficient than the original test when proficiency scores were above 1.0. At Cut 3, this new test was about 40% more efficient than the original test.

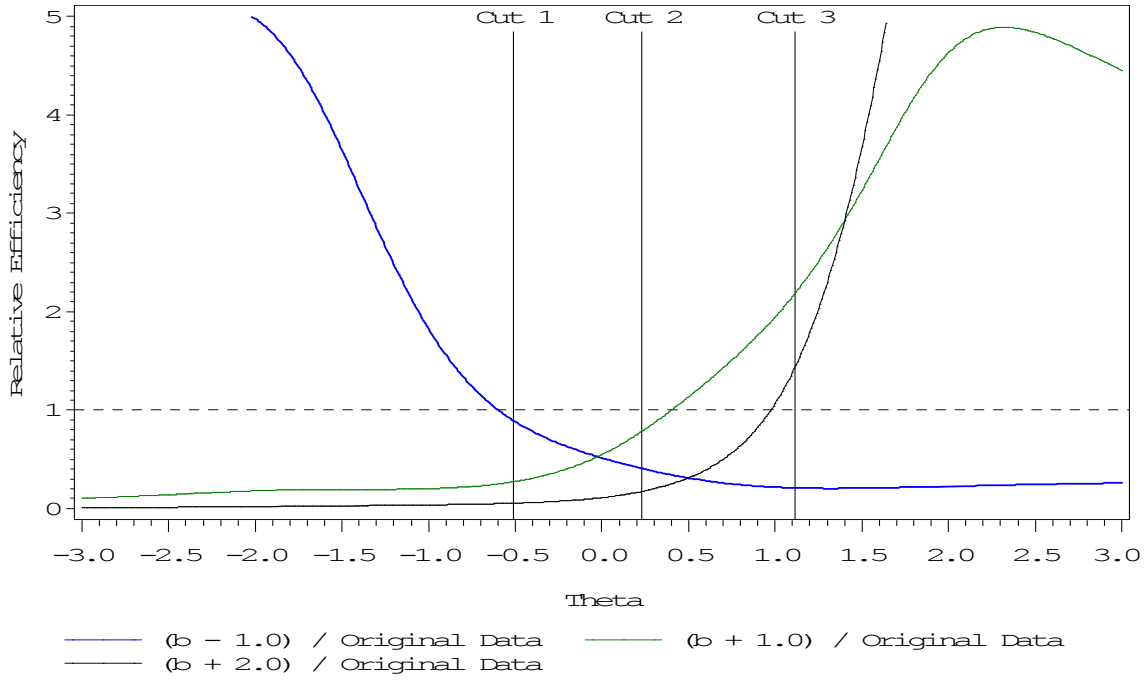


Figure 4.41 Middle school Mathematics test – manipulating difficulty level: Relative efficiency based on multiple choice items only

4.3.1.2 Effects of Changing Difficulty Level on the Short Answer Items

Figures 4.42 and 4.43 present the information functions and the conditional standard errors for the SA items in middle school Mathematics test. Each figure contains information or conditional standard errors based on the original item parameter estimates and three manipulations of average difficulty for the SA items: $b - 1.0$; $b + 1.0$; and $b + 2.0$. Figure 4.44 reports the relative efficiency of the SA items for the original test compared to the three tests.

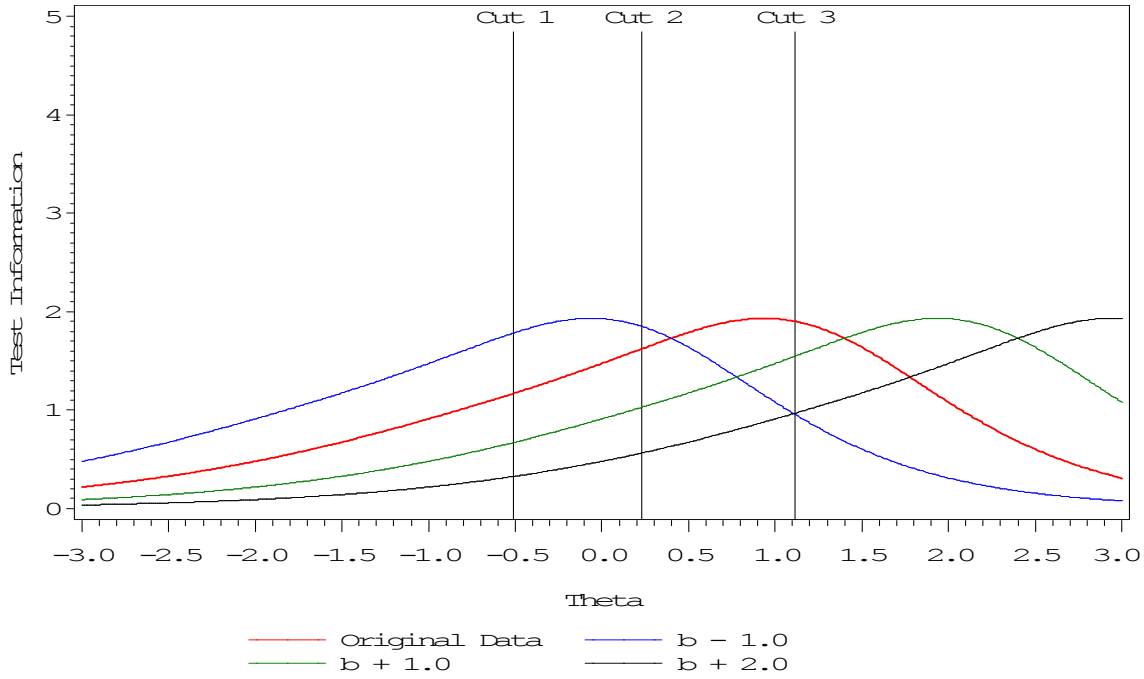


Figure 4.42 Middle school Mathematics test – manipulating difficulty level: Test information based on short answer items only (5 items, maximum score = 5)

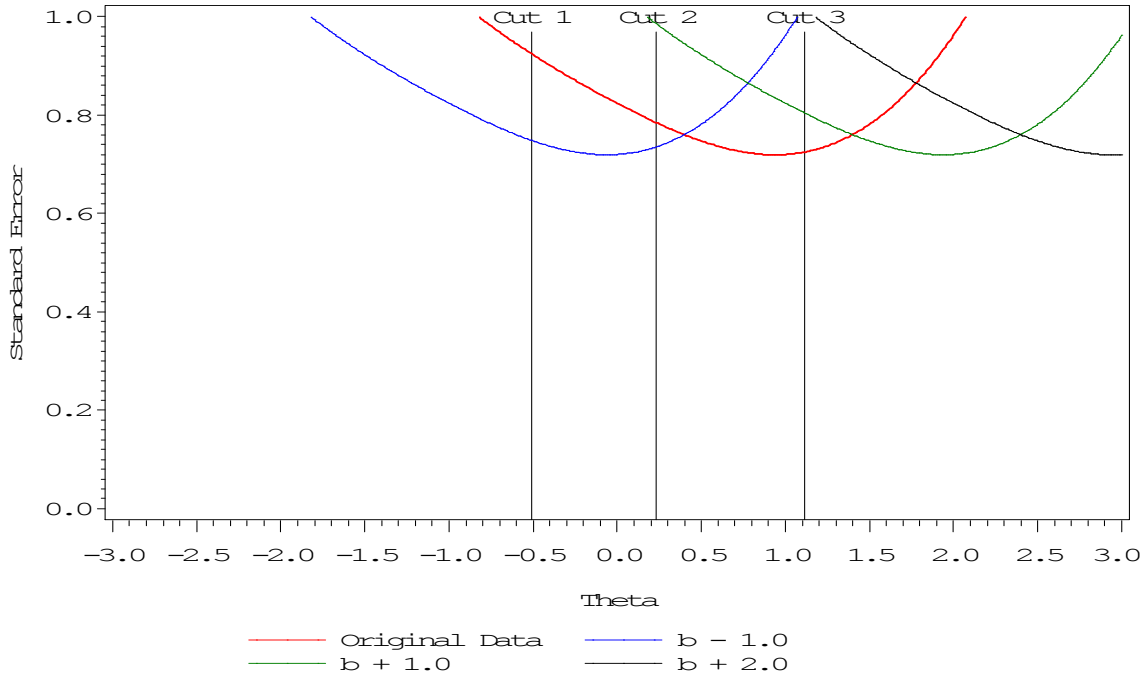


Figure 4.43 Middle school Mathematics test – manipulating difficulty level: Conditional standard error of measurement based on short answer items only

The level of information at the three cutscores for SA items based on the original item parameter estimates was approximately 1.1, 1.5 and 1.8, and their corresponding

standard errors of measurement were .95, .82, and .75. Low level of information and high standard error rate were expected as the results were only based on 5 dichotomously scored items. These items were calibrated by the 2-PLM model, with zero *c*s, information function peaked at a much higher proficiency score (at around $\theta = .90$) than the average difficulty level of these items (average *b* for SA items is .16).

Decreasing the difficulty level of the SA items by 1.0 shifted the information function to the left of the proficiency scale by 1.0, thus, the information function peaked at about -.10 on the proficiency scale whereas the original information function peaked at about .90. In addition, by making the SA items easier, this made the information higher for the lower two cutscores: information at Cut 1 became 1.8 and the corresponding standard error of measurement was .75; and information at Cut 2 became 1.9, the corresponding standard error of measurement was .73. However, information at Cut 3 became 50% lower compared to the original item parameter estimates. With the easier test, information at Cut 3 was only .90 and its standard error of measurement was 1.05.

Based on the results presented in Figure 4.44, decreasing the item difficulty by 1.0 on the SA items only made the new test more efficient at the lower two cutscores: Cut 1 was 60% more efficient and Cut 2 was 20% more efficient than the original test. However, at Cut 3, this new test was 40% less efficient than the original test. When item difficulties were increased by 1.0 or 2.0, the new tests became less efficient than the original test across all three cutscores. These new tests only became more efficient than the original SA items when proficiency scores were above 1.4 (when difficulty was increased by 1.0) and when proficiency scores were above 1.8 (when difficulty was increased by 2.0).

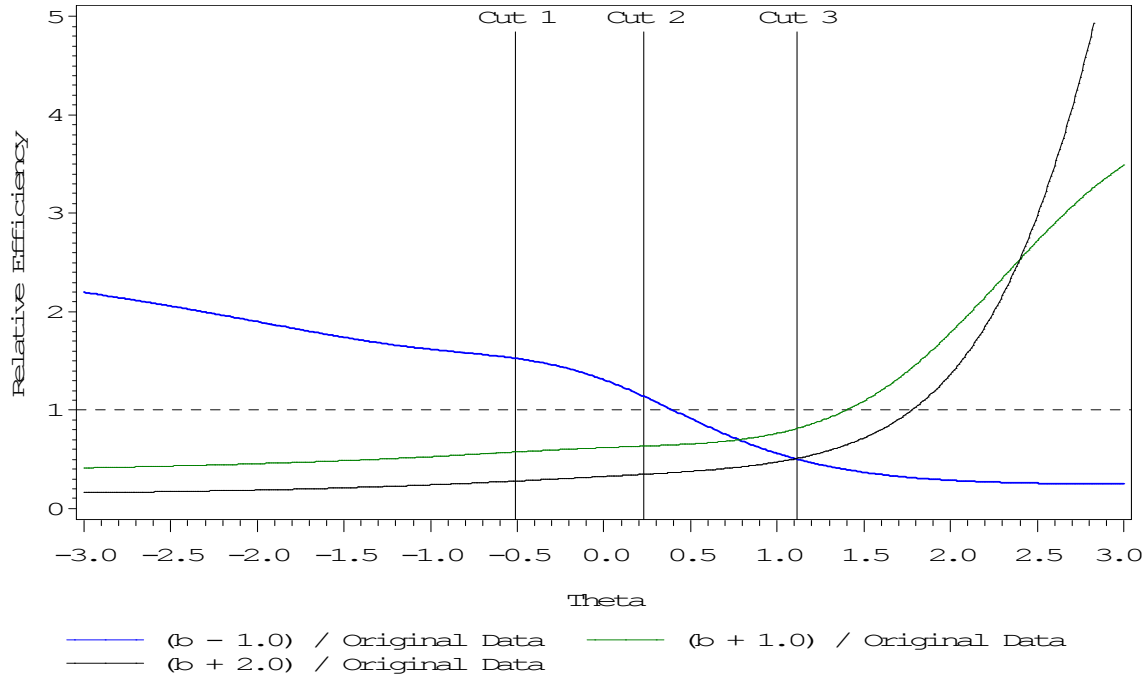


Figure 4.44 Middle school Mathematics test – manipulating difficulty level: Relative efficiency based on short-answer items only

4.3.1.3 Effects of Changing Difficulty Level on the Constructed Response Items

Figure 4.45 reports the information functions for the CR items from the middle school Mathematics test. The amount of information provided by the first cutscore was quite similar between the original parameter estimates and when the difficulty of the CR items was increased by 1.0. The amount of information at Cut 1 was considerably less when items were less difficult (i.e., when $b - 1.0$) or when item were substantially more difficult (i.e., when $b + 2.0$). At Cut 2, increasing item difficulties by 1.0 increased the amount of information by about .30 from the original parameter estimates. The amount of information provided by the original test and from the extremely difficult set of CR items (i.e., $b + 2.0$) were roughly equal. However, making the items easier (i.e., $b - 1.0$) decreased the amount of information at this cutscore. Finally, increasing the difficulty of the items increased the amount of information at Cut 3 but when difficulty of the items was decreased (i.e., $b - 1.0$), the amount of information at this cutscore also decreased.

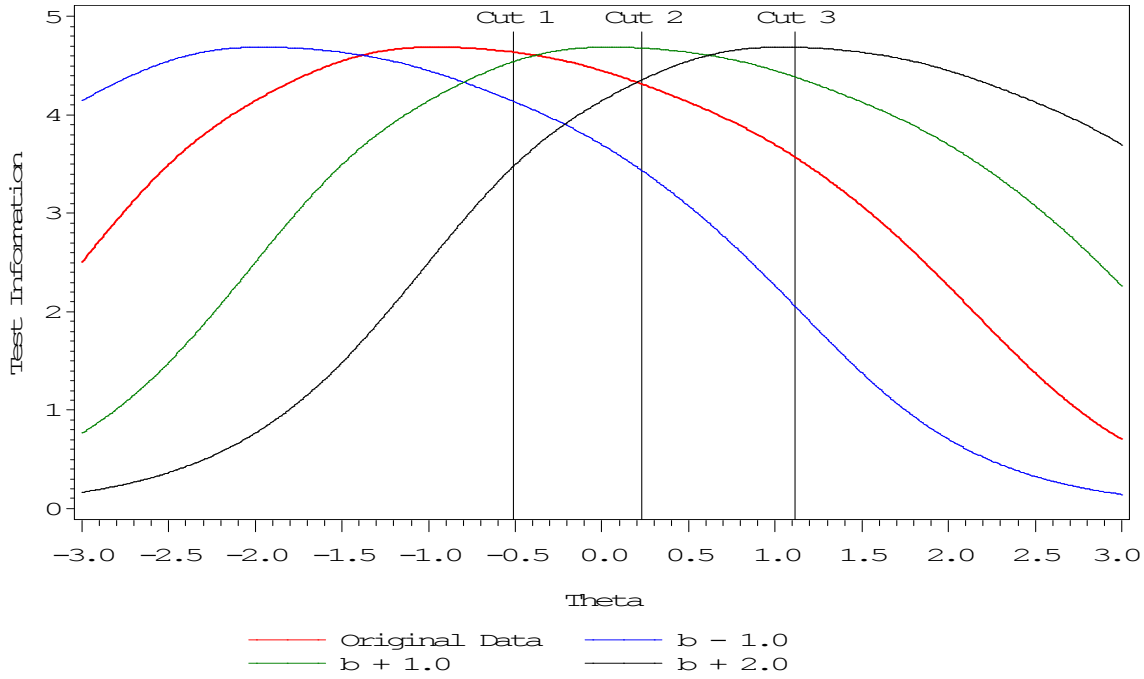


Figure 4.45 Middle school Mathematics test – manipulating difficulty level: Test information based on constructed response items only (5 items, maximum score = 20)

Figure 4.46 displays the conditional standard errors of measurement for the CR items. The conditional standard errors of measurement were comparable across different versions of tests. At Cut 1, errors were a little bit higher when the test was most difficult (i.e., $b + 2.0$). The errors obtained from the original test and when the average difficulty was 1.0 unit higher were almost identical at this cutscore. Errors were only slightly higher when the CR items were easier. At Cut 2, the conditional standard errors of measurement were identical to those obtained from the original test and when there was an extreme increase in the item difficulty (i.e., $b + 2.0$). The standard error rate slightly decreased when item difficulties were increased by 1.0, but the standard error was higher than the original test when items were easier. Increasing item difficulties improved the measurement precision at Cut 3; however, making these items easier lowered the measurement precision at this particular cutscore.

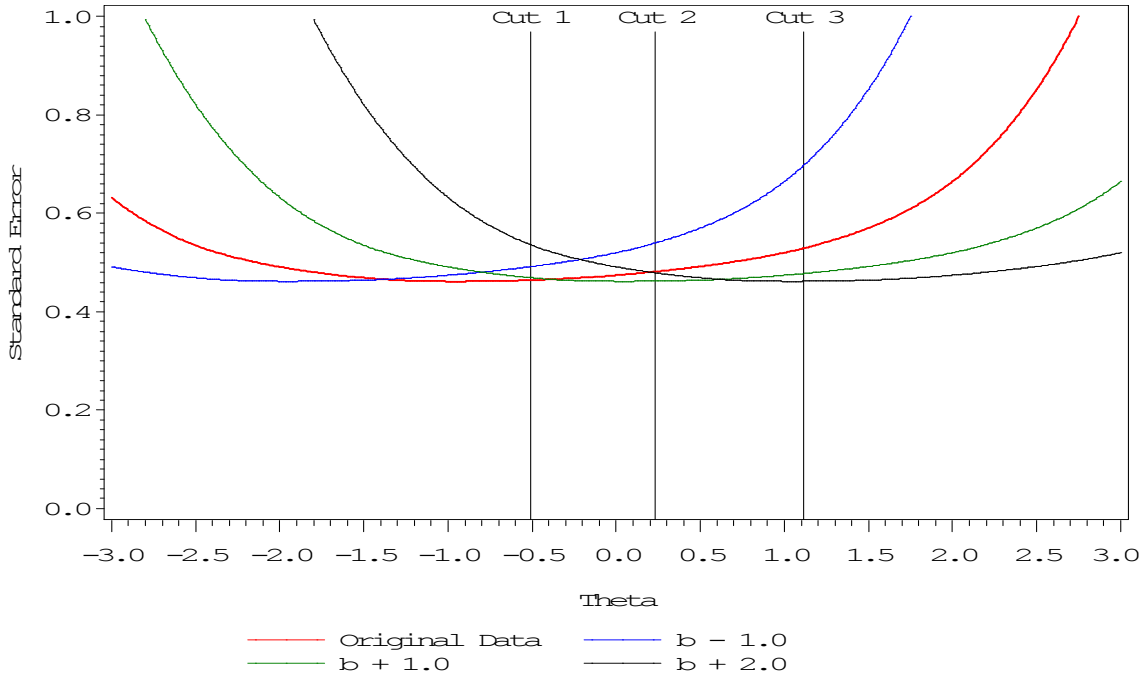


Figure 4.46 Middle school Mathematics test – manipulating difficulty level: Conditional standard error of measurement based on constructed response items only

In general, decreasing the difficulty of the CR items was not recommended as this made the CR items less effective at all cutscores (See Figure 4.47 below). Increasing the difficulty of the CR items, on the other hand, made the new test as effective or more effective compared to the original test. For example, when the average difficulty of the CR items was increased by 1.0 unit, at Cut 1, the new test was as effective as the original test. However, at Cut 2 and Cut 3, the new test was 10% and 20%, respectively, more effective than the original test. With an extreme increase in item difficulty, the original test functioned better at Cut 1, the new test functioned as good as the original test at Cut 2, but 30% more effective than the original test at Cut 3.

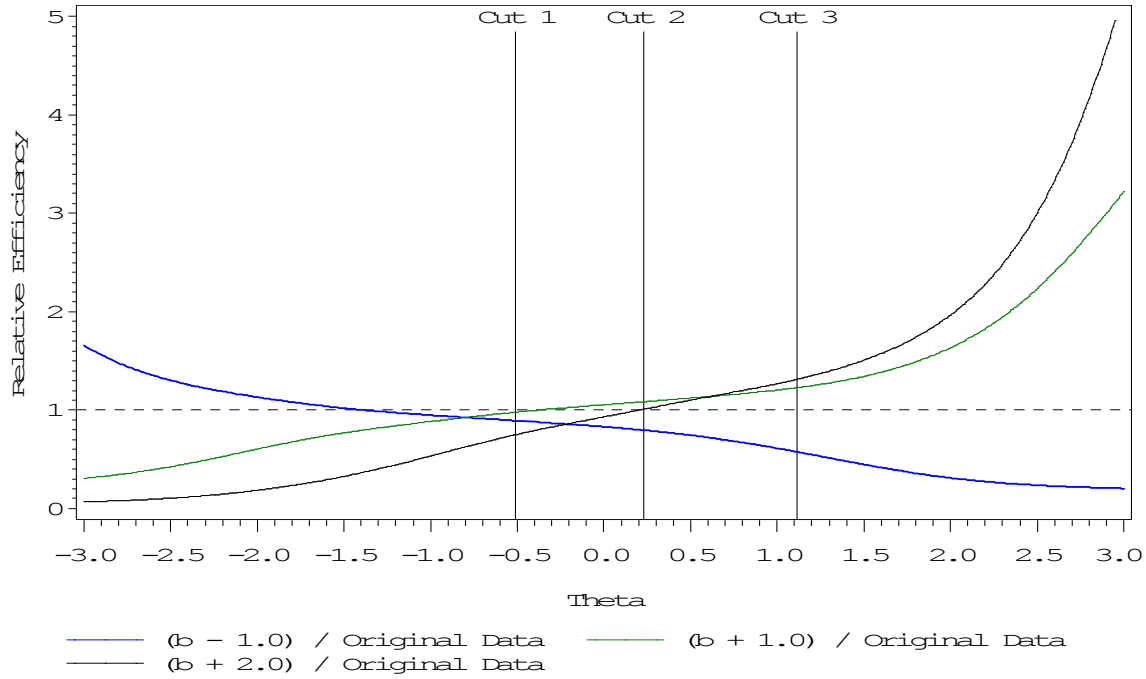


Figure 4.47 Middle school Mathematics test – manipulating difficulty level: Relative efficiency based on constructed response items only

4.3.1.4 Effects of Changing Difficulty Level on the Overall Test

Figures 4.48 and 4.49 compare the information functions and the standard errors for the overall test based on original parameter estimates and when the item difficulty values were decreased by 1.0, increased by 1.0 and 2.0.

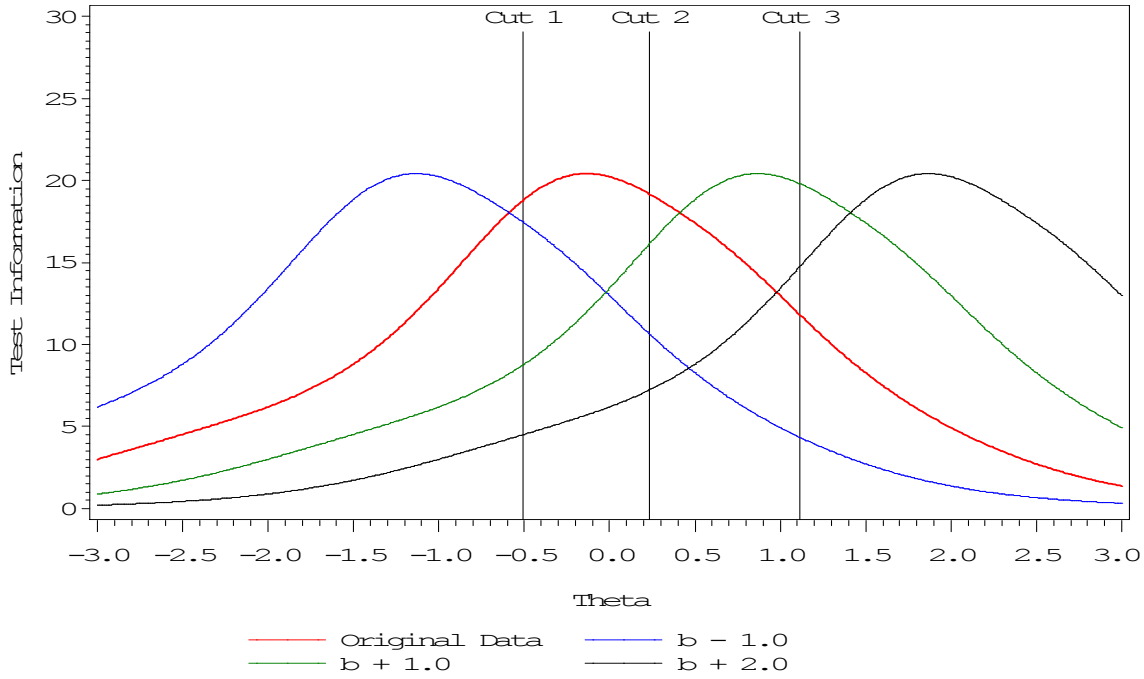


Figure 4.48 Middle school Mathematics test – manipulating difficulty level: Test information for the overall test and three variations of test difficulties (39 items, maximum score = 54)

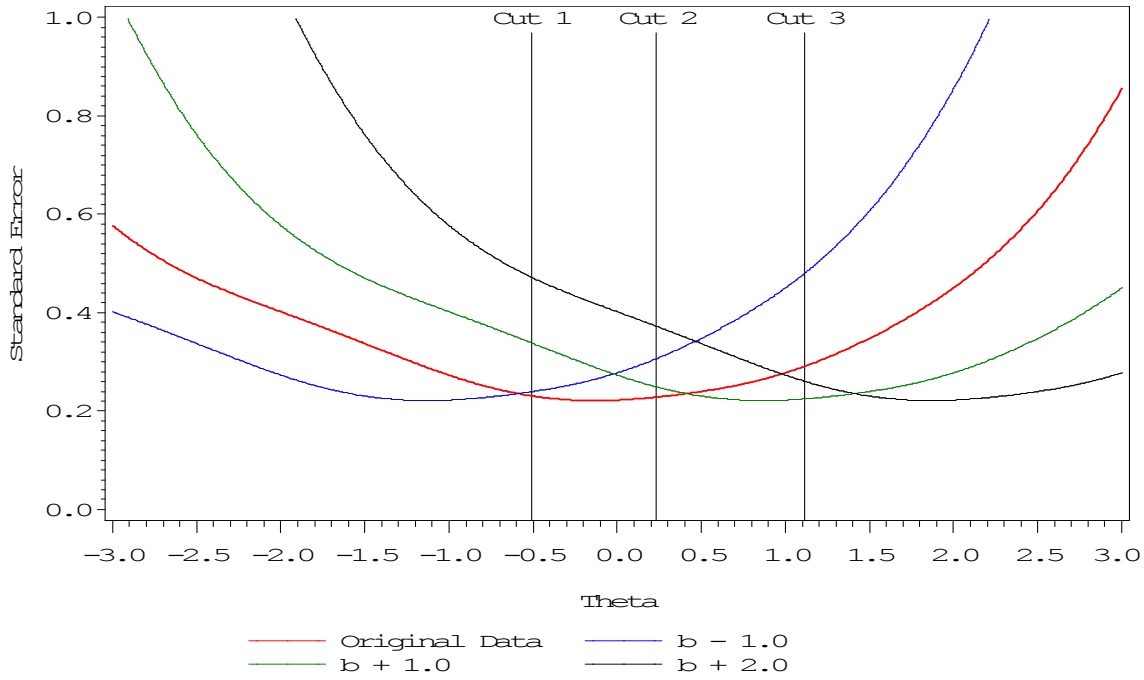


Figure 4.49 Middle school Mathematics test – manipulating difficulty level: Conditional standard error of measurement for the overall test and three variations of test difficulties

Test information for the original Mathematics test was approximately 19.0, 19.5, and 12.0 at Cut 1, Cut 2 and Cut 3, and their respective conditional standard errors of measurement was about .23, .23, and .29. Decreasing the difficulty of the test by 1.0 unit made the information for the new test lower at all cutscores, especially at the third cutscore. Information was only about 5.0 at Cut 3 and the corresponding conditional standard error of measurement was .45. The results of the easier test being less efficient compared to the original test can be observed in Figure 4.50, where the relative efficiency between the easier test (i.e., $b - 1.0$) and the original test was below the reference line (i.e., relative efficiency = 1.0). Although increasing the difficulty of the test pushed the information at the third cutscore to be higher compared to the original test, information was over 50% lower at the first cut and about 15% lower at the second cut when test difficulty was increased by 1.0 unit. Further increase in test difficulties (i.e., $b + 2.0$) actually made the test least effective at the first two cutscores and the third cutscore to be about 20% more effective than the original test.

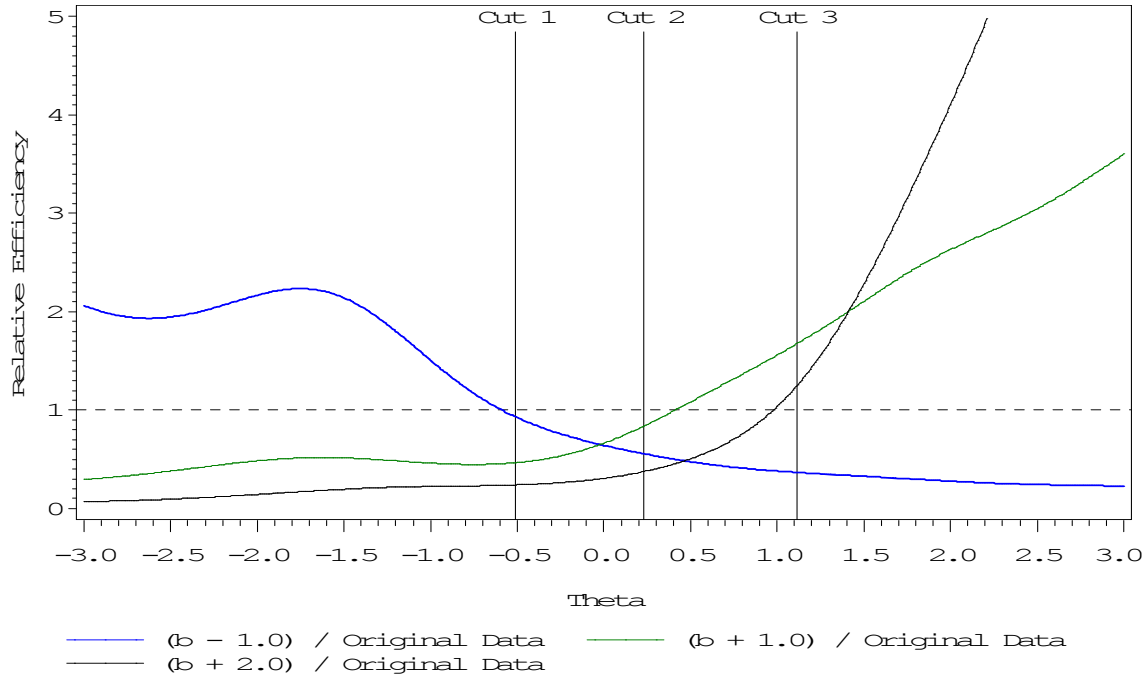


Figure 4.50 Middle school Mathematics test – manipulating difficulty level: Relative efficiency for the overall test and three variations of test difficulties

4.3.1.5 Summary

As concluded in the previous section, MC items in the middle school Mathematics test provided modest amount of information at the first two cutscores, and the information function was quite well centered. Making the items easier shifted the information function to the left of the proficiency continuum, thus, making the test more informative for those with lower proficiency scores. When MC items were more difficult, more information could be obtained for those at the higher end of the proficiency scale. For SA items, again, making the items easier or harder only shifted the information function to the left or right of the proficiency continuum scale. Since the original SA items appeared to be slightly harder than needed (as Cut 2 categorized examinees into either passing or failing category), making these items easier would be more appropriate for this test. The information function based on the original parameter estimates for the CR items peaked at Cut 1, by making these items a little bit harder (i.e., b + 1.0) shifted

the information function to the right and thus providing more information at Cut 2. At the overall test level, the original test information was rather well centered. Replacing some of the easier items with some moderately difficult items is recommended as the maximum information for the original information function was slightly off from the second cutscore.

4.3.2 High School English Language Arts (ELA) Test

Based on the summary of the item parameter estimates for the high school ELA test as presented in Table 3.2, constructed response (CR) items are the most difficult item format in the test. Difficulty of the multiple choice (MC) items and the essay items (EI) are comparable; however, the spread of the MC items was much wider than the EI items.

4.3.2.1 Effects of Changing Difficulty Level on the Multiple Choice Items

Figure 4.51 displays four information functions for the MC items in the high school ELA test, which is the information function based on the original item parameter estimates and three manipulations of the difficulty level: b - 1.0; b + 1.0; and b + 2.0.

The level of information at the three cutscores based on the original item parameter estimates of the MC items only was approximately 14.5, 17.0 and 7.0, which corresponded to the standard errors of measurement of .26, .24, and .38 at Cut 1, Cut 2 and Cut 3, respectively. The information function peaked at a proficiency score about 0. Decreasing the average difficulty of the MC items by 1.0 shifted the information function to the left of the proficiency continuum by 1 unit, thus the new information function peaked at around -1.0. The information at the three cutscores became 15.5, 7.5 and 1.5 which corresponded to standard errors of measurement of .25, .37, and .82, respectively. When the average difficulty of the MC items was increased by 1.0 unit, information at the three cutscores became 2.5, 11.5, and 17.0, and their corresponding standard errors of

measurement were .63, .29, and .24. The information function was shifted to the right of the proficiency continuum and it peaked at a proficiency score at 1.0. Extreme increase in the average difficulty of the MC items (i.e., *b* + 2.0) further shifted the information function to the right, and the information function now peaked at about a proficiency score of 2.0. Information at the three cutscores was .00, 1.5 and 12.0.

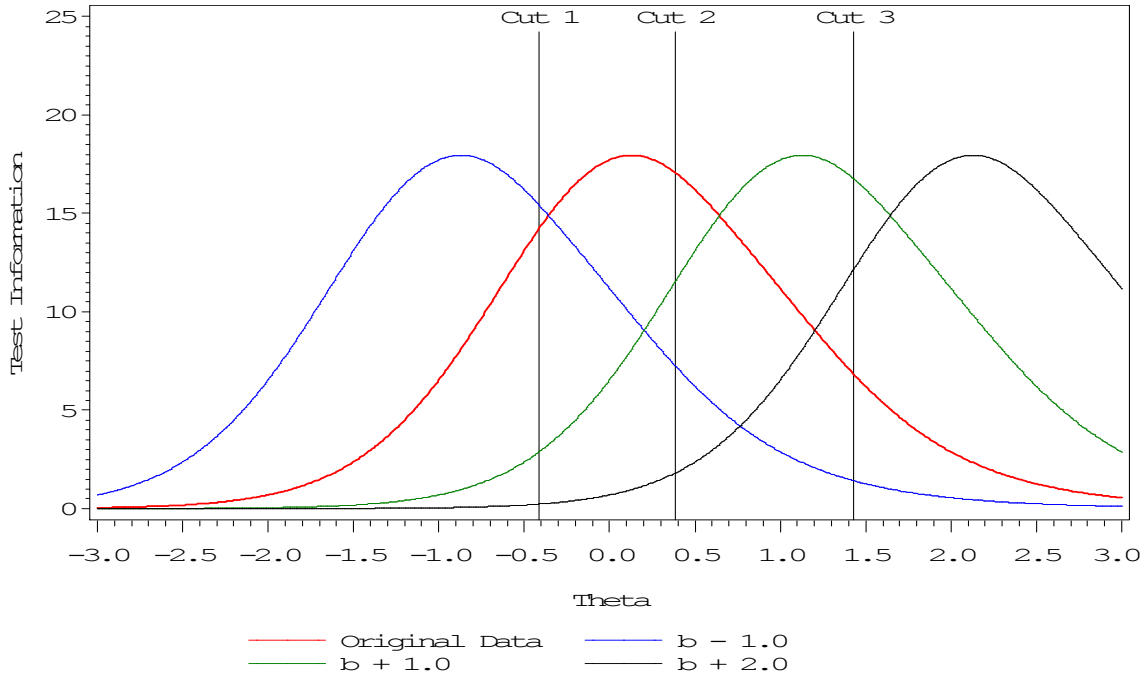


Figure 4.51 High school ELA test – manipulating difficulty level: Test information based on multiple choice items only (36 items, maximum score = 36)

Figure 4.52 displays the conditional standard errors for the MC items. Based on the results of the conditional standard error of measurement, decreasing the average difficulty of the MC items by 1.0 unit lowered the measurement error of the proficiency scores below the first cutscore, which is -.414; measurement errors were higher when the proficiency scores were above Cut 1. When average difficulty was increased by 1.0 unit, measurement errors were higher when proficiency scores were below .70, which is above the second cutscore. Extreme increase in the average difficulty of the MC items (i.e., *b* + 2.0) only made the measurement errors lowered for proficiency scores above 1.6.

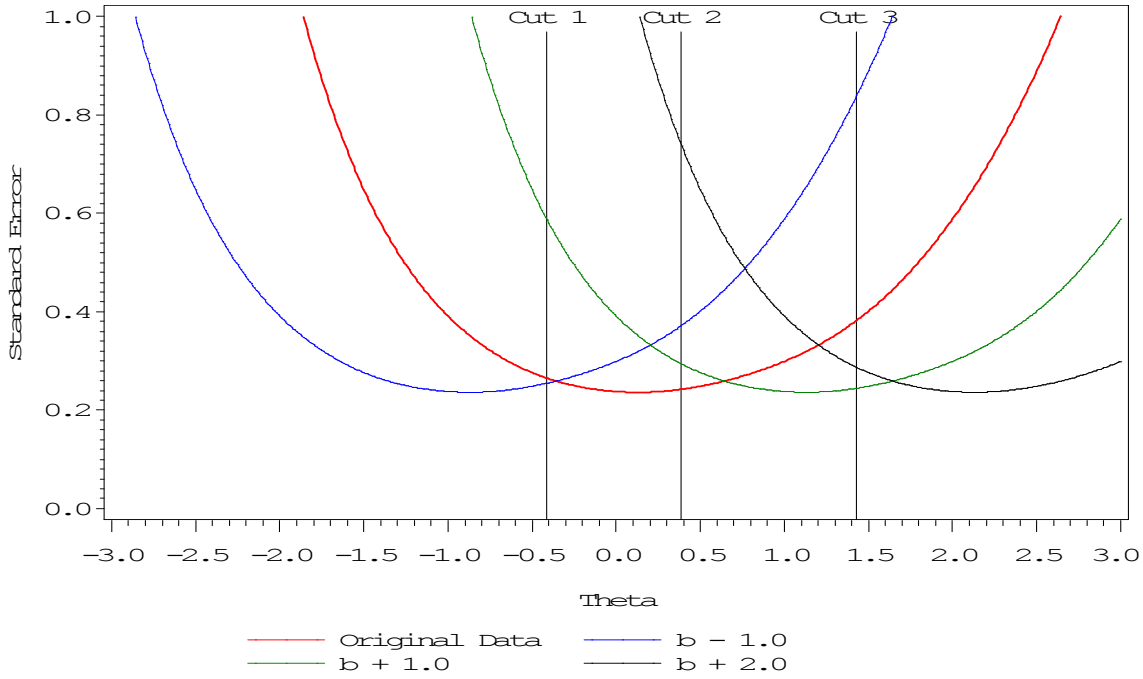


Figure 4.52 High school ELA test – manipulating difficulty level: Conditional standard error of measurement based on multiple choice items only

Figure 4.53 displays the relative efficiency of each of the three variations of tests versus the original test. When considering only MC items in the high school ELA test, decreasing the *b*-parameter estimates by 1.0 made the new test less efficient than the original test at the second and the third cutscore; however, at Cut 1, this easier test was only about 10% more efficient than the original MC items. An increase of 1.0 unit in the overall *b* for the MC items made the new test less efficient compared to the original test at Cut 1 and Cut 2; however, at Cut 3, this new test was about 2.5 times more efficient than the original MC items. Increasing the average difficulty of the MC items by 2.0 units made the new test extremely inefficient at Cut 1 and Cut 2. At Cut 3, the new test was about 80% more efficient compared to the original MC items.

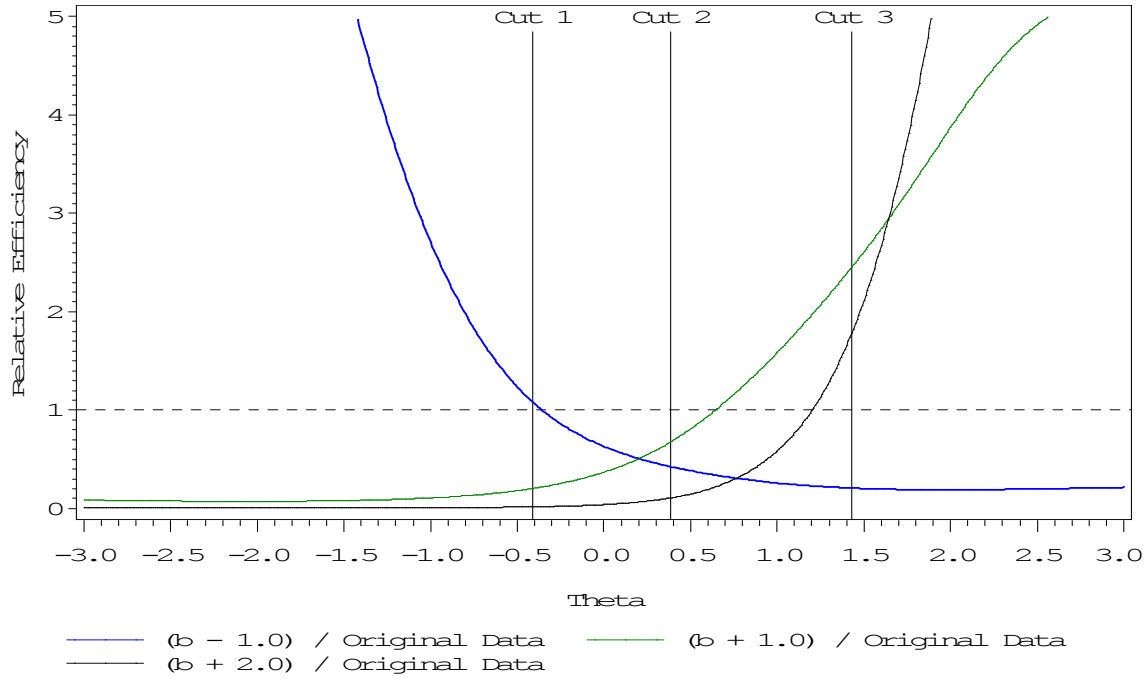


Figure 4.53 High school ELA test – manipulating difficulty level: Relative efficiency based on multiple choice items only

4.3.2.2 Effects of Changing Difficulty Level on the Constructed Response Items

Figures 4.54 and 4.55 present the information functions and the conditional standard errors for the CR items in the high school ELA test. Each of the figures contains the information or conditional standard errors based on the original item parameter estimates and three manipulations of average difficulty for the CR items: $b - 1.0$; $b + 1.0$; and $b + 2.0$.

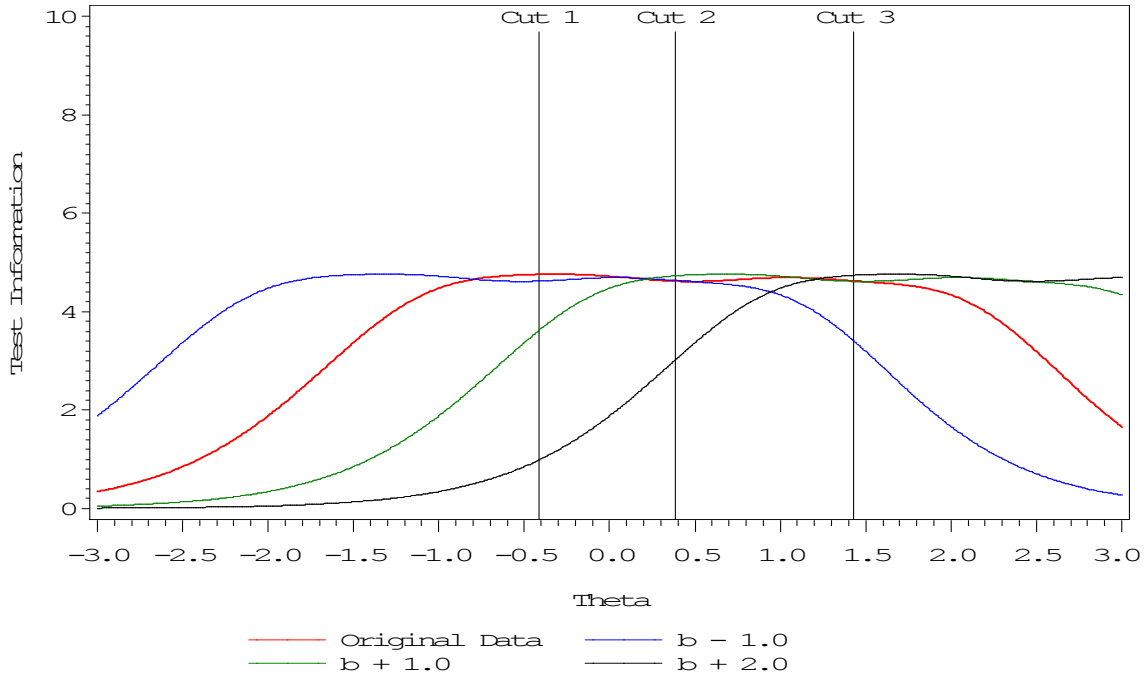


Figure 4.54 High school ELA test – manipulating difficulty level: Test information based on constructed response items only (4 items, maximum score = 16)

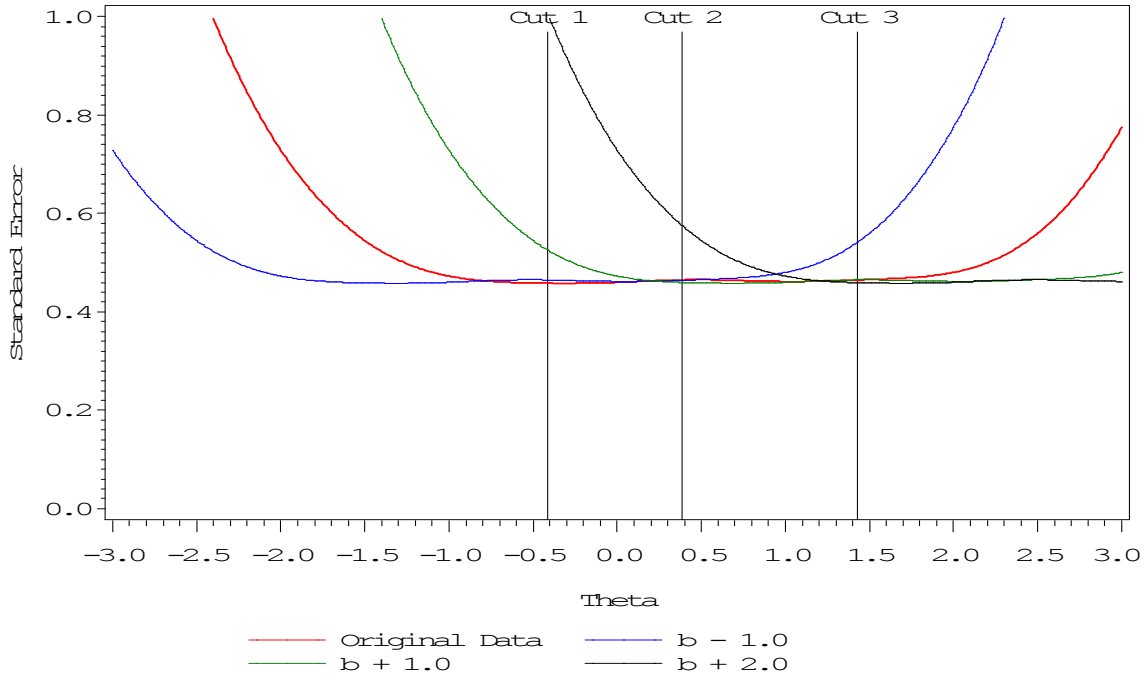


Figure 4.55 High school ELA test – manipulating difficulty level: Conditional standard error of measurement based on constructed response items only

The amount of information provided by the first two cutscores were quite similar between the original parameter estimates and when the difficulty of the CR items were decreased by 1.0 unit. The amount of information at Cut 1 was considerably less when items were more difficult (i.e., when $b + 1.0$) or when items were substantially more difficult (i.e., when $b + 2.0$). At Cut 2, the amount of information provided by the original parameter estimates, the easier test and more difficult test (i.e., $b + 1.0$) was about the same, but the amount of information provided by the most difficult test at this cutscore was substantially less. The amount of information provided by the original test and the two more difficult tests at Cut 3 were about the same, but the amount of information provided by the easier test was considerably lower.

The conditional standard errors of measurement were comparable across different versions of tests, except at Cut 1 where the errors were higher when average *b* for CR items was increased by 1.0 unit and it was much higher when the average *b* for the CR items was increased by 2.0 units. At Cut 2, measurement errors were highest when *b* was increased by 2.0 units. At Cut 3, measurement errors were highest for the easiest test.

Figure 4.56 reports the relative efficiency of the CR items for the original test compared to the three tests. In general, comparing to the original test, the easier version of the test was as efficient as the original test at Cut 1 and Cut 2, but at Cut 3, the new test only functioned at about 70% of the original test. When test difficulty was increased by 1.0 unit, Cut 2 and Cut 3 functioned about the same as the original test, but this test only functioned at about 80% of the original test at Cut 1. Finally, when test difficulty was increased by 2.0 units, it only performed as well as the original test at Cut 3, but the test was less efficient at Cut 1 and Cut 2 compared to the original test.

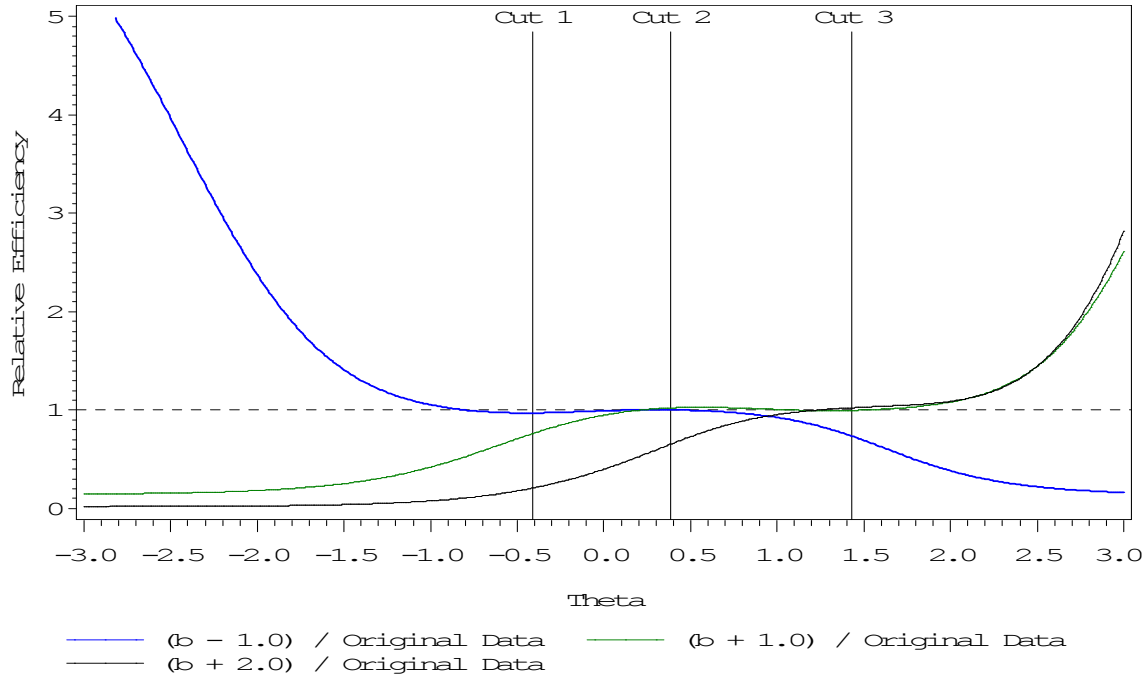


Figure 4.56 High school ELA test – manipulating difficulty level: Relative efficiency based on constructed response items only

4.3.2.3 Effects of Changing Difficulty Level on the Essay Items

Figures 4.57 and 4.58 present the information functions and the conditional standard errors for the essay items (EI) in high school ELA test based on the original parameter estimates and when $b - 1.0$, $b + 1.0$, and $b + 2.0$.

Information functions for EI were at least bimodal, regardless of the difficulty level. The information function based on the original parameter estimates peaked at around proficiency scores -1.2 and .40. When the average difficulty for the EI was decreased by 1.0 unit, the information function was shifted to the left and was peaked at -2.2 and -.60. When the average difficulty was increased by 1.0 unit, the information function was then peaked at -.20 and 1.4. Further increase in difficulty shifted the information function further to the right, making it peaked at .80 and 2.4.

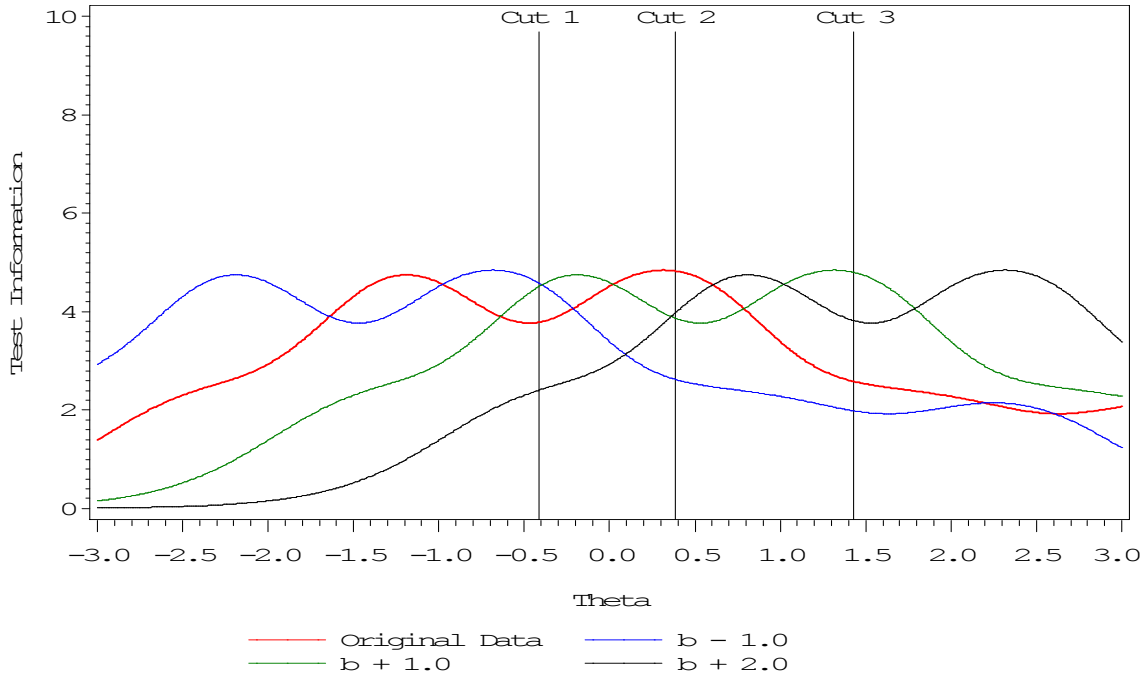


Figure 4.57 High school ELA test – manipulating difficulty level: Test information based on essay items only (2 items, maximum score = 16)

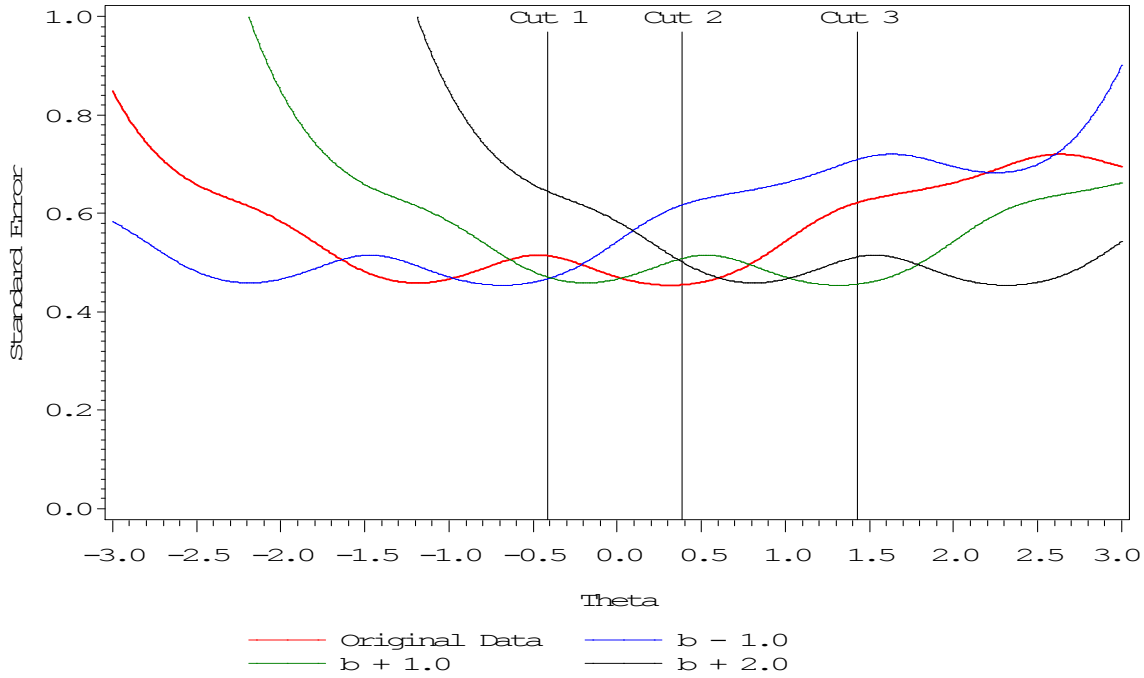


Figure 4.58 High school ELA test – manipulating difficulty level: Conditional standard error of measurement based on essay items only

Figure 4.59 reports the relative efficiency of the EI for the original test compared to the three tests. Except for the extreme increase in difficulty, the other two tests were more efficient than the original test at Cut 1. At Cut 2, none of the modified tests were as efficient as the original test. At Cut 3, except for the easier test, increasing the difficulty by 1.0 or 2.0 units both made the tests more efficient than the original test, but the efficiency from increasing the difficulty by 1.0 unit was higher between the two.

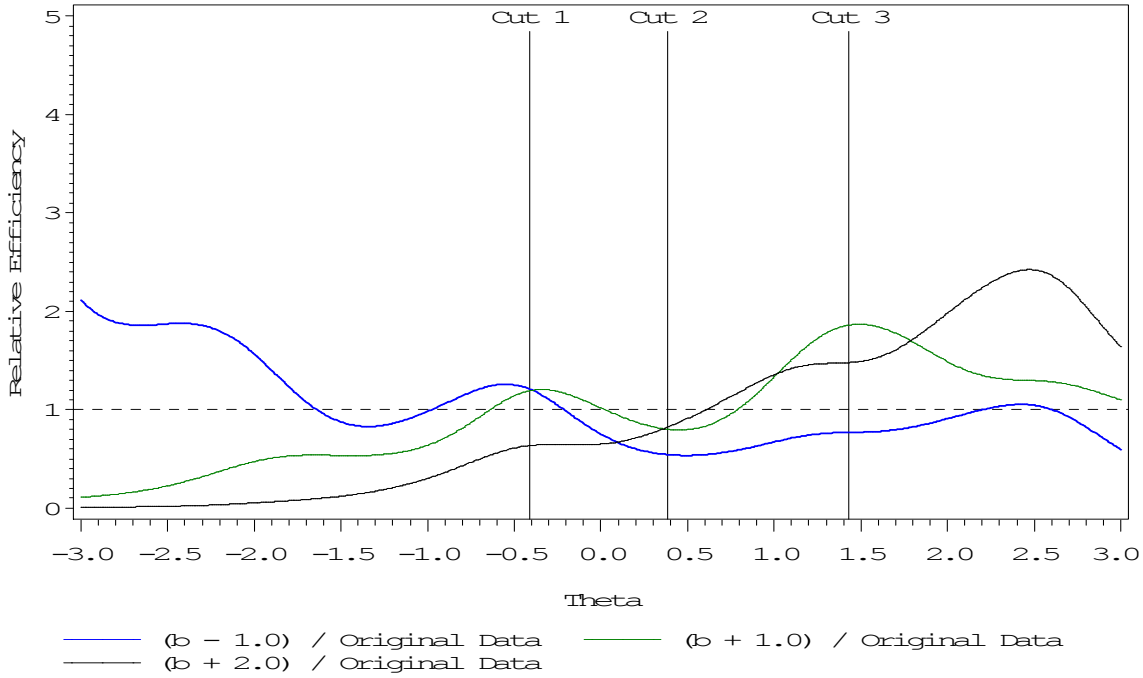


Figure 4.59 High school ELA test – manipulating difficulty level: Relative efficiency based on essay items only

4.3.2.4 Effects of Changing Difficulty Level on the Overall Test

Figures 4.60 and 4.61 compare the information functions and the standard errors for the overall test based on the original parameter estimates and when the item difficulty values were decreased by 1.0, increased by 1.0 and 2.0.

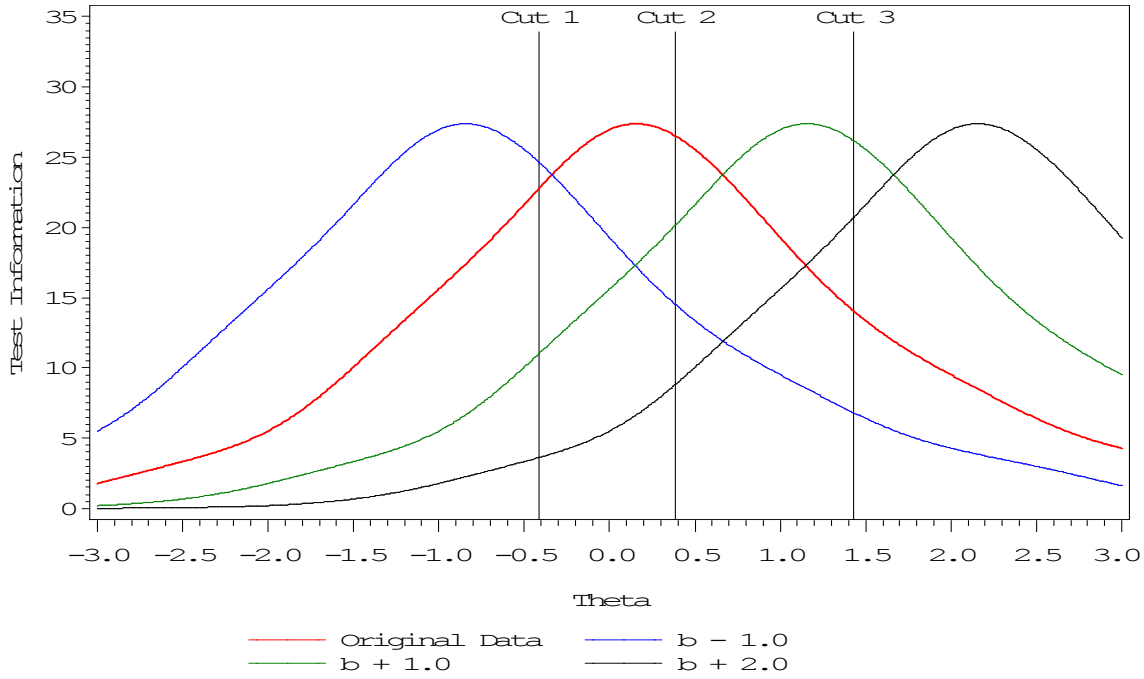


Figure 4.60 High school ELA test – manipulating difficulty level: Test information for the overall test and three variations of test difficulties (42 items, maximum score = 68)

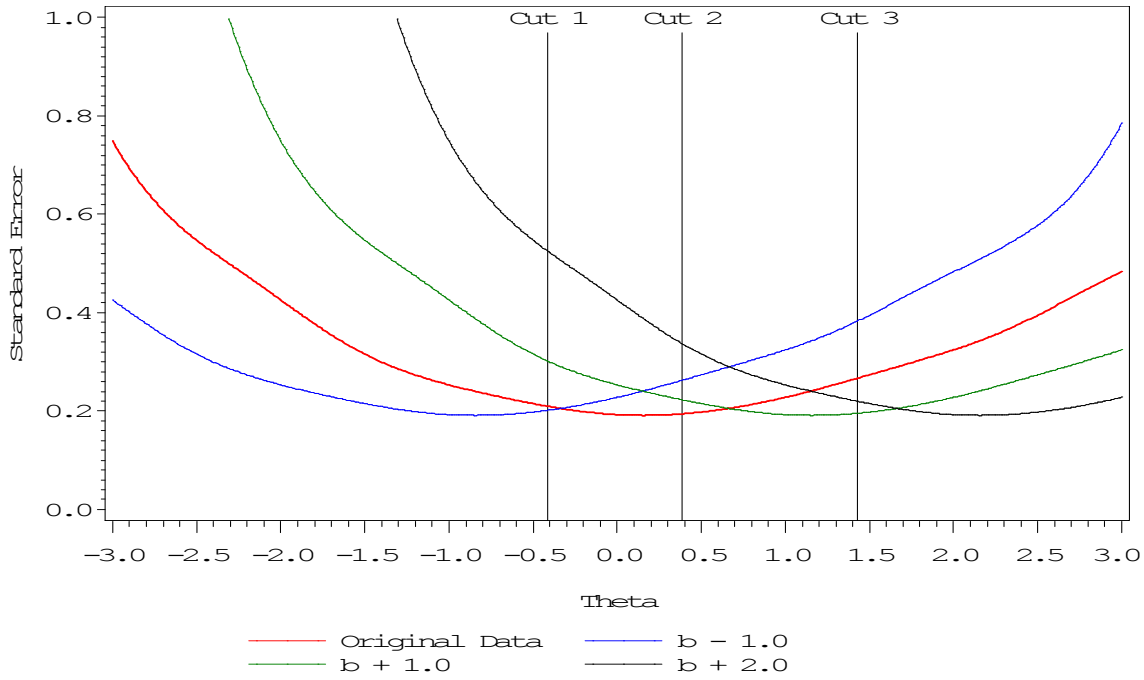


Figure 4.61 High school ELA test – manipulating difficulty level: Conditional standard error of measurements for the overall test and three variations of test difficulties

Test information for the original ELA test was approximately 22.5, 26.0, and 14.0 at Cut 1, Cut 2 and Cut 3, respectively, and their corresponding conditional standard errors of measurement were about .21, .20, and .27. Decreasing the difficulty of the test by 1.0 unit made the information at Cut 1 slightly higher but lower at the other two cutscores, especially at the third cut (information was only about 7.0 at Cut 3 and the corresponding conditional standard error of measurement was .38).

Figure 4.62 displays the relative efficiency of each of the three tests versus the original test. The easier test was less efficient at Cut 2 and Cut 3 compared to the original test. Relative efficiency between the easier test (i.e., $b - 1.0$) and the original test was below the reference line (i.e.: relative efficiency = 1.0). Although increasing the difficulty of the test pushed the information at the third cut to be higher compared to the original test, information was over 50% lower at the first cut and about 25% lower at the second cut when test difficulty was increased by 1.0 unit. Further increase in item difficulty (i.e., $b + 2.0$) actually made the test least effective at the first two cutscores and the third cutscore to be about 50% more effective than the original test. However, it was clear that increasing the overall difficulty by 1.0 unit almost doubled the information at the third cutscore when comparing to the original test.

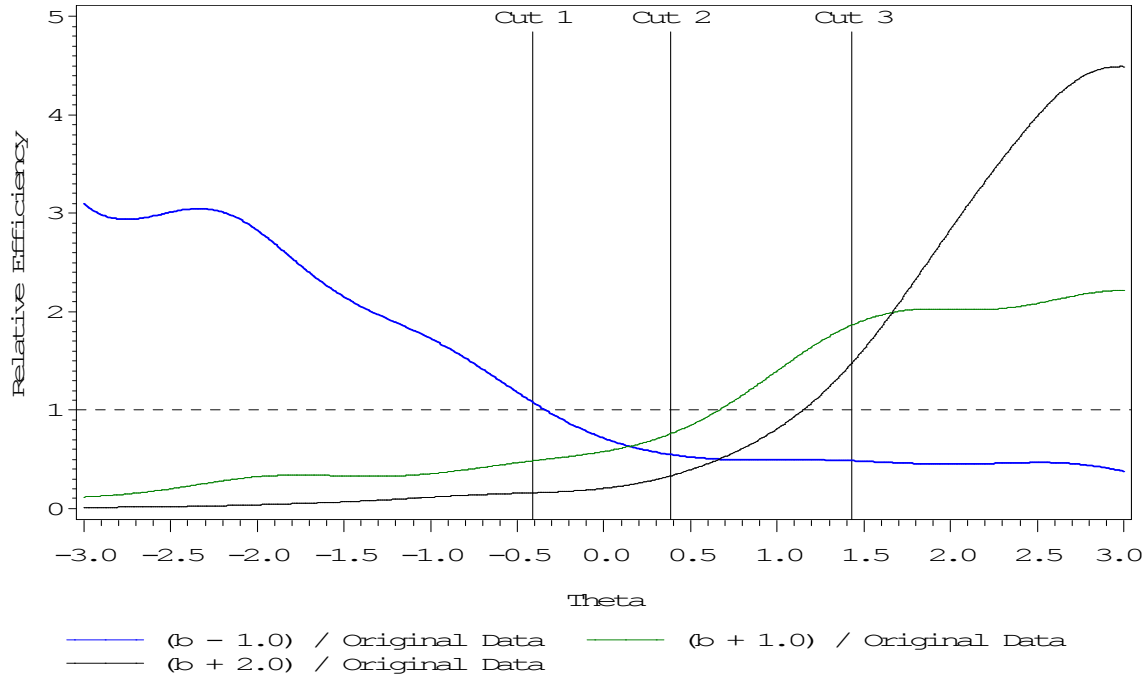


Figure 4.62 High school ELA test – manipulating difficulty level: Relative efficiency for the overall test and three variations of test difficulties

4.3.2.5 Summary

In conclusion, MC items in the high school ELA test provided modest amount of information at the first two cutscores, and the information function was quite well centered. However, the amount of information at Cut 3 was slightly insufficient (using information = 10 as the criterion). The information function for the CR items was quite flat, meaning that these items provided relatively the same amount of information across the proficiency continuum. Making the items easier or harder only shifted the information function to the left or right of the proficiency scale. Information functions for EI were at least bimodal. The information function based on the original parameter estimates peaked at proficiency scores about -1.2 and .40. Increasing the difficulty of these items is recommended so that information would peak at the cutscores, especially at Cut 2, where this is the cutscore that distinguishes students into passing and failing categories. At the overall test level, the original test information was rather well centered; however, more

difficult items to be included in the test would be recommended as the amount of information at the lowest cut is excessive. At the same time, by including more difficult items in the test would shift the information to the right of the proficiency continuum, making the measurement error lowered at the pass and fail cut (i.e., Cutscore 2).

4.4 Building the Optimal Test

Individual item information functions based on the original parameter estimates and variations of item discriminations and item difficulties for the middle school Mathematics test and the high school ELA test are presented in Appendix A and B, respectively.

Each item was evaluated based on the original location of maximum information and how the item information function shifted when the *a*- or *b*-parameter changed. The optimal test for each content area was built by choosing the appropriate parameter estimates for each item that could maximize the information at the three cutscores. This is to mimic the process that test developers pick the most appropriate items, statistically, in different regions of the proficiency scale in order to maximize information at a particular region of interest, for example, near the pass/fail cutscore; however, the process should not affect the content validity of the test. In this study, variations were limited to either increasing the discriminating power or changing the difficulty of the item, but in reality, test developers could replace items with both high discriminating value and different levels of difficulty to fit their purposes.

4.4.1 Middle School Mathematics Test

Table 4.5 below summarizes the number of items by item type that were being "modified" for the creation of the optimal test. Table 4.6 presents the summary of the item parameter estimates by item type for the optimal test.

Table 4.5 Number of Modified Item Parameter Estimates for Middle School Mathematics Test by Item Type.

Changes in Item Parameter Estimates		Item Type[1]		
		MC	SA	CR
a	+ .05	0	0	0
	+ .10	0	0	0
	+ .30	4	3	2
b	- 1.0	0	0	0
	+ 1.0	5	1	0
	+ 2.0	1	1	1
No changes made		19	0	2

[1] MC – Multiple choice items, SA – Short answer items, CR – Constructed response items

Table 4.6 Summary of Item Parameter Estimates for Optimal Middle School Mathematics Test by Item Type.

Item Type[1]	*n*	Parameter	Mean	SD	Min	Max
MC	29	*a*	1.11	.28	.67	1.82
		b	-.05	.43	-.84	.71
		c	.18	.08	.05	.36
SA	5	*a*	.96	.33	.55	1.38
		b	.76	.33	.37	1.14
CR	5	*a*	1.16	.19	.92	1.13
		b	-.20	.70	-.85	.80

[1] MC – Multiple choice items, SA – Short answer items, CR – Constructed response items

The discriminating parameters of 4 MC items were increased by .30, which increased the averaged discriminating power of the optimal test by .04 from the original test. The average difficulty of the MC items were also increased by .25 from the original test, after increasing the difficulty of 5 MC items by 1.0 unit and another MC item by 2.0 units. Three out of five SA items were modified to have higher discriminating power (i.e., *a* + .30), which increased the average discriminating power to .96 for the SA items in the optimal test, and difficulties of the other 2 items were increased. For the CR items, 2 items were modified to have higher discriminating power and only 1 item was modified

to become much more difficult, which made the average difficulty of this item type increased by .40 and the average discriminating power increased by .12.

Figures 4.63 and 4.64 present the information functions and the standard errors of the optimal test and the original test. Test information for the original Mathematics test was approximately 19.0, 19.5, and 12.0 at Cut 1, Cut 2 and Cut 3. After modifications to the item parameters for the creation of the optimal test, the test information became 19.0, 23.0 and 16.5. This optimal test is now centered at Cut 2, which is the most significant cutscore for the examinees because it is used to categorize examinees into either passing or failing category.

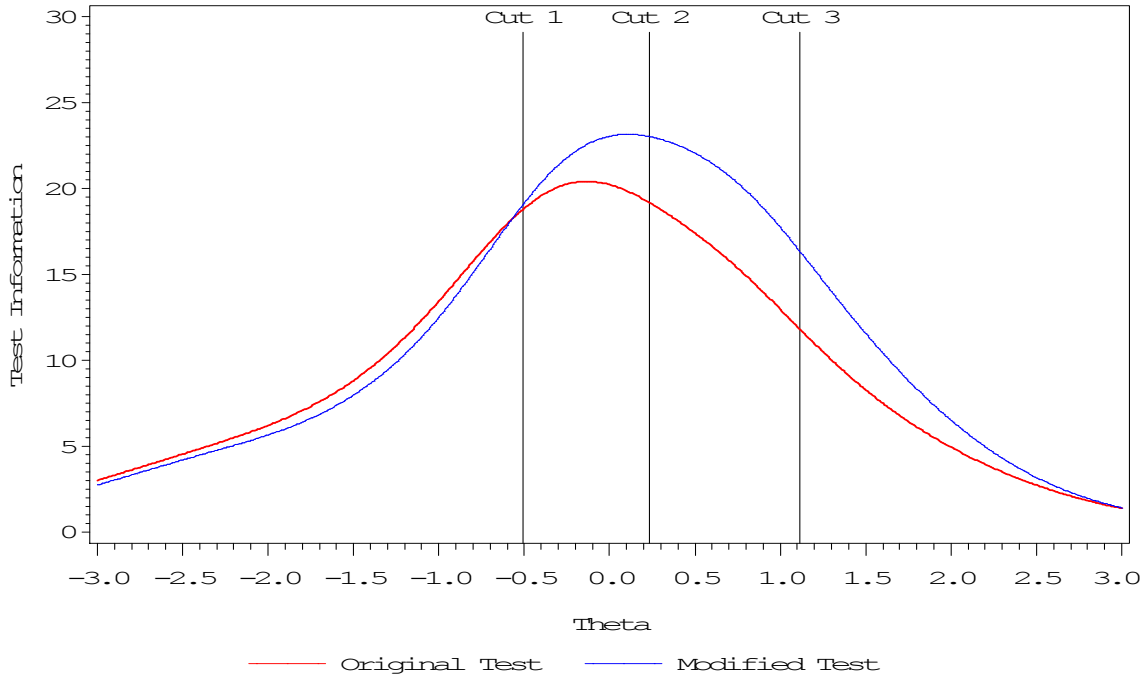


Figure 4.63 Middle school Mathematics test: Test information for the original test and the optimal test (39 items, maximum score = 54)

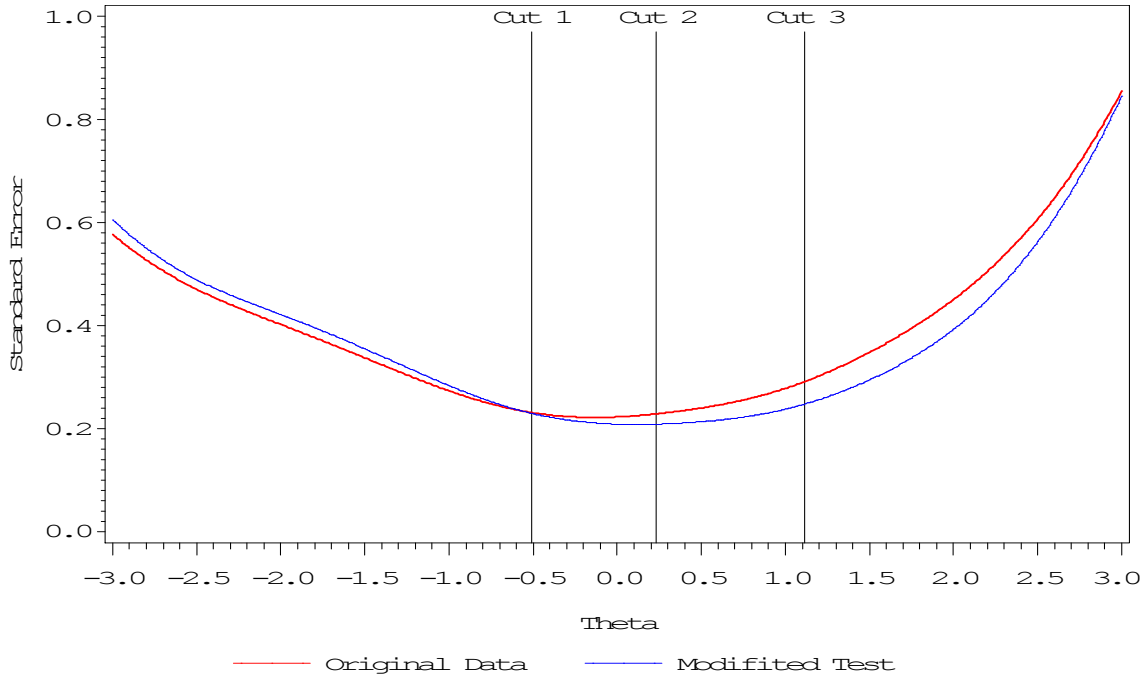


Figure 4.64 Middle school Mathematics test: Conditional standard error of measurement for the original test and the optimal test

Figure 4.65 displays the relative efficiency of the optimal test versus the original test. The optimal test effectively lowered the standard errors of measurement at Cut 2 and Cut 3, compared to the original test. This test was about 20% more efficient than the original test at Cut 2 and 35% more efficient at Cut 3. At Cut 1, the optimal test was as efficient as the original test, but slightly less efficient for proficiency scores below Cut 1.

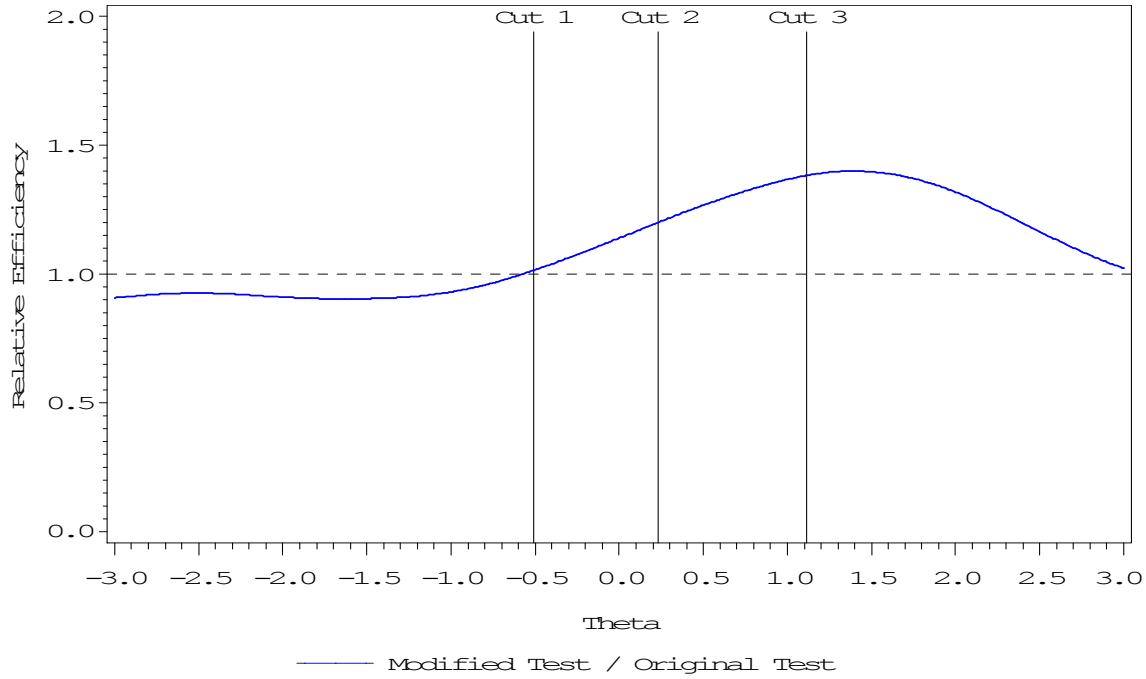


Figure 4.65 Middle school Mathematics test: Relative efficiency of the original test versus the optimal test

4.4.2 High School English Language Art (ELA) Test

The test information function based on original parameter estimates met reasonable expectations at all three cutscores; however, the level of information at Cut 3 was lower, compared to the other two cuts. Test improvement could be made by replacing a few items from the low and middle difficulty categories to the higher levels of difficulty or improving the discriminating powers of items.

Table 4.7 below summarizes the number of items by item type that were being "modified" for the creation of the optimal high school ELA test. Table 4.8 presents the summary of the item parameter estimates by item type for the optimal test.

Table 4.7 Number of Modified Item Parameter Estimates for High School ELA Test by Item Type.

Changes in Item Parameter Estimates		Item Type[1]		
		MC	CR	EI
a	+ .05	0	0	0
	+ .10	0	0	0
	+ .30	0	4	0
b	- 1.0	0	0	0
	+ 1.0	4	0	2
	+ 2.0	0	0	0
No changes made		32	0	0

[1] MC – Multiple choice items, CR – Constructed response items, EI – Essay items

Table 4.8 Summary of Item Parameter Estimates for Optimal High School ELA Test by Item Type.

Item Type[1]	*n*	Parameter	Mean	SD	Min	Max
MC	36	*a*	1.12	.29	.59	1.96
		b	.07	.48	-.79	.97
		c	.22	.06	.11	.38
CR	4	*a*	1.48	.11	1.39	1.64
		b	.42	.17	.27	.57
EI	2	*a*	1.67	.14	1.57	1.77
		b	.98	.51	.62	1.34

[1] MC – Multiple choice items, CR – Constructed response items, EI – Essay items

Item difficulties for four MC items were increased by 1.0, making the average difficulty of the MC portion of the optimal test .11 higher than the original test. Item discriminating powers for all the CR items were increased by .30. Difficulties for all essay items were increased by 1.0.

Figures 4.66 and 4.67 present the information functions and the standard errors of the optimal and the original ELA test.

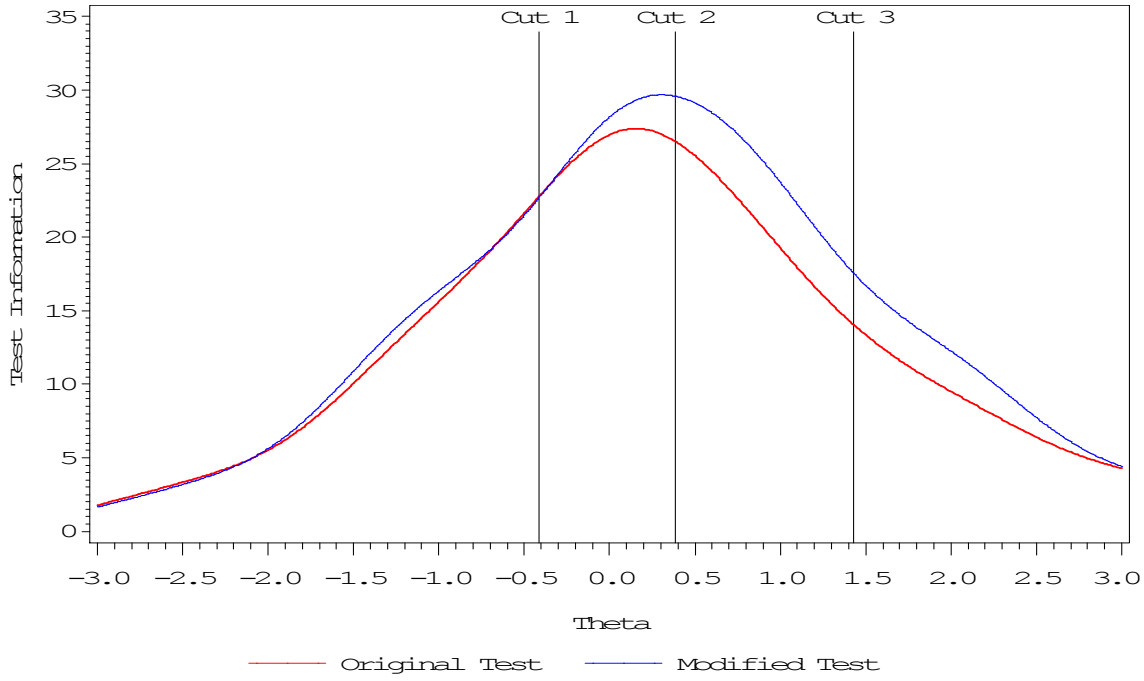


Figure 4.66 High school ELA test: Test information for the original test and the optimal test (42 items, maximum score = 68)

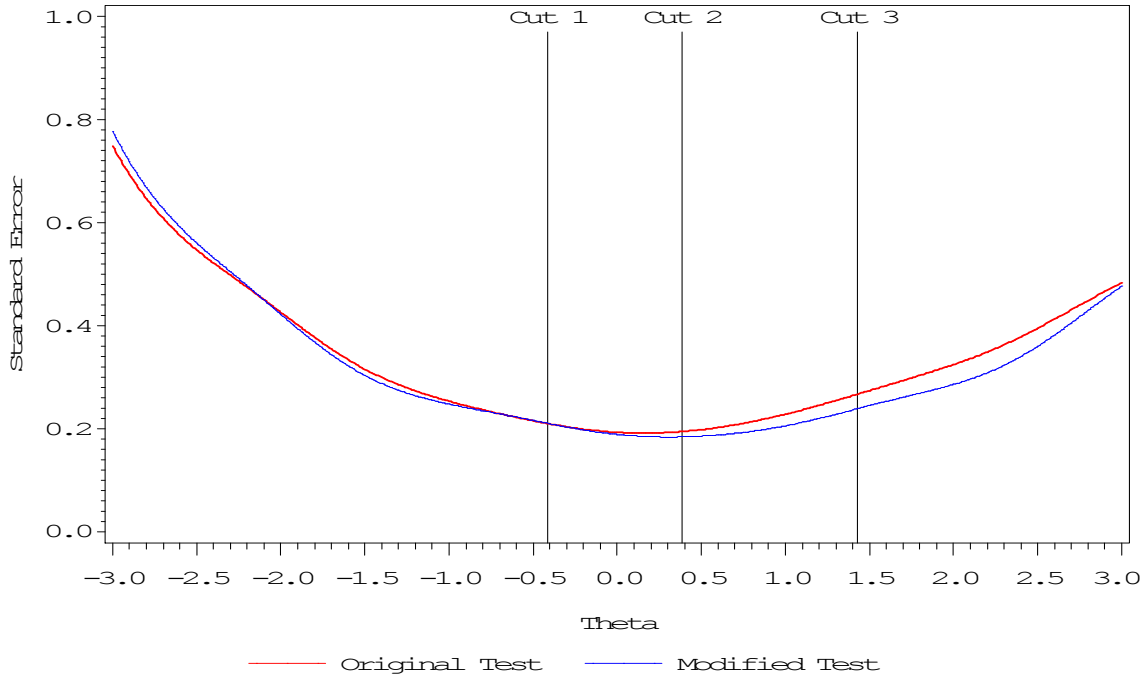


Figure 4.67 High school ELA test: Conditional standard error of measurement for the original test and the optimal test

Test information for the original ELA test was approximately 22.5, 26.0, and 14.0 at Cut 1, Cut 2 and Cut 3, respectively. After modifications to the item parameters for the creation of the optimal test, test information became 22.5, 29.5 and 17.5. The location of maximum information was now centered at Cut 2. In addition, the optimal test slightly lowered the standard errors of measurement at Cut 2 and Cut 3 compared to the original test. Conditional standard errors of measurement at Cut 1 for the original and the optimal test were almost identical; however, the errors for proficiency scores from the original test were slightly lower at the lower end of the proficiency continuum.

Figure 4.68 displays the relative efficiency of the optimal test versus the original test. The optimal test was about 8% more efficient than the original test at Cut 2 and about 25% more efficient at Cut 3. At Cut 1, the optimal test was as efficient as the original test.

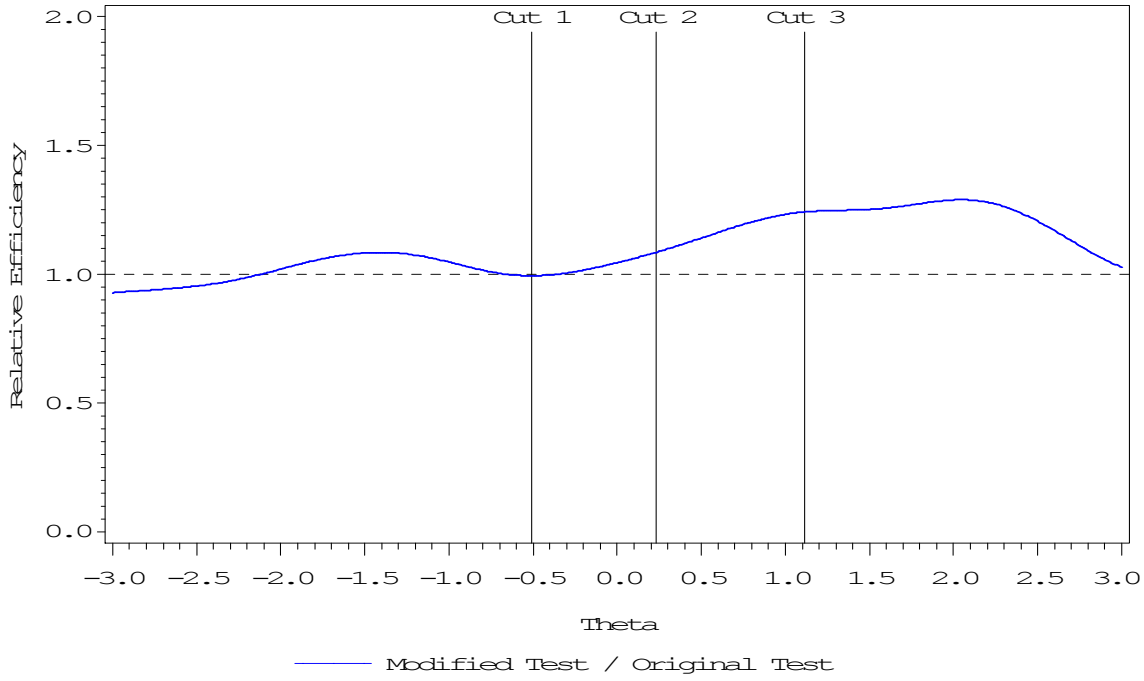


Figure 4.68 Middle school Mathematics test: Relative efficiency of the original test versus the optimal test

4.5 Decision Consistency, Decision Accuracy, and Expected Information

When proficiency classifications and other performance standards are set on educational tests, it is important to estimate the degree to which the classifications are accurate and reliable. As the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) state, "when a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure …" (p. 35).

For many educational tests, such as those used for high school graduation or for the accountability demands associated with the No Child Left Behind (NCLB) Act, the most important interpretations of students' test performance are based on proficiency classifications. As is evident from the aforementioned standard, providing evidence of decision consistency is one way to demonstrate such classifications can be reliably made on the basis of test scores. It is also important to provide evidence that classification decisions made on the basis of test scores are not only reliable, but also accurate. For these reasons, the concepts of decision consistency (DC) and decision accuracy (DA) are important for understanding the technical quality and utility of tests that classify students into different achievement level categories.

Although the information function provides a clear demonstration of the information at each point on the ability scale; however, it does not provide a measure of congruence of the information function to the ability distribution of the population of interest. The expected information, proposed by Donoghue (1994), is a weighted average

based upon the distribution of examinee ability; therefore, it reflects how well an item is focused.

As seen in the previous sections, increasing the discriminating power of test items or changing the difficulty of tests definitely affect the information provided by the test at various locations on the proficiency scale, hence, it could also affect the consistency and accuracy of proficiency classifications, and also the expected information when information shifts along the proficiency continuum. This section presents the results of DC and DA for all the conditions studied. In addition, expected information will be provided for the original test; when the average discriminating power of the test was increased by .05, .10 and .30; when average test difficulty was decreased by 1.0, increased by 1.0 and 2.0; and finally the optimal test.

4.5.1 Middle School Mathematics Test

4.5.1.1 Decision Consistency

Table 4.9 below summarizes the results of decision consistency (DC) and *Kappa* statistics for the middle school Mathematics test for all conditions studied in the previous sections.

Table 4.9 Summary of Decision Consistency (DC) and *Kappa* Statistics for Middle School Mathematics Test.

		Increase average *a* by…			Increase low *a* by…		
	Original Test	.05	.10	.30	.05	.10	.30
DC	68.09%	68.07%	68.84%	72.24%	67.57%	67.65%	68.26%
Kappa	54.26%	54.27%	55.45%	60.46%	53.51%	53.55%	54.49%

	Increase low and medium *a* by…			Change in average *b* by…			
	.05	.10	.30	-1.00	+1.00	+2.00	Optimal Test
DC	67.54%	68.06%	69.41%	65.58%	62.62%	53.81%	68.81%
Kappa	53.46%	54.23%	56.28%	48.43%	47.65%	36.71%	55.68%

It can be observed from the above table that when the average item discriminating power of the test was increased, both the index of DC and the *Kappa* statistics were also increased. However, if only increasing the discriminating power of the low discriminating items or those items with low or medium discriminating value, higher increase in the discriminating power would be needed (for example, increasing *a* by more than .10 for the low discriminating items).

DC obtained from the easier test (i.e., $b - 1.0$) and slightly more difficult test (i.e., $b + 1.0$) were comparable to the original test; however, more decision agreement were due to chance as their *Kappa* values were lower. When the test was substantially more difficult than needed, DC dropped closer to the minimum accepted level (i.e., .50) (Crocker & Algina, 1986, p. 200).

The *Kappa* statistic takes into account the chance agreement in decision consistency, if the two forms are strictly parallel; *Kappa* has a maximum value of 1. Based on the results presented in Table 4.9, the *Kappa* statistics for all variations of tests were between .45 to .60, indicating a moderate level of agreement from two administrations of the same test. *Kappa* was highest when the overall discriminating value of the test was increased by .30 (60.46%); and lowest when the overall difficulty of test was increased by 2.0 units (36.71%), indicating only a fair level of agreement.

For the optimal test, where the item discriminating power for 4 MC, 3 SA and 2 CR items were increased by .30, item difficulties for 5 MC and 1 SA items were increased by 1.0, and item difficulty level for one item out of each item type were increased by 2.0, DC was increased by about 1.06% from the original test and the *Kappa* statistic was increased by about 2.62%.

4.5.1.2 Decision Accuracy, False Positive and False Negative Error Rate

Table 4.10 summarizes the results of decision accuracy (DA) at the overall test level and false positive (FP) and false negative (FN) error rate at each of the cutscores.

Table 4.10 Summary of Decision Accuracy (DA), False Positive (FP) and False Negative (FN) Error Rate for Middle School Mathematics Test.

	DA	False Positive			False Negative		
		Cut1	Cut2	Cut3	Cut1	Cut2	Cut3
Original Test	70.76%	17.47%	15.22%	13.24%	14.99%	16.85%	19.06%
Increase average *a* by…							
0.05	71.02%	16.61%	14.34%	12.53%	14.07%	16.04%	18.56%
0.10	71.87%	15.78%	13.51%	11.87%	13.13%	15.27%	18.04%
0.30	75.07%	12.90%	10.61%	9.72%	10.11%	12.92%	16.23%
Increase low *a* by…							
0.05	70.60%	17.21%	14.98%	13.01%	14.75%	16.58%	18.78%
0.10	69.88%	16.91%	14.73%	12.72%	14.50%	16.32%	18.48%
0.30	70.92%	15.89%	13.73%	11.77%	13.51%	15.27%	17.50%
Increase low and medium *a* by…							
0.05	71.02%	17.11%	14.90%	12.90%	14.65%	16.53%	18.70%
0.10	70.48%	16.71%	14.58%	12.63%	14.31%	16.15%	18.41%
0.30	71.69%	15.25%	13.30%	11.49%	12.95%	14.97%	17.19%
Change in average *b* by…							
-1.00	59.80%	15.68%	20.57%	23.87%	18.08%	25.16%	30.54%
+1.00	69.09%	26.58%	20.06%	9.45%	22.68%	16.25%	10.81%
+2.00	60.26%	32.67%	28.46%	16.23%	29.91%	25.22%	10.91%
Optimal Test	72.90%	17.85%	13.63%	9.82%	14.21%	14.02%	15.05%

Similar to the DC results, when the average discriminating power of the test increased, the rate of accurate decision classification also increased. In addition, higher increase in the discriminating power was needed for the improvement in DA if only increasing the discriminating power of the low or low and medium discriminating items. When difficulty of the test did not align with the ability of the target population, DA drops drastically, especially for the case where the test was easier (i.e., $b - 1$) and when the test was much more difficult than was needed (i.e., $b + 2$). DA for the optimal test

was about 3% higher than the original test, the improvement might seem small; however, for a typical large scale statewide assessment which can range from about 50,000 to 200,000 examinees in a grade level, 3% improvement in DA could affect about 1,000 to over 4,000 examinees!

Both false positive (FP) and false negative (FN) rates decreased when average discriminating power increased when comparing the original test to three levels of increase in the overall discriminating power of the test.

When the increase in discriminating power only applied to the low discriminating items, FP rate decreased at a lower rate: if the low discriminating items were replaced by those with discriminating power of .05 higher, FP rate for the three cutscores were about 1 to 2% lower than the original test; when discriminating power of the low *a* items were increased by .10, FP rate decreased by about 3 to 4% at the cutscores; and when the average *a* for the low discriminating items were increased by .30, FP rate decreased by 9 to 11% at the cutscores.

When the increase in discriminating power applied to the low and medium discriminating items, the rate of decrease in FP was slightly higher than those obtained from only increasing *a* of the low discriminating group.

When the test was easier (i.e., $b - 1$), FP rate at Cut 1 was lower than the original test; however, FP rate at Cut 2 and Cut 3 was higher than the original test. When the test was harder (i.e., $b + 1$), only the FP rate at Cut 3 was lower than the original test. When the test was much harder (i.e., $b + 2$), FP rates at the three cuts were all higher than the original test. For the optimal test, FP rate for Cut 1 was slightly higher than the original

FP rate from the original test but the FP rate for Cut 2 and Cut 3 was 10% and 26% lower than the original test.

Increasing the average discriminating power of the overall test or only increasing the discriminating power of the low *a* or low and medium *a* items also decreased FN. When the test was easier (i.e., *b* – 1), FN rate increased at the three cutscores but was most severe at Cut 3 (increased from 19.06% based on the original test to 30.54%). When the difficulties of all items in the original test was shifted by +1.0, FN rate at Cut 1 increased but decreased at the higher cutscores. The decrease was more prominent at Cut 3 (decreased from 19.06% to 10.81%). When the difficulties of the original test was shifted by +2.0, FN rate increased by 100% at the lowest cutscore and increased by 50% at Cut 2, but decreased by 43% at the third cutscore. FN rate at the three cutscores for the optimal test were all lower than the original test: FN rate at Cut 1 decreased by 5%, 17% decreased at Cut 2 and 21% decreased at Cut 3.

4.5.1.3 Expected Information

Figure 4.69 presents the average information function provided by each item type based on original data. It can be seen that CR items provided more information than the MC and the SA items throughout the ability continuum. However, this type of item was considered to be quite easy for the target population as this type of item provided the most information for those examinees between -1.50 and -.50 on the proficiency scale. MC items provided more information than the SA items, on average, for ability score below .50; but SA items provided more information at the third cutscore than the MC items.

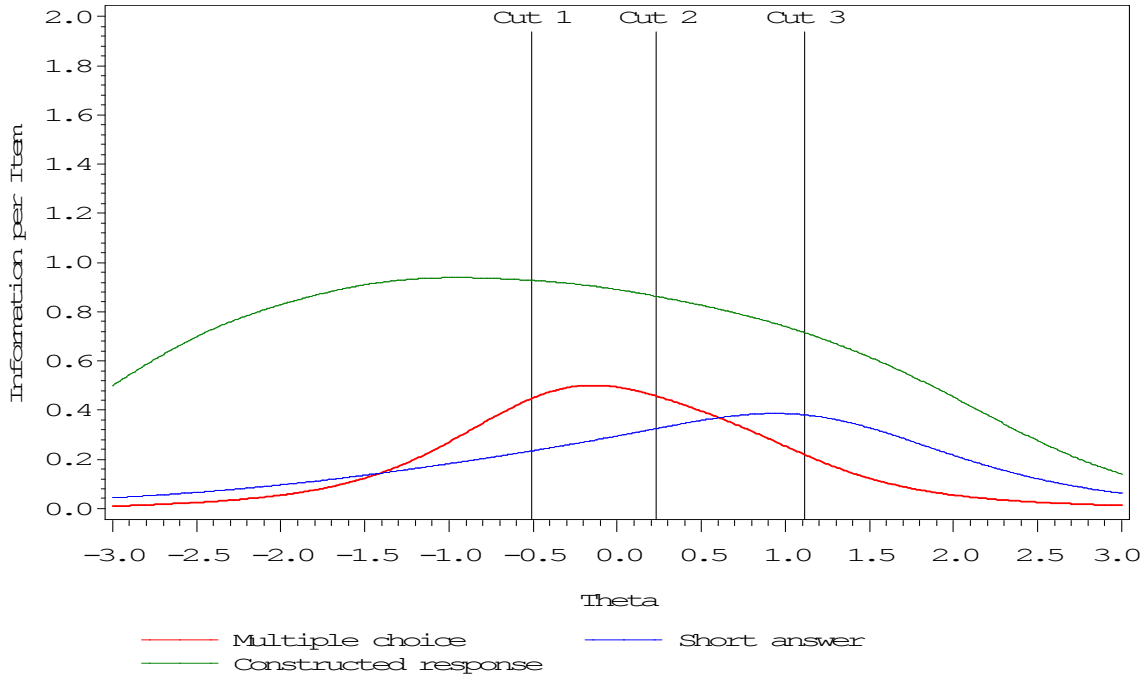


Figure 4.69 Middle school Mathematics test: Average information per item by item type

Expected information was used to determine how well these items functioned to the target population, and the results are summarized in Table 4.11 below.

As expected, increasing the average discriminating power of the test provided more information and thus the expected information would also be higher for each item type. When the test was easier, expected information for MC and CR were lower but the expected information for SA was higher. When the test became more difficult (i.e., b + 1), expected information for CR was higher but lower for MC and SA. And when test was much more difficult than needed, there were less information provided by all item types compared to the original test. Since the optimal test was built to target the ability of the population of interest, information provided by each item type would definitely be higher than those obtained from the original test.

Table 4.11 Descriptive Statistics for the Expected Information by Item Type for Middle School Mathematics Test.

	Item Type[1]	*n*	Mean	SD	Min	Max
Original Test	MC	29	.33	.15	.11	.65
	SA	5	.27	.09	.14	.37
	CR	5	.83	.13	.67	1.02
Increase average *a* by…						
.05	MC	29	.35	.15	.13	.68
	SA	5	.30	.09	.16	.39
	CR	5	.90	.14	.73	1.10
.10	MC	29	.37	.15	.13	.71
	SA	5	.32	.09	.18	.41
	CR	5	.97	.14	.79	1.19
.30	MC	29	.45	.17	.15	.82
	SA	5	.42	.10	.25	.49
	CR	5	1.27	.17	1.06	1.53
Change average *b* by…						
-1.00	MC	29	.25	.13	.04	.67
	SA	5	.29	.17	.09	.53
	CR	5	.68	.18	.50	.99
+1.00	MC	29	.26	.12	.07	.47
	SA	5	.19	.06	.15	.29
	CR	5	.86	.13	.69	1.02
+2.00	MC	29	.12	.07	.01	.28
	SA	5	.11	.07	.04	.19
	CR	5	.74	.14	.62	.95
Optimal Test	MC	29	.35	.14	.16	.65
	SA	5	.36	.13	.18	.49
	CR	5	1.03	.35	.67	1.53

[1] MC – Multiple choice items, SA – Short answer items, CR – Constructed response items

Table 4.12 presents the relative information for various item types for the conditions shown in Table 4.11. Relative information is the ratio of average expected information for polytomous items divided by the average expected information for each type of the dichotomous items.

Table 4.12 Relative Information for Middle School Mathematics Test.

	Item Type[1]	Relative Information
Original Test	CR/MC	2.52
	CR/SA	3.07
Increase average *a* by…		
.05	CR/MC	2.57
	CR/SA	3.00
.10	CR/MC	2.62
	CR/SA	3.03
.30	CR/MC	2.82
	CR/SA	3.02
Change average *b* by…		
-1.00	CR/MC	2.72
	CR/SA	2.34
+1.00	CR/MC	3.31
	CR/SA	4.53
+2.00	CR/MC	6.17
	CR/SA	6.73
Optimal Test	CR/MC	2.94
	CR/SA	2.86

[1] MC – Multiple choice items, SA – Short answer items, CR – Constructed response items

It can be seen that polytomous items always provided more information compare to the dichotomously scored items. In the original test, CR items yielded about 2.5 times more information than the MC items; and the CR items also yielded about 3 times more information than the SA items. When average *a* of the test was increased, the ratio between CR/MC and CR/SA were similar to those obtained from the original test. When the test difficulty did not match the ability of the target population, CR items still yielded more information than the MC and SA items; however, the patterns were somewhat different: when the test was easier, SA items provided more information than MC items and thus the ratio of CR/MC was bigger than CR/SA. One thing to note is that when the test was extremely difficult (i.e., $b + 2$), the ratio of the expected information for CR/MC

and CR/SA were over 6, this is because MC items and SA items became very inefficient in providing information about the examinees (average expected information were .12 and .11 for MC and SA items, respectively). For the optimal test, CR items yielded about 3 times more information than the MC items and also the SA items.

4.5.2 High School English Language Arts (ELA) Test

4.5.2.1 Decision Consistency

Table 4.13 summarizes the results of decision consistency (DC) and the *Kappa* statistics for the high school ELA test for all conditions.

Table 4.13 Summary of Decision Consistency (DC) and *Kappa* Statistic for High School ELA Test.

	Original Test	Increase average *a* by…			Increase low *a* by…		
		.05	.10	.30	.05	.10	.30
DC	72.21%	72.39%	73.57%	76.28%	71.53%	71.87%	72.72%
Kappa	59.13%	59.51%	61.20%	65.38%	58.20%	58.68%	59.93%

	Increase low and medium *a* by…			Change in average *b* by…			Optimal Test
	.05	.10	.30	-1.00	+1.00	+2.00	
DC	72.03%	72.40%	74.23%	71.63%	66.50%	55.81%	72.34%
Kappa	58.99%	59.45%	62.13%	56.85%	51.83%	37.44%	59.80%

As observed from the middle school Mathematics test, when average discriminating power of the overall test was increased, DC and *Kappa* also increased. In addition, if only increasing the discriminating power of the low discriminating items or those items with low or medium discriminating value, higher increase in the discriminating power would be needed (for example, increase *a* by more than .10 if only low discriminating items were considered; and increase *a* by at least .10 if low and medium discriminating items were considered).

Since the difficulty of the original high school ELA test aligned quite well with the ability of the target population, changing the difficulty of the test did not help to improve the rate of DC and *Kappa*, in fact, making the test easier or more difficult decreased both DC and *Kappa*.

The optimal test was built by increasing the discriminating parameters by .30 for all CR items, and increasing the difficulty level by 1.0 unit for 4 MC and all EI, by so doing slightly improved DC by .50% (from 72.71% based on the original test to 72.34% based on the optimal test) and *Kappa* by 1.13% (from 59.13% to 59.80%).

4.5.2.2 Decision Accuracy, False Positive and False Negative Error Rate

The results of decision accuracy (DA) at the overall test level and false positive (FP) and false negative (FN) error rate at each of the cutscore for the high school ELA test are summarized in Table 4.14 below.

When average discriminating power of the test increased, the rate of accurate decision classification also increased. In addition, the higher increase in the discriminating power was needed (at least greater than .10 increase in *a*) for the improvement in DA if only increasing the discriminating power of the low or low and medium discriminating items.

Table 4.14 Summary of Decision Accuracy (DA), False Positive (FP) and False Negative (FN) Error Rate for High School ELA Test.

	DA	False Positive			False Negative		
		Cut1	Cut2	Cut3	Cut1	Cut2	Cut3
Original Test	76.31%	17.40%	15.49%	13.48%	14.59%	16.71%	18.12%
Increase average *a* by…							
0.05	76.63%	16.70%	14.69%	12.90%	13.75%	15.95%	17.60%
0.10	76.86%	16.03%	13.94%	12.39%	12.99%	15.34%	17.19%
0.30	80.05%	13.62%	11.31%	10.61%	10.34%	12.81%	15.60%
Increase low *a* by…							
0.05	75.90%	17.37%	15.45%	13.45%	14.56%	16.67%	18.08%
0.10	76.00%	17.34%	15.41%	13.40%	14.53%	16.63%	18.06%
0.30	77.05%	17.24%	15.23%	13.25%	14.39%	16.42%	18.03%
Increase low and medium *a* by…							
0.05	75.92%	17.25%	15.29%	13.39%	14.39%	16.53%	18.09%
0.10	76.10%	17.00%	15.09%	13.29%	14.12%	16.35%	18.08%
0.30	77.24%	16.20%	14.27%	13.00%	13.37%	15.72%	18.13%
Change in average *b* by…							
-1.00	69.17%	14.00%	21.30%	21.89%	16.25%	24.48%	26.74%
+1.00	72.78%	26.76%	20.63%	8.17%	22.21%	17.75%	9.86%
+2.00	62.83%	35.96%	30.42%	13.20%	32.95%	26.01%	9.43%
Optimal Test	78.17%	17.44%	14.57%	10.93%	14.41%	14.91%	15.32%

Since test difficulty of the original test aligned quite well with the target population, any variations in the overall test difficulty would have a negative impact on DA. The optimal test increased DA from 76.31% based on the original test to 78.17%. If a state has 100,000 examinees tested in this subject area, this 2.44% increase in DA would affect about 2,000 more examinees being correctly classified into a performance category.

Similar patterns of improved false positive (FP) error rates were observed in the high school ELA test: increasing the overall discriminating power by .05 decreased the FP rate by about 4 to 5%; when the average discriminating power was increased by .10,

FP rate improved by about 8 to 10%; and when the average discriminating power was increased by .30, the FP rate increased by 21 to 27% at the three cutscores.

Only three multiple choice items were categorized in the low discriminating group (i.e. $a_i < .80$), FP error rate was improved by .20 to about 2%, depending on the level of increase in the discriminating power and also location of the cutscores.

When both low and medium discriminating items were considered in the redesign, depending on the level of increase in the discriminating power and also location of the cutscores, FP error rate decreased by 1 to 8%.

When the test was easier (i.e., $b - 1$), FP rate decreased at the first cutscore, but the FP rate at Cut 2 and Cut 3 increased by 38% and 62%, respectively. When the test became more difficult, FP rates were higher than the original test at the lower two cutscores, but lower at the third cutscore. For the optimal test, FP rate decreased at all cutscores, but the decrease was the most prominent at Cut 3.

Increasing the average discriminating power of the overall test slightly decreased the FN error rate for all cutscores. When only low discriminating items were considered, again, due to the low number of items considered in the group, the improvement in FN was at most 2%, depending on the level of increase in a and also the location of the cutscores. When both low and medium discriminating items were considered, the improvement in FN rate ranged from .1 to 6%.

When the test was easier (i.e., $b - 1$), FN rates for all cutscores increased. When the test was more difficult than is needed, FN rates were higher at Cut 1 and Cut 2 but lower at Cut 3. FN rates for the three cutscores in the optimal test were all lower than the original test, but the improvement was more obvious for the highest cutscore.

4.5.2.3 Expected Information

Average information function by item type based on original data is presented in Figure 4.70.

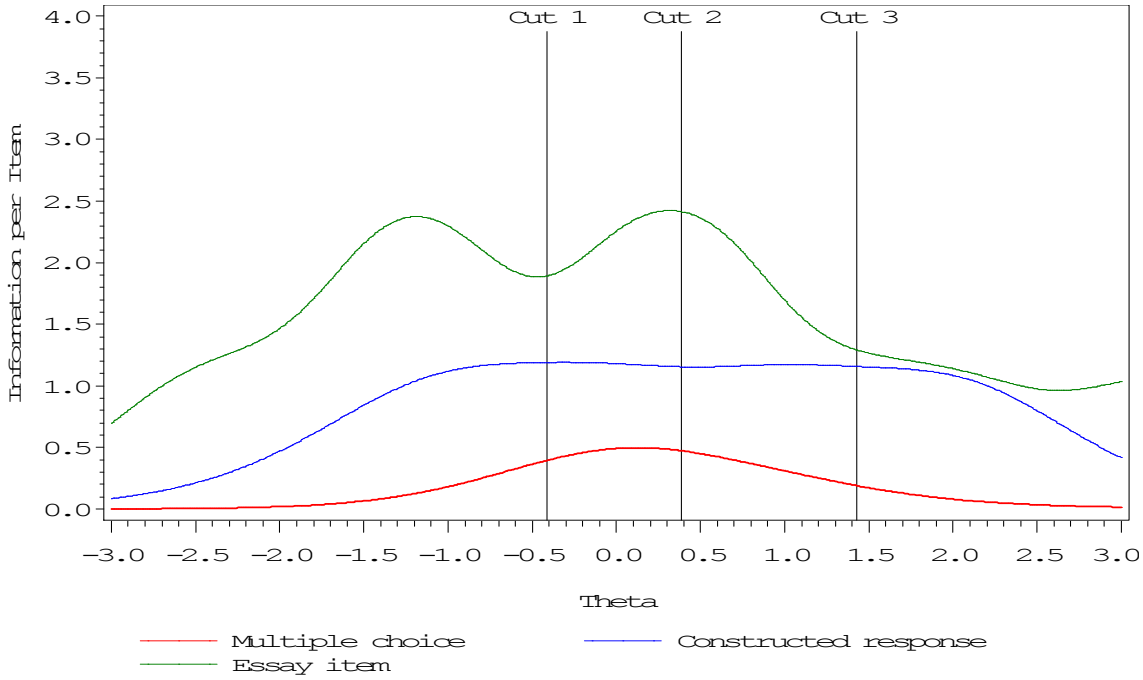


Figure 4.70 High school ELA test: Average information per item by item type

On average, EI provided more information than the MC and CR items throughout the ability continuum. EI provided more information at the low end of the ability continuum (-1.5 $\leq \theta \leq$ -1.0) and at around Cut 2 (0.0 $\leq \theta \leq$ 0.5). CR items provided relatively the same amount of information across the three cutscores, and MC items provided the least amount of information.

Results for the expected information are summarized in Table 4.15 below. Same as those results obtained from middle school Mathematics test, increasing the average discriminating power of the test provided more information and thus the expected information was also higher for each item type. When the test became easier, expected

information for all item types were lower. When the test became more difficult, there was less information provided by all item types compared to the original test.

Table 4.15 Descriptive Statistics for the Expected Information by Item Type for High School ELA Test.

	Item Type[1]	*n*	Mean	SD	Min	Max
Original Test	MC	36	.32	.11	.14	.64
	CR	4	1.11	.18	.95	1.36
	EI	2	1.99	.07	1.94	2.05
Increase average *a* by…						
.05	MC	36	.34	.12	.16	.66
	CR	4	1.19	.18	1.03	1.45
	EI	2	2.09	.07	2.05	2.14
.10	MC	36	.36	.12	.17	.68
	CR	4	1.27	.18	1.12	1.54
	EI	2	2.19	.06	2.15	2.24
.30	MC	36	.43	.12	.23	.76
	CR	4	1.62	.19	1.46	1.89
	EI	2	2.60	.04	2.58	2.63
Change average *b* by…						
-1.00	MC	36	.27	.12	.13	.56
	CR	4	1.09	.19	.96	1.38
	EI	2	1.70	.39	1.42	1.97
+1.00	MC	36	.21	.09	.07	.44
	CR	4	.92	.13	.76	1.08
	EI	2	1.91	.02	1.90	1.92
+2.00	MC	36	.08	.05	.01	.20
	CR	4	.55	.08	.44	.62
	EI	2	1.39	.34	1.15	1.63
Optimal Test	MC	36	.31	.11	.14	.64
	CR	4	1.62	.19	1.46	1.89
	EI	2	1.99	.07	1.94	2.05

[1] MC – Multiple choice items, CR – Constructed response items, EI – Essay items

Average expected information for MC items in the optimal test was only slightly lower than the original test; but expected information for CR items was higher and the expected information for the EI items remained the same.

Table 4.16 presents the relative information for different item types.

Table 4.16 Relative Information for High School ELA Test.

	Item Type[1]	Relative Information
Original Test	EI/MC	6.22
	CR/MC	3.47
	EI/CR	1.79
Increase average *a* by…		
.05	EI/MC	6.15
	CR/MC	3.50
	EI/CR	1.76
.10	EI/MC	6.08
	CR/MC	3.53
	EI/CR	1.72
.30	EI/MC	6.05
	CR/MC	3.77
	EI/CR	1.60
Change average *b* by…		
-1.00	EI/MC	6.30
	CR/MC	4.04
	EI/CR	1.56
+1.00	EI/MC	9.10
	CR/MC	4.38
	EI/CR	2.08
+2.00	EI/MC	17.38
	CR/MC	6.88
	EI/CR	2.53
Optimal Test	EI/MC	6.42
	CR/MC	5.23
	EI/CR	1.23

[1] MC – Multiple choice items, CR – Constructed response items, EI – Essay items

It can be seen that EI items always provided more information than the MC and also CR items. In the original test, EI items yielded more than 6 times more information than the MC items; the CR items yielded more than 3 times more information than the MC items; and the EI items also yielded about 1.8 times more information than the CR

items. When average *a* of the test was increased, the ratio between EI and MC decreased and the ratio between EI and CR also decreased; however, the ratio between CR and MC increased. This is because the rate of increase in information for MC and CR was higher than EI but the rate of information increase in CR was higher than MC.

When the test difficulty did not match the ability of the target population, EI and CR items still yielded more information than the MC items. However, MC items became almost useless in measuring the ability of the examinees (when overall test difficulty was increased by 1.0 unit, the average expected information for MC items was only .21 and .08 when test difficulty was increased by 2.0 units). Similar to the original test, EI items yielded about 6.4 times more information than the MC items based on the optimal test, CR items yielded about 5.2 times more information than the MC items, and the EI items yielded about 1.2 times more information than the CR items.

CHAPTER 5

DISCUSSION

Results of the two main studies were reported in detail in the previous chapter. This chapter provides a summary of the findings, limitations of the study, directions for future research, and conclusions.

5.1 Summary of Findings

The first study was intended to investigate the impact of various test quality aspects on test information, conditional standard errors of measurement and relative efficiency, based on empirical data. Specifically, the impact of better test quality by means of including higher discriminating items in the test (either at the overall test level or just by replacing some low or low and medium discriminating items with higher discriminating items), and also the impact of having a test that does not align with the ability of the target population, were studied. The goal for the second study was to examine the relationship between test quality and decision consistency, decision accuracy and also expected information, through simulated data.

5.1.1 Summary of Test Information, Conditional Standard Error of Measurement, and Relative Efficiency Results

At any given ability level, information and the conditional standard error of measurement has an inverse relationship: conditional standard error of measurement is simply the reciprocal of the square root of the information; therefore, the location where the information function peaks is where the location of the conditional standard error is at its lowest. Information functions usually peak at somewhere around the mid-range of the proficiency scale and taper off towards the extremes. Some observations based on the literature review and also results of the analyses regarding item and test information are:

- Information increases when the number of items in the test increases;
- Level of information also depends on the IRT model(s) chosen for the data;
- Item information functions tend to look bell-shaped;
- The shape of the information function depends on the distribution of the difficulty parameters of the test items and also the distribution of the discriminating parameters of the test items;
- High discriminating items have tall and narrow information functions but only contribute to a narrow range of proficiency scale; on the other hand, low discriminating items provide less information but to a broader range of ability level;
- When the average discriminating power of the test items increase, the level of test information will generally increase;
- Test information peaks at a point on the ability scale where item difficulties are clustered around that ability level and the maximum amount of information depends on the discriminating parameters;
- For dichotomously scored items calibrated with 3-PLM, when $a > 1.0$ and with minimal c generally have higher amount of item information;
- If the item bank does not have enough good quality items (i.e., high discriminating items), by only replacing the low discriminating items with those better discriminating items, the amount of information can be increased, thus lowering the measurement error;
- Increase in a has a bigger effect in information at the lower cutscore;

- However, increasing discriminating power could have a slightly adverse effect (i.e., higher measurement error) at the tails of the proficiency scale as observed in the middle school Mathematics test (when the average *a* of the overall test was increased by .30), and in the multiple choice portion of the high school ELA test (when the average *a* for the MC items was increased by .30);
- For the two tests included in this study, increasing the average discriminating power of the overall test by .05 would increase the effective length of the test near the cutscores by 5 to 6%; increasing the average discriminating power of the test by .10 increased the effective length of the test near the cutscores by 13 to 14%; and increasing the average discriminating power of the overall test by .30 increased the effective length to about 35 to 41%;
- However, manipulating difficulty of the test had different effects on the relative efficiency at different cutscores;
- Relative efficiency is a very helpful concept in comparing the effectiveness of the newly built test to the original test for the evaluation of how the improved test function at various locations of the proficiency continuum, for example, at the cutscores. Based on the results of the relative efficiency, test developers could either shorten the test and still achieved the same level of predefined measurement precision; or they could choose to further modify the test until it fits their purposes.

5.1.2 Summary of Decision Consistency, Decision Accuracy and Expected Information Results

As mentioned in the previous section, test length and item characteristics affect the amount of information provided at different points on the proficiency scales which would ultimately affect the decision consistency (DC) and decision accuracy (DA) of proficiency classifications. Below is a list of observations based on the results obtained from the simulated data:

- When the average discriminating power of a test is increased, DC, *Kappa* statistic, and DA are also increased;
- If only a portion of items in a test are replaced due to their low discriminating power, discriminating value for those replaced items need to be at least .10 higher than their original values in order to improve DC, *Kappa*, and DA;
- DC and *Kappa* are negatively affected when difficulty of the test does not match with the ability of the intended population. However, the effect is less severe if the test is easier than needed for mixed format tests. Results of DC and *Kappa* suffered most when test difficulty is much more difficult than needed (i.e., $b + 2.0$);
- When a test is substantially more difficult than is needed, result of DC is close to the minimum possible value (i.e., .50);
- When test difficulties are somewhat aligned with the ability of the target population, false positive (FP) error rate decreased but false negative error (FN) rate increased when cutscores moved up on the ability scale, regardless of the discriminating power of the test;

- Increasing the average discriminating power of a test lowered FP and FN rates at the three cutscores in both tests in this study;
- DA decreases if test difficulty does not align with the ability of the target population. However, the impact was less severe when the test is slightly more difficult than needed in the two subjects studied;
- When a test is easier than needed, it only improved the FP rate at the lowest cutscore;
- When a test is more difficult than is needed, FP rate increased dramatically. Except for the FP rate for the third cutscore from the two tests in this study where overall test difficulty is increased by 1.00 unit;
- In general, FN rate increases when a test is easier than needed;
- When tests are more difficult than needed, FN rates decrease at the highest cutscore but increase at the lower cutscores;
- Polytomous items tend to yield more information than dichotomously scored items, regardless of the discriminating power and difficulty of the items;
- The more score categories an item has, the more information it can provide.

5.2 Limitations of the Study

Since only two empirical tests were examined in this dissertation, findings from these two tests can only be extended to those mixed format tests that are similar in the composition of the item format and characteristics of the items. Although data were simulated to mimic the realistic response data, other possible conditions could have

happened in an operational testing program, for example, errors in the scoring rubric, which could seriously affect the results of the expected information and also the relative information provided by each item format.

5.3 Directions for Future Research

As noted in the early chapters, item parameters and item information functions are well-studied for dichotomously scored items; however, still not too much is known about how the item parameters affect the level of information provided by polytomous items. Therefore, it would be beneficial to the measurement field if simulation studies can be conducted to examine the patterns of step difficulties on the level of information provided by different polytomous IRT models. For example, the study can examine different magnitudes of the step difficulties for the graded response model and its effect on the item information function. For instance, fixing the overall item difficulty to be equal, manipulation could be done on the distance of individual step difficulties: steps are of equal distances; distances of initial steps are closer than latter steps and vice versa. Similar simulation studies could be done for the partial credit models with an extra condition that step difficulties are not necessarily in order as these models assume each of the two adjacent categories in a polytomously scored item as dichotomous case. Therefore, study conditions can also include scenarios when step difficulties are out of order, for example, how the information functions appear when step difficulties are out of order in the initial steps versus when the out of order happened at the latter steps.

This study only considered tests with over 50% of the test scores coming from multiple choice items. As shown in the results of these studies, polytomous items always provide more information, hence better measurement precision than multiple choice

items, it would be interesting to know if more polytomous items could be included in a test to replace dichotomously scored items, how it would affect the rate of DC and DA.

In addition, IRT information functions are only useful when the model fits the data, a study can be conducted to examine the effect of model misfit on information function, and how it would affect DC and DA.

5.4 Conclusion

IRT information functions have a critical role in test construction and evaluation as it reflects the test's reliability by providing overall test precision information. It has the capability to produce a test that has the desired precision of measurement for any defined proficiency scale when sufficient number of test items are available. This feature is extremely useful when the information is used for decision making, for instance, whether an examinee passes a test or obtains a licence from a certified professional exam. Some examples for using IRT information functions in test construction are automated test assembly and computerized adaptive testing (CAT).

Since test information is simply the sum of item information of all the items in a test, plots of individual item information can be used to examine where on the proficiency scale the item contributes and how much information it can contribute. With a large item bank, measurement error can be controlled very precisely at various locations on the proficiency scale by shaping the test information function. Therefore, except for the constraint of content validity, test developers have full control to select items independently from other items and would be able to know the consequences of selecting a particular item to include in a test.

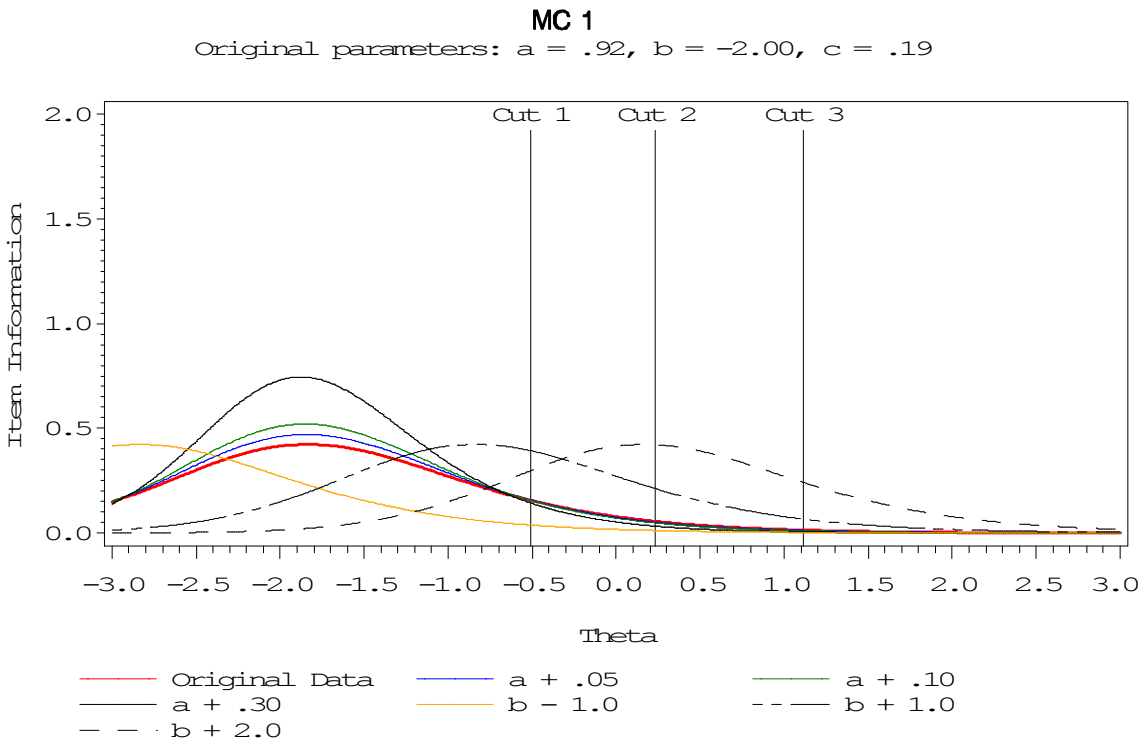
Based on the findings from this study, it is obvious that item quality has a very important role in information function and thus measurement precision. Specifically, discriminating power of the item plays a crucial role in ensuring high accuracy and consistency of proficiency classification decisions. Therefore, it is important to have more good quality items included in the bank. This study also produced some interesting results which can provide test developers with ideas about the impact of including high discriminating items and/or including items that are too easy or too difficult in a test on decision consistency and decision accuracy. The results indicated that it is important to have more items that are targeted at the population of interest. Otherwise, no matter how good the quality of the items may be, they are of less value in test development when they are not targeted to the distribution of candidate ability or at the cutscores.

So how can the quality of test items be improved? In testing practice, in order to improve the quality of items in an item bank, a number of options are available: improving item writer training, cloning the best items, and improving and extending field testing of new items, offer the potential for improving the statistical characteristics of items. However, high quality test items are by no means easy to obtain. Apart from better training of item writers and more use of cloning, in order to optimize the item bank, there is also a need to study how to write items to fulfill particular statistical specifications. If more could be learned about what makes items difficult or discriminating, better quality items can be produced. When items are better targeted statistically, the utility of an item bank could be enhanced. With more good items, two options are available – increasing test information without increasing test length, or maintaining test information and lowering the level of item exposure because more forms can be constructed with

additional test items. However, it is important to remember that effective use of IRT test information requires a diverse and high quality item pool, IRT model that fits the data, and also item statistics that are estimated with good precision. Nevertheless, item writers need to be conscientious about the fact that the information function is merely a statistical tool in building a good test, other criteria should also be considered, for example, content balancing and content validity.

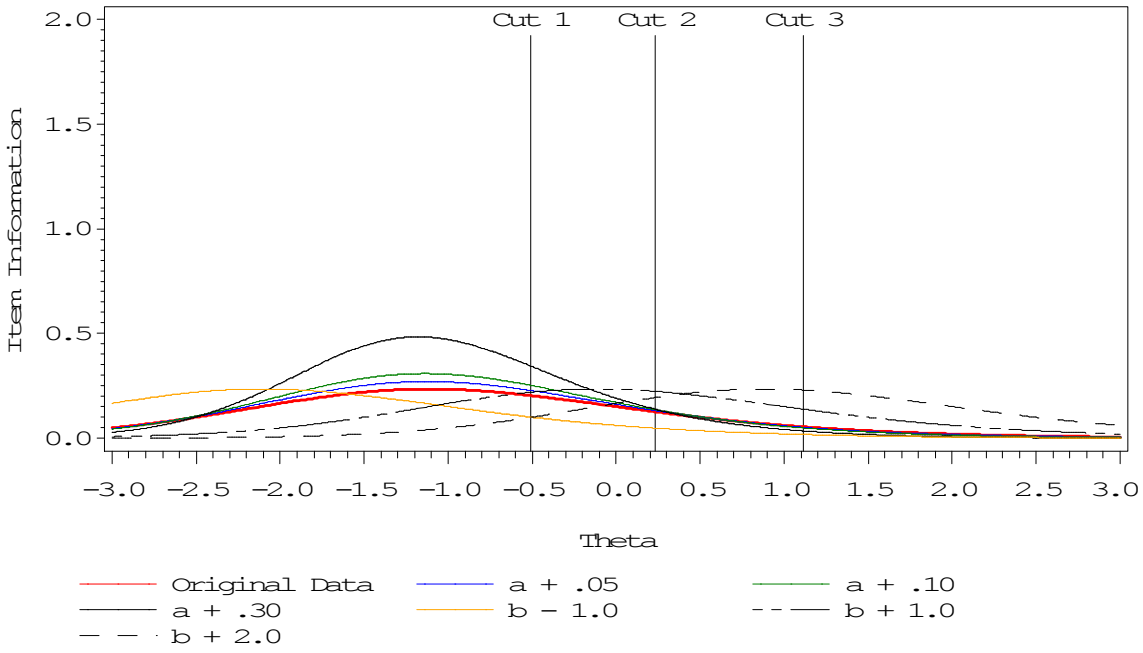
APPENDIX A.

ITEM INFORMATION FUNCTIONS FOR A MIDDLE SCHOOL MATHEMATICS TEST



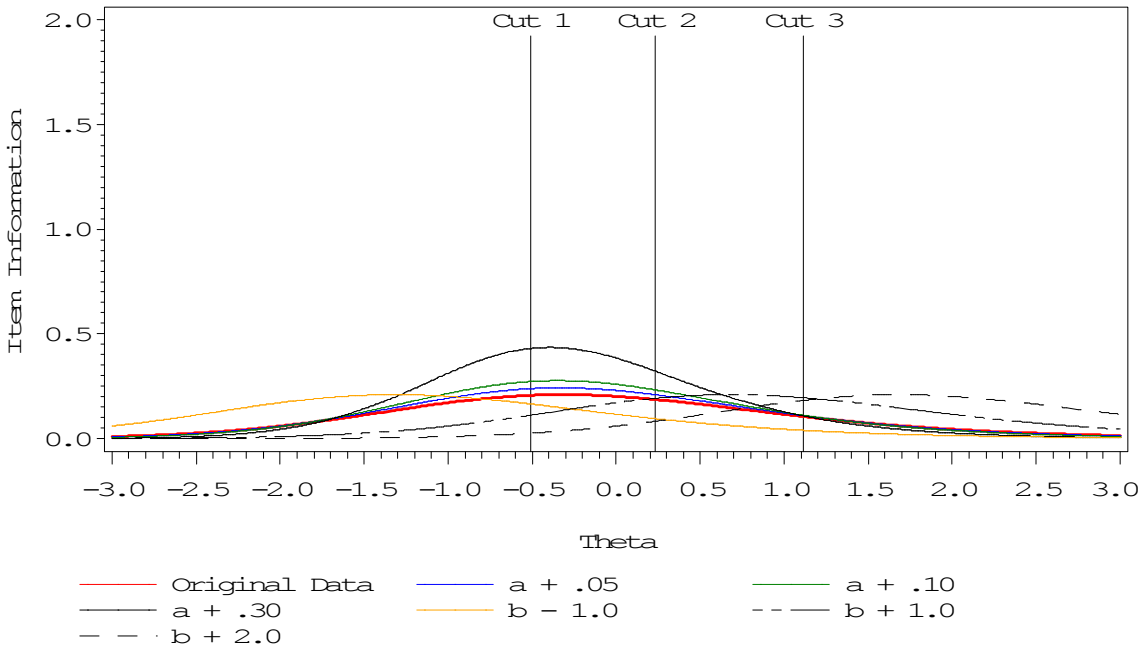
MC 2

Original parameters: a = .69, b = −1.34, c = .20



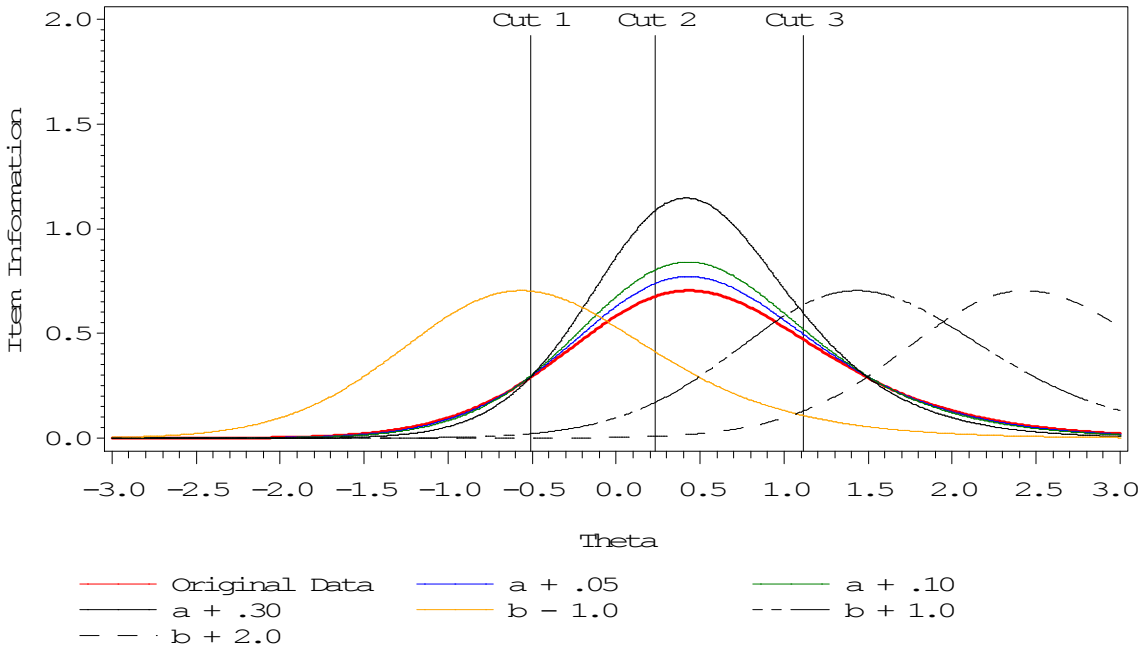
MC 3

Original parameters: a = .69, b = −.58, c = .25



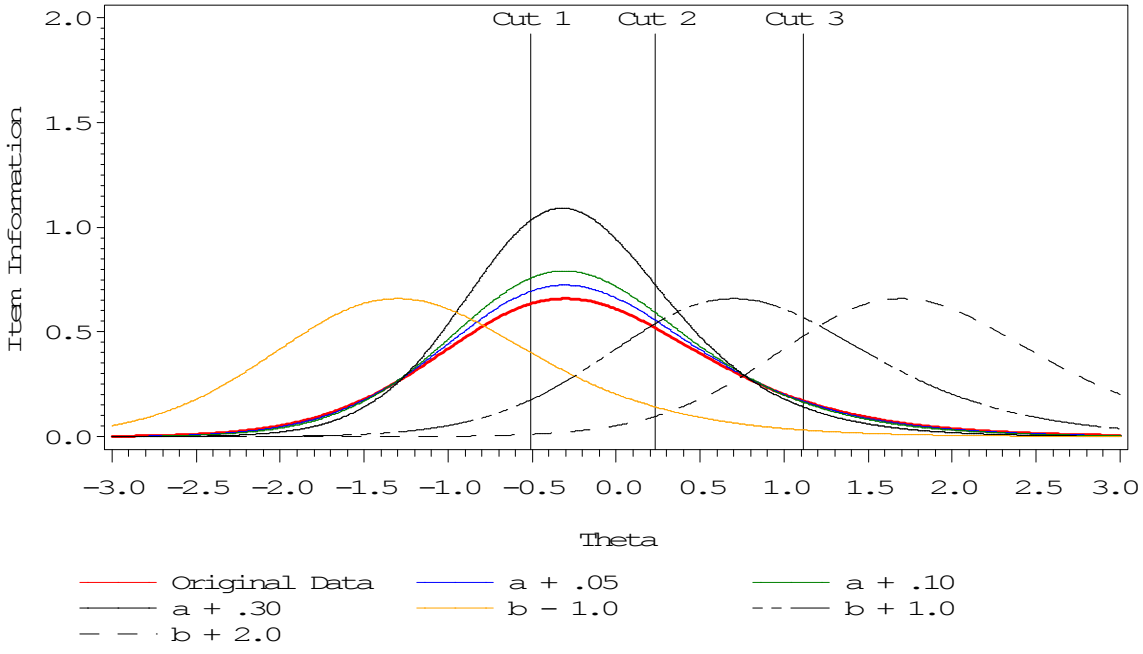
MC 4

Original parameters: a = 1.09, b = .35, c = .10



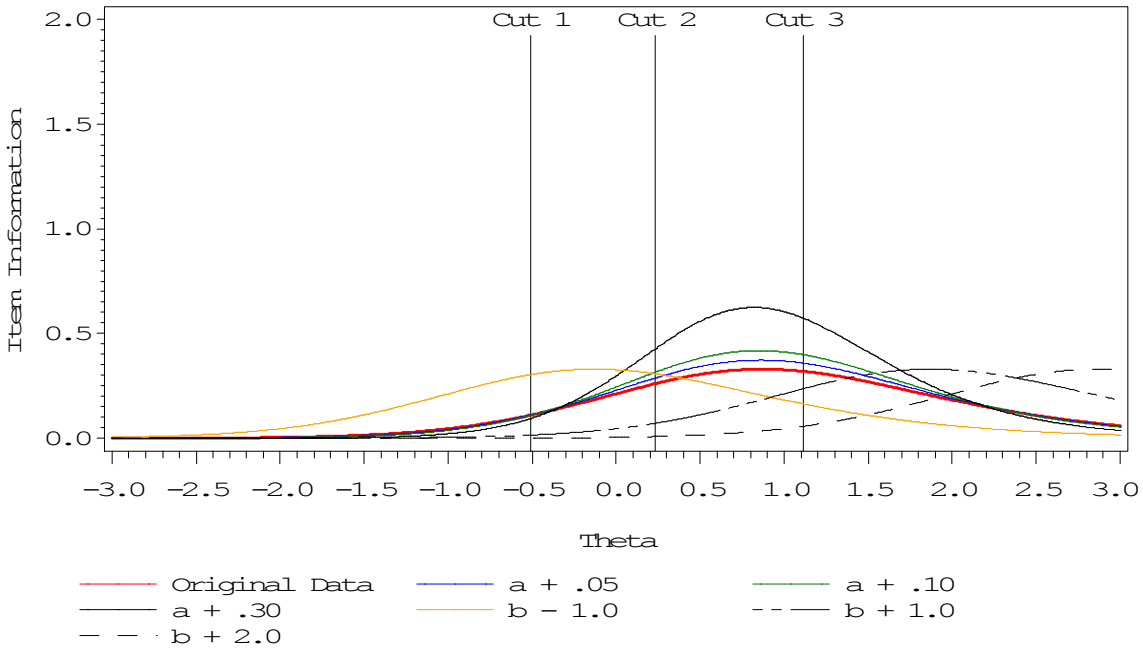
MC 5

Original parameters: a = 1.05, b = -.39, c = .10



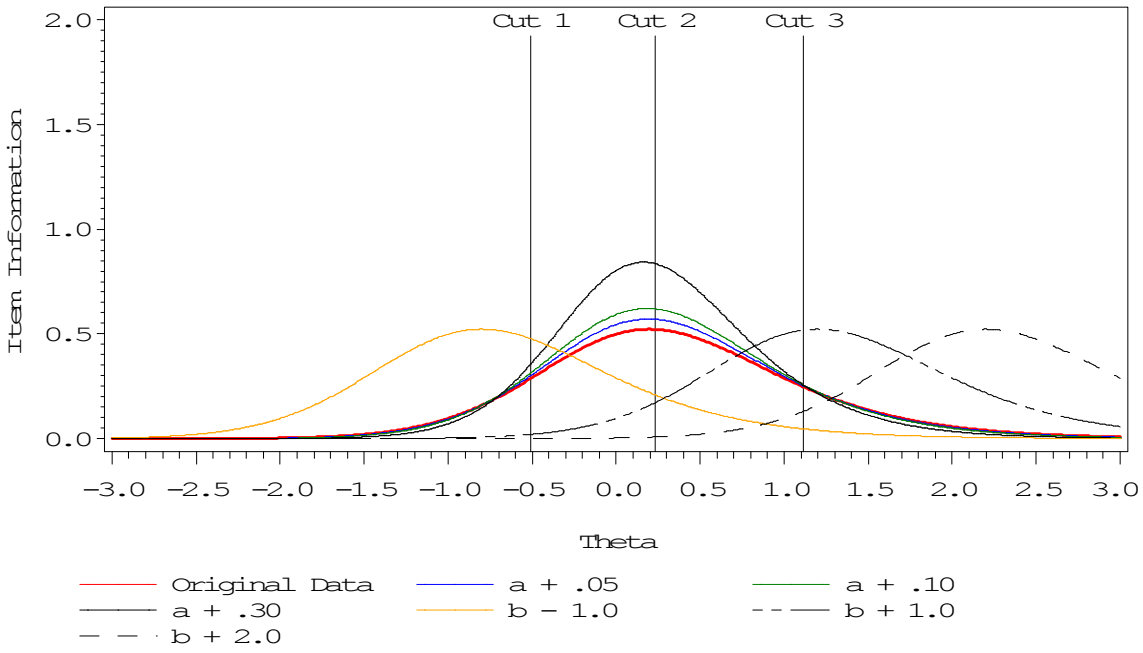
MC 6

Original parameters: a = .80, b = .69, c = .18



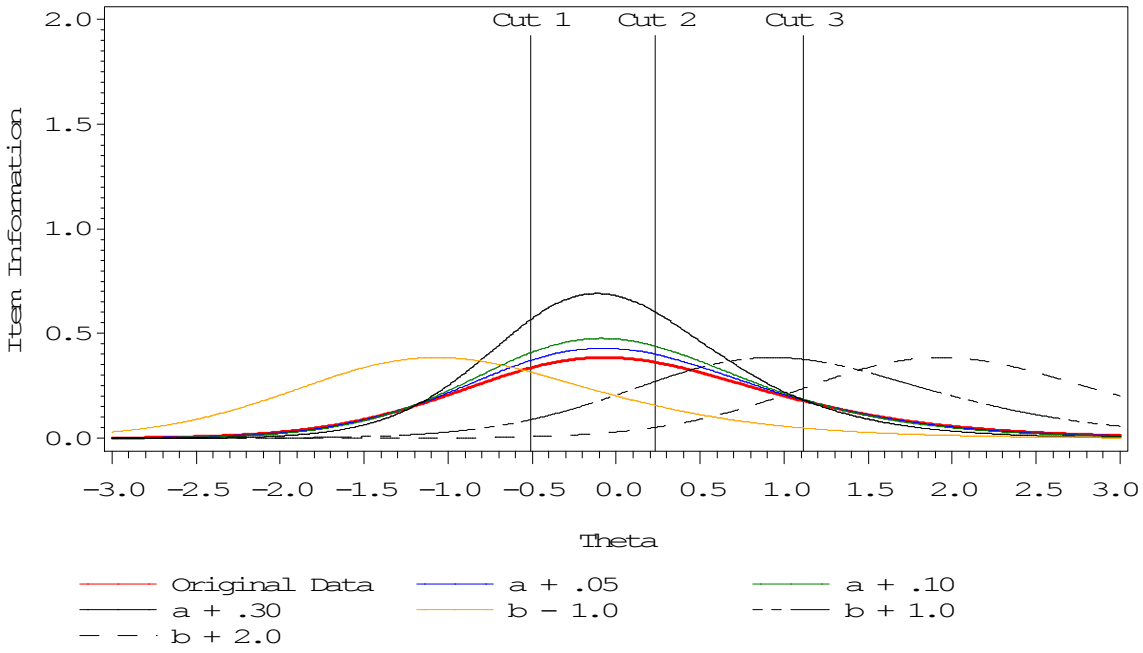
MC 7

Original parameters: a = 1.11, b = .02, c = .28



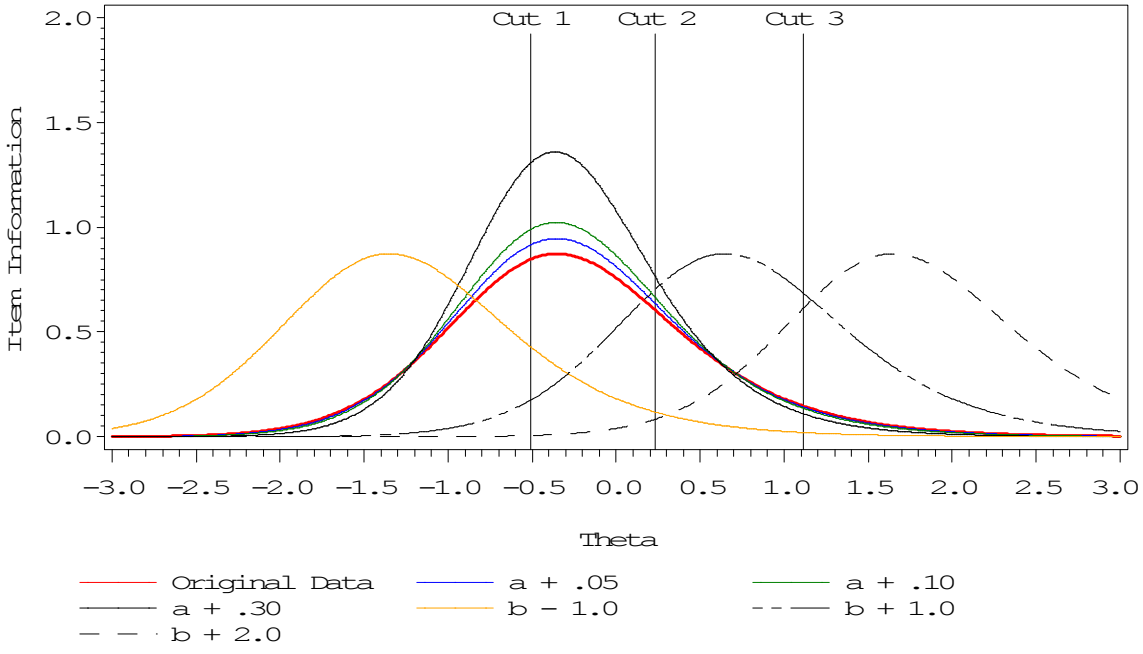
MC 8

Original parameters: a = .88, b = −.25, c = .20



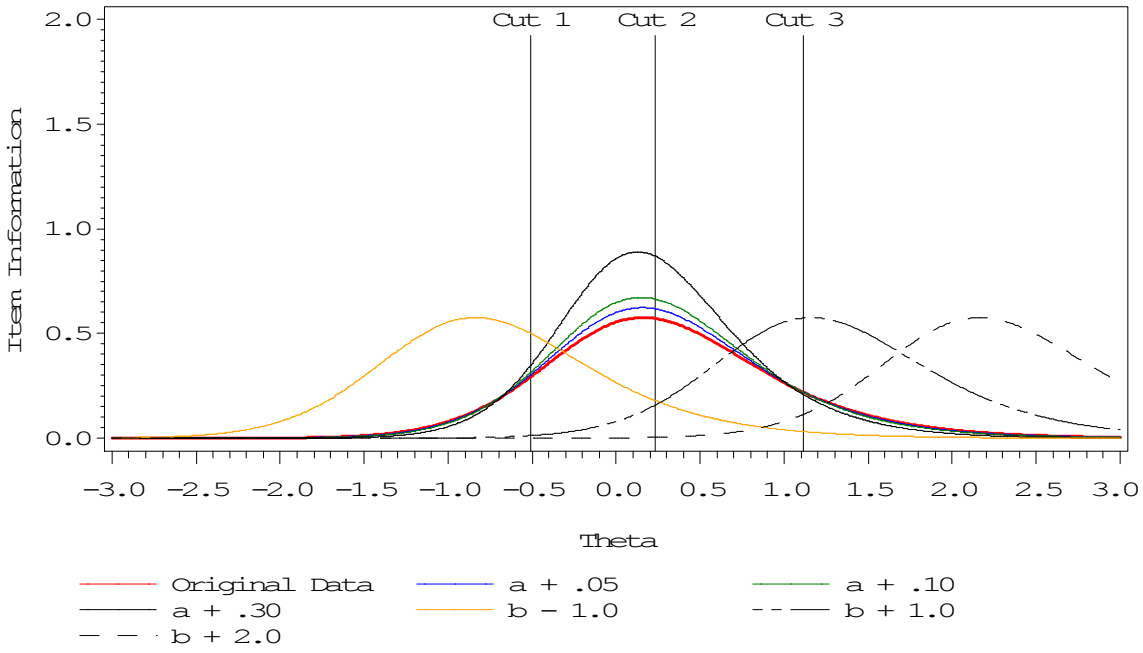
MC 9

Original parameters: a = 1.21, b = −.43, c = .10



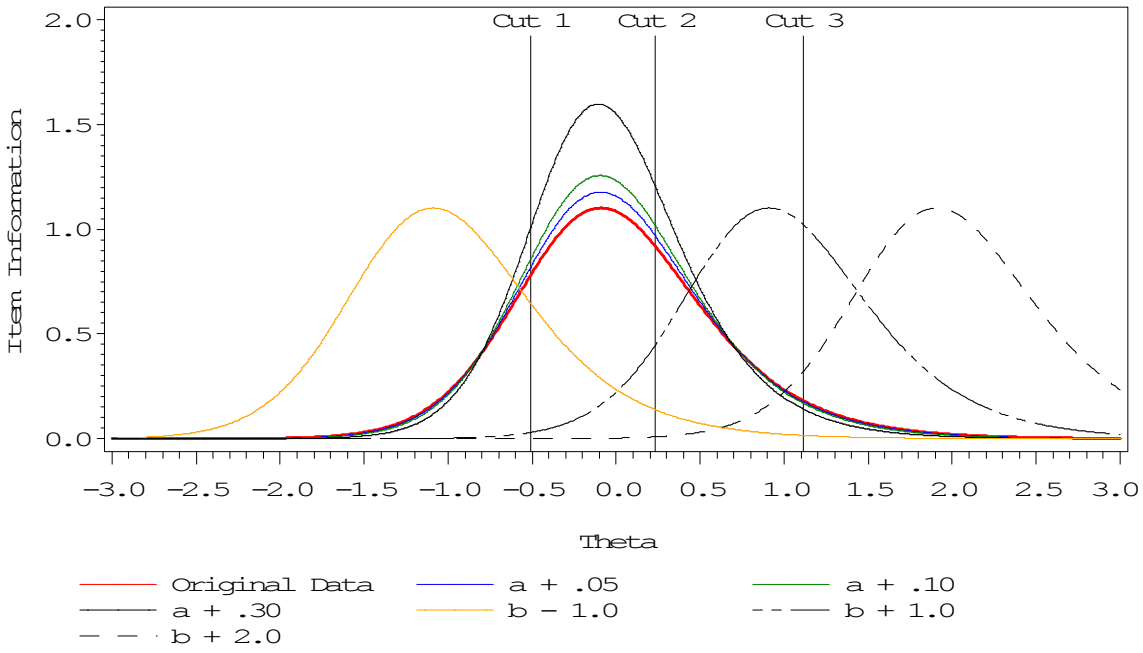
MC 10

Original parameters: a = 1.24, b = −.02, c = .34



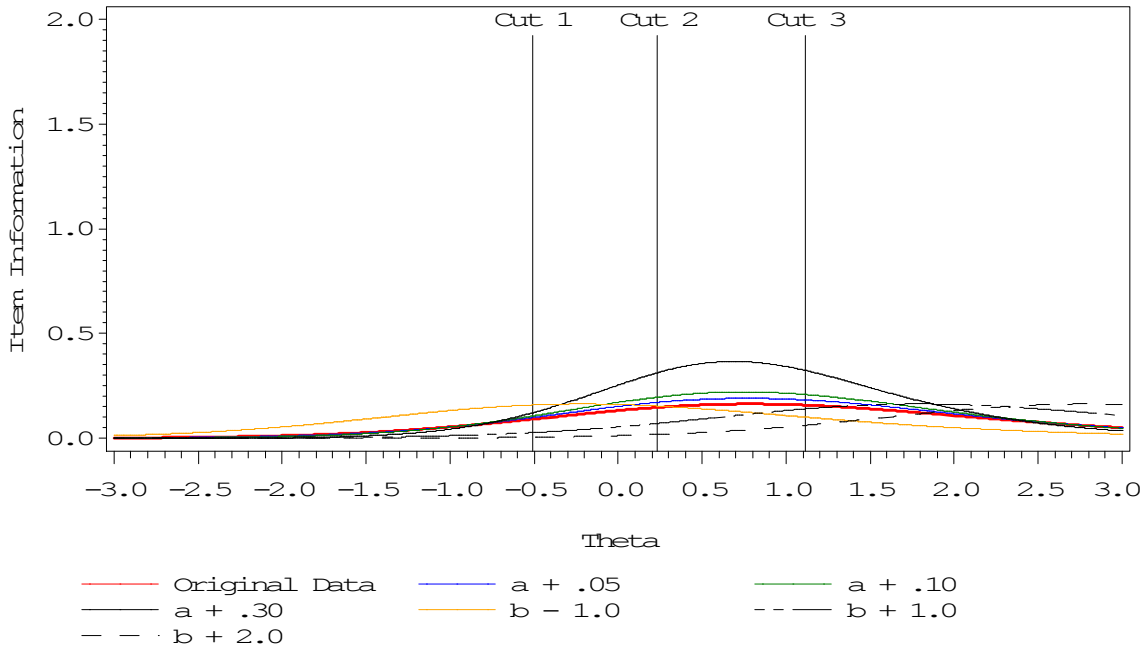
MC 11

Original parameters: a = 1.47, b = −.19, c = .18



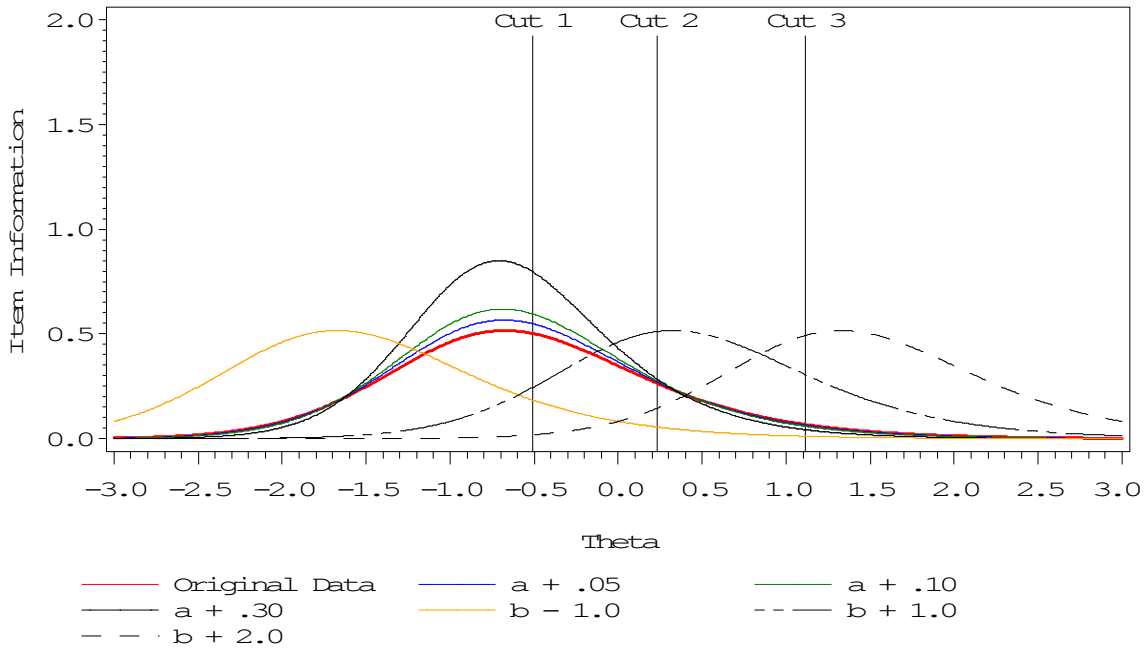
MC 12

Original parameters: a = .60, b = .48, c = .25



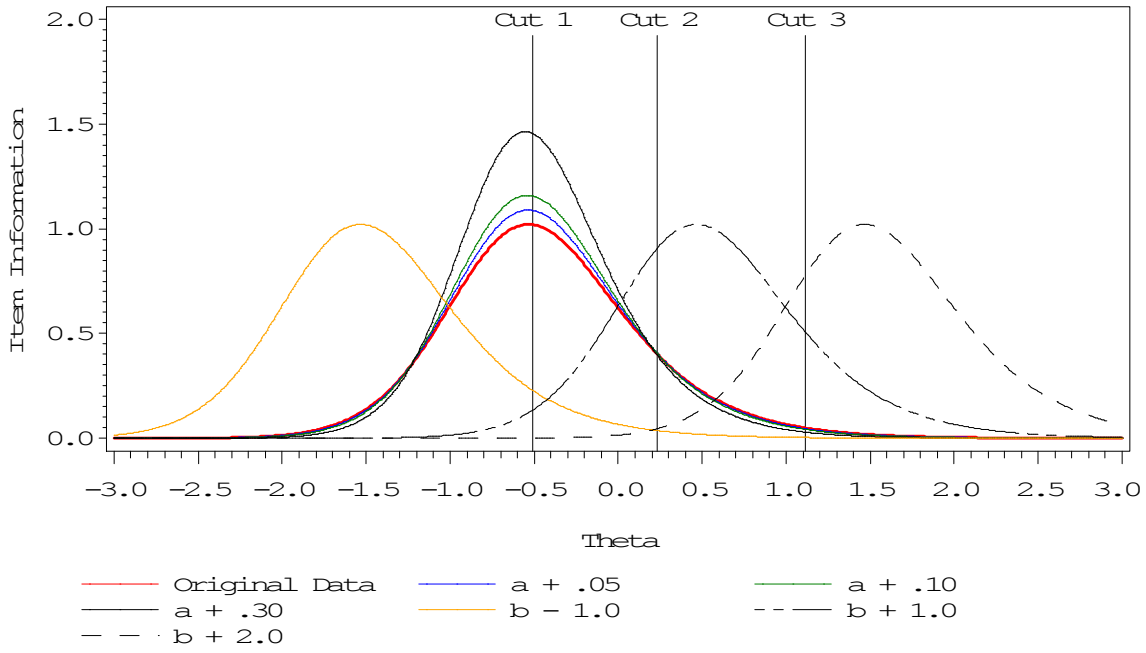
MC 13

Original parameters: a = 1.06, b = -.84, c = .24



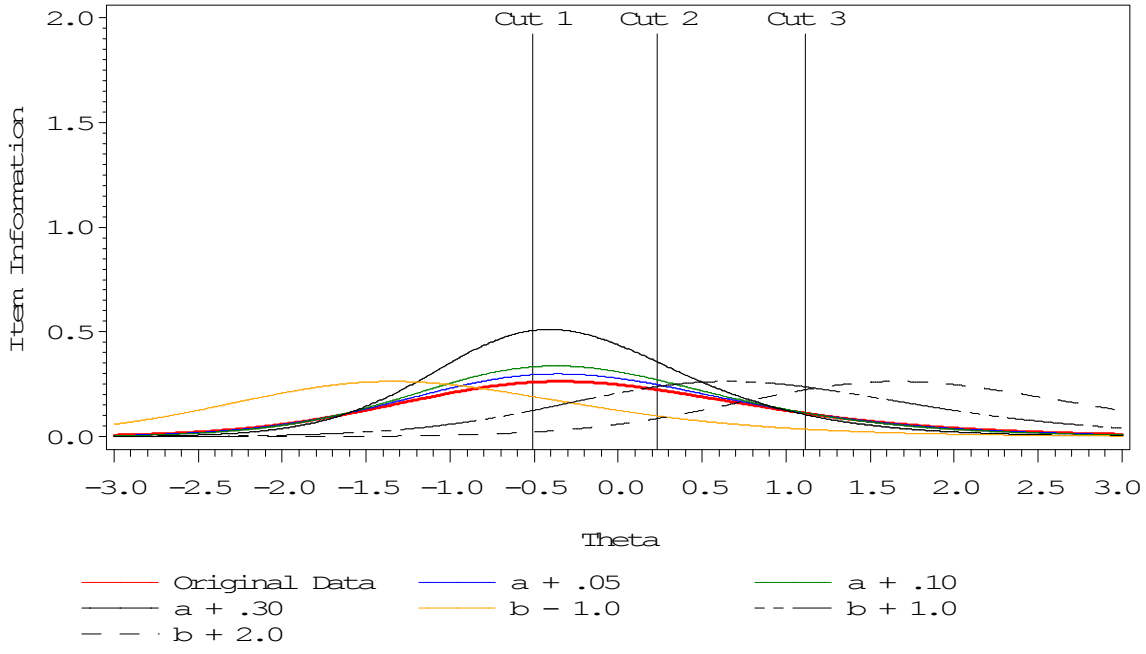
MC 14

Original parameters: a = 1.52, b = −.66, c = .26



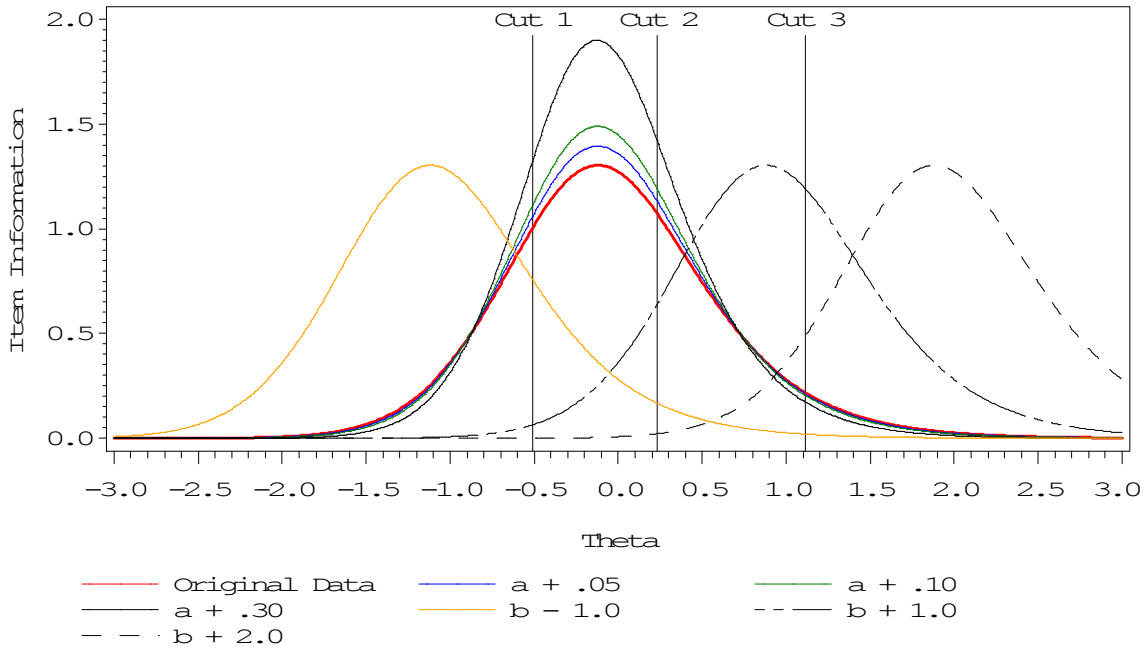
MC 15

Original parameters: a = .76, b = −.58, c = .24



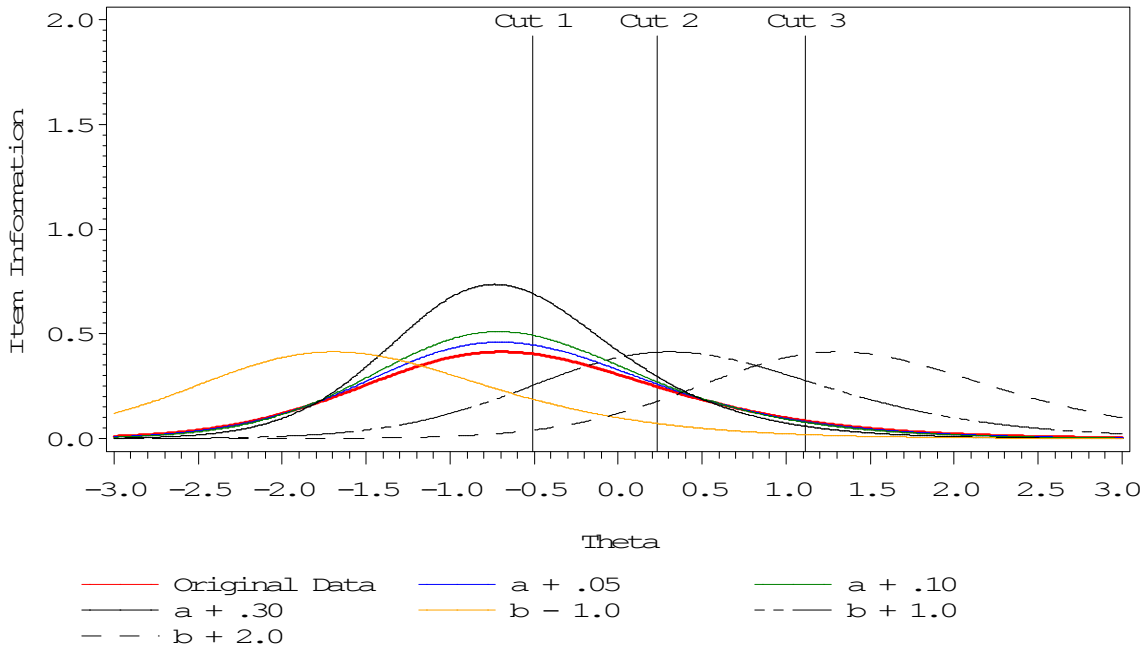
MC 16

Original parameters: a = 1.44, b = .17, c = .08



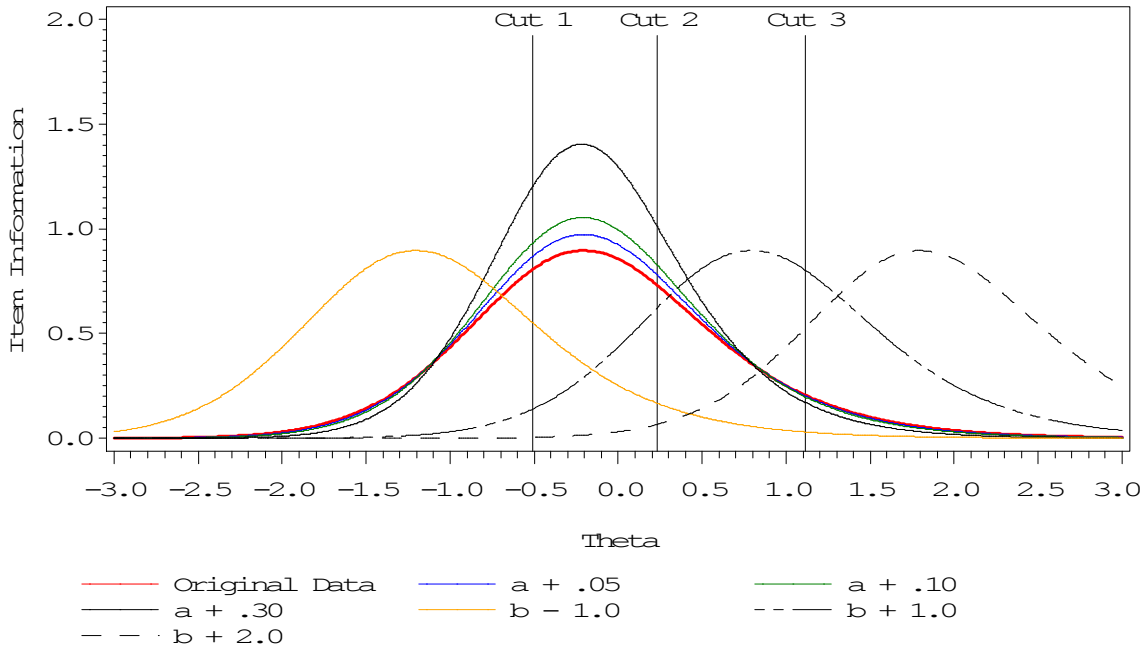
MC 17

Original parameters: a = .90, b = −.86, c = .18



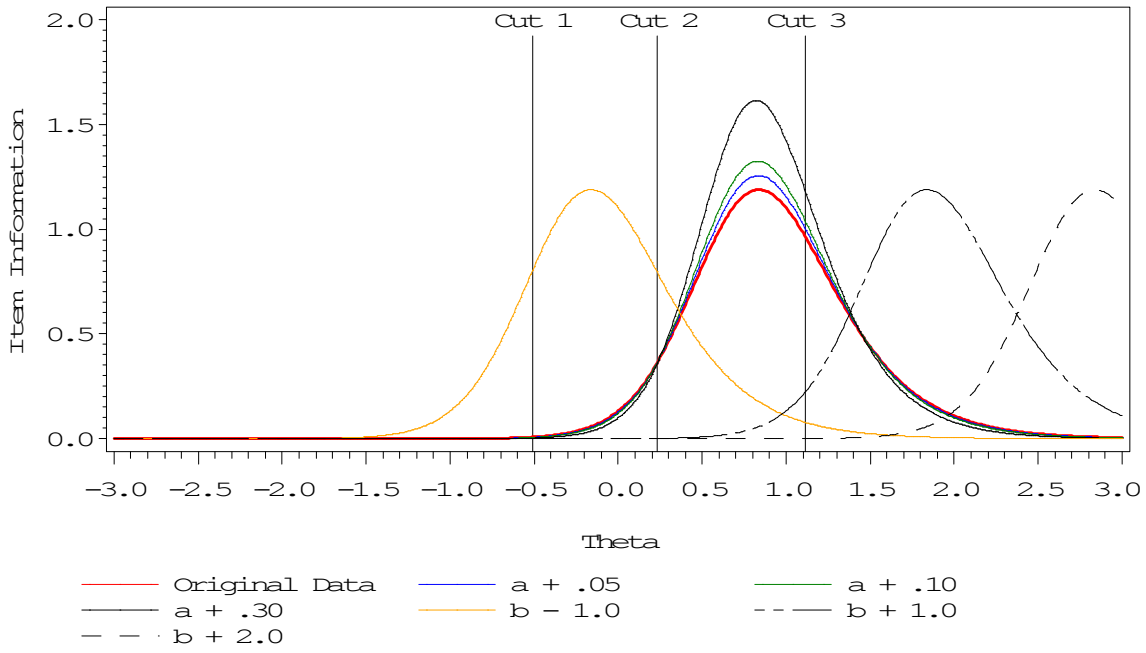
MC 18

Original parameters: a = 1.19, b = −.26, c = .07



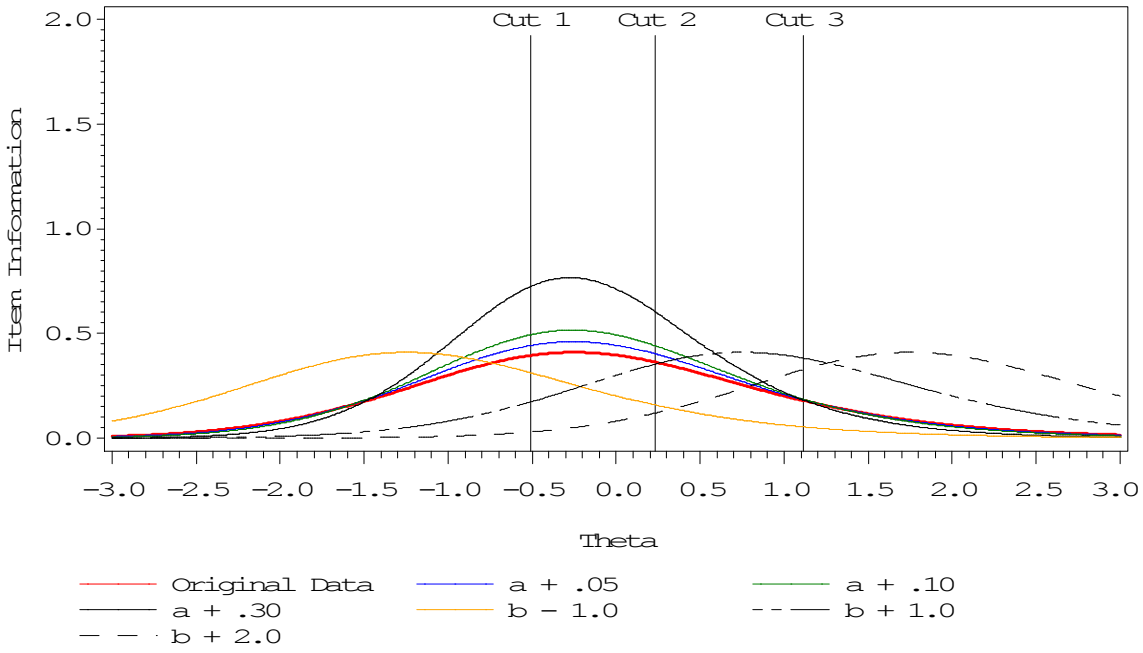
MC 19

Original parameters: a = 1.82, b = .71, c = .36



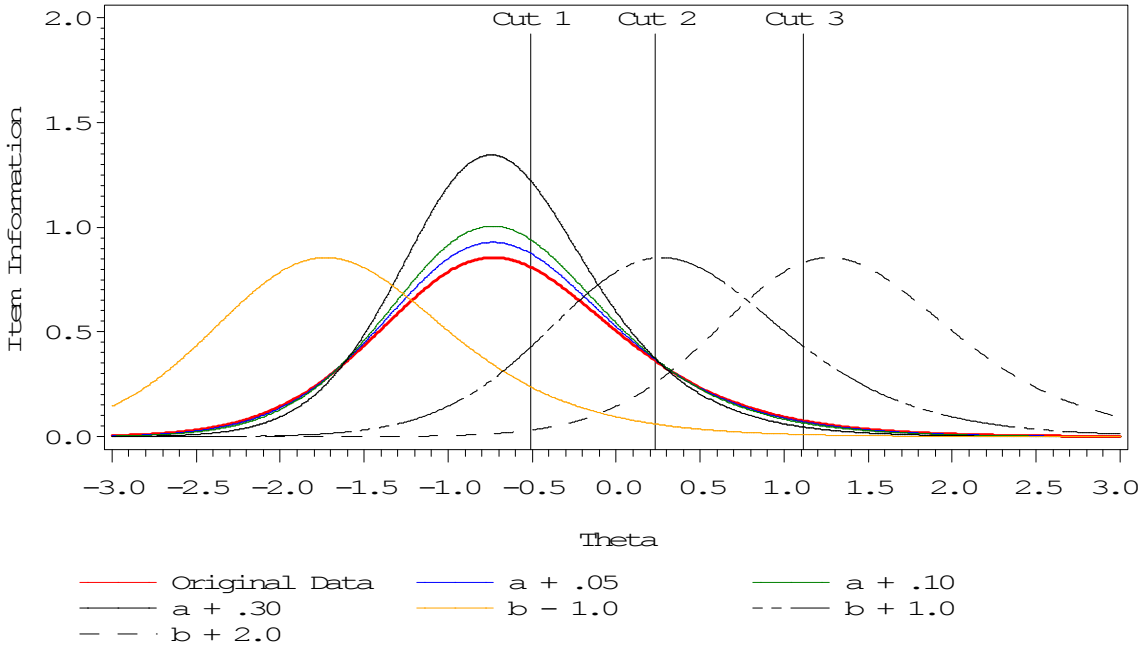
MC 20

Original parameters: a = .81, b = −.35, c = .08



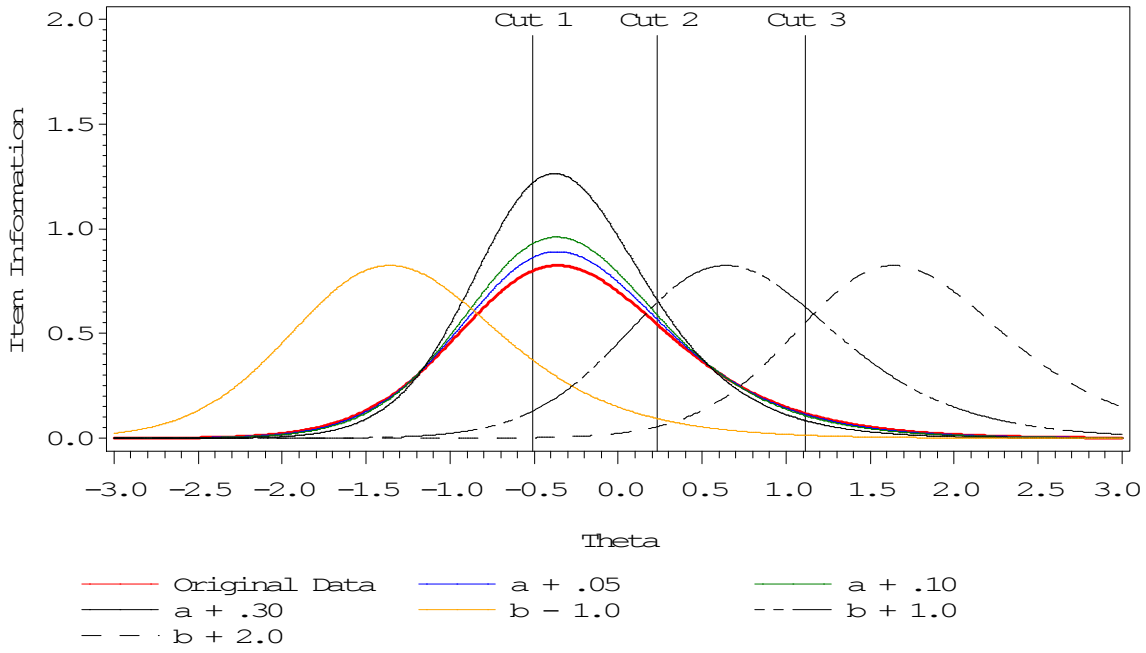
MC 21

Original parameters: a = 1.18, b = −.80, c = .08



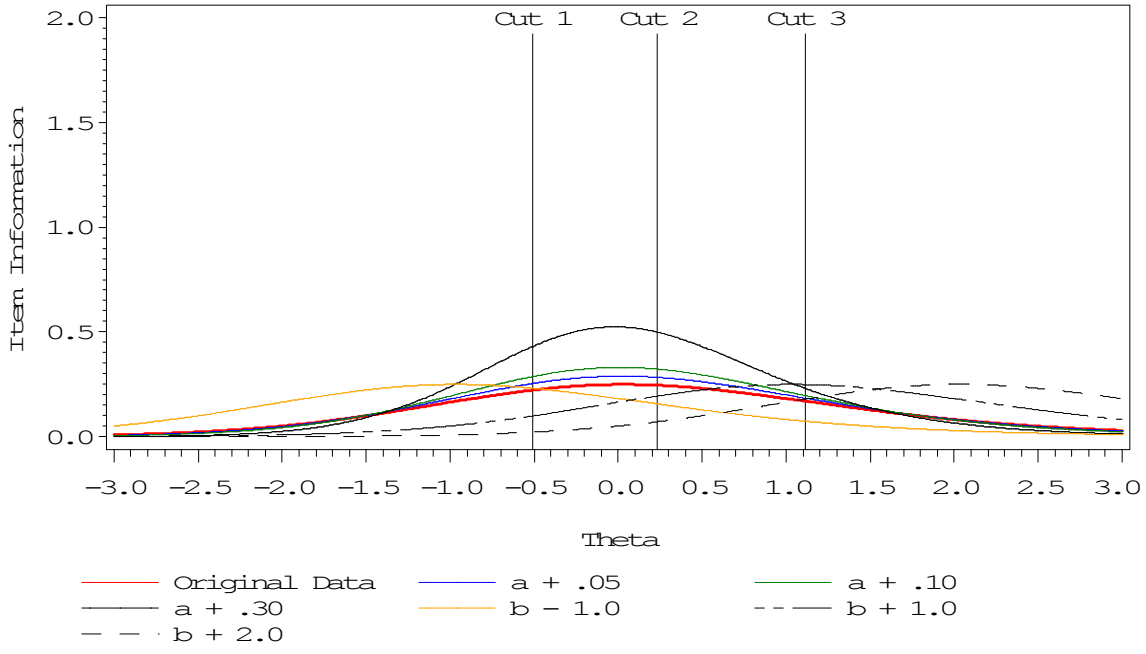
MC 22

Original parameters: a = 1.26, b = −.47, c = .17



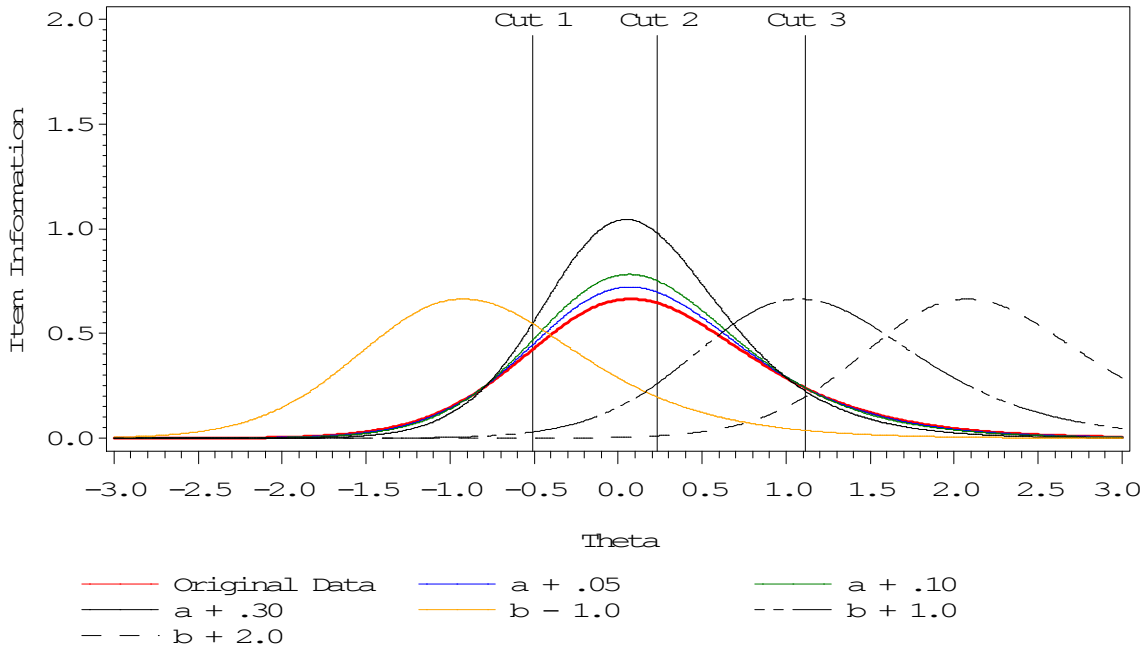
MC 23

Original parameters: a = .67, b = −.14, c = .13



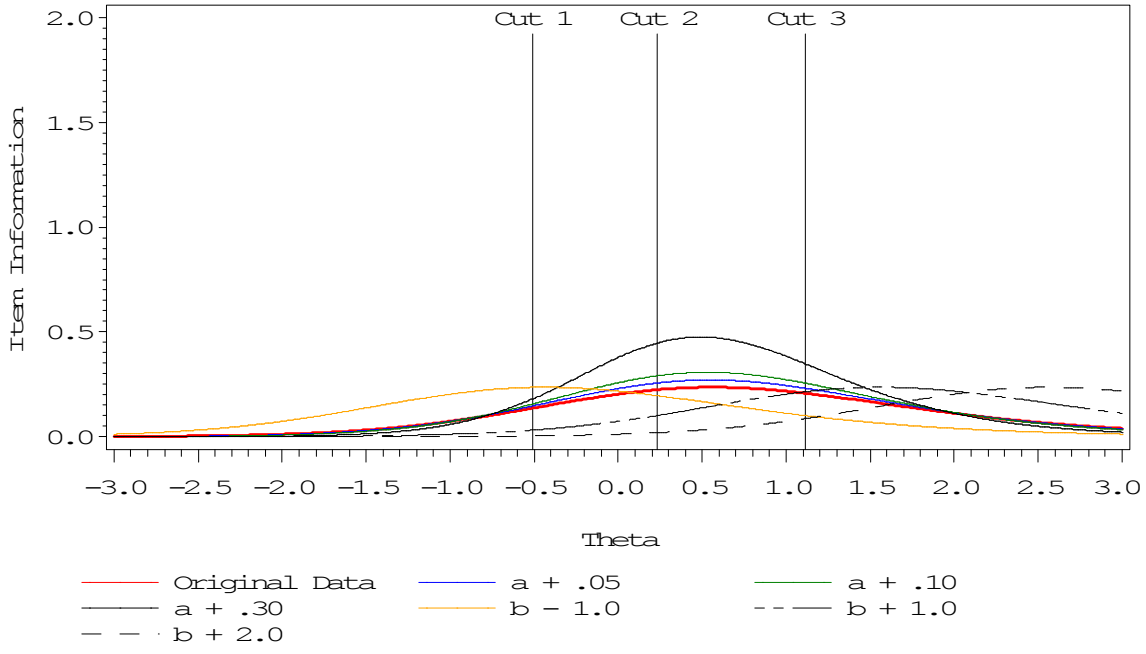
MC 24

Original parameters: a = 1.19, b = −.07, c = .22



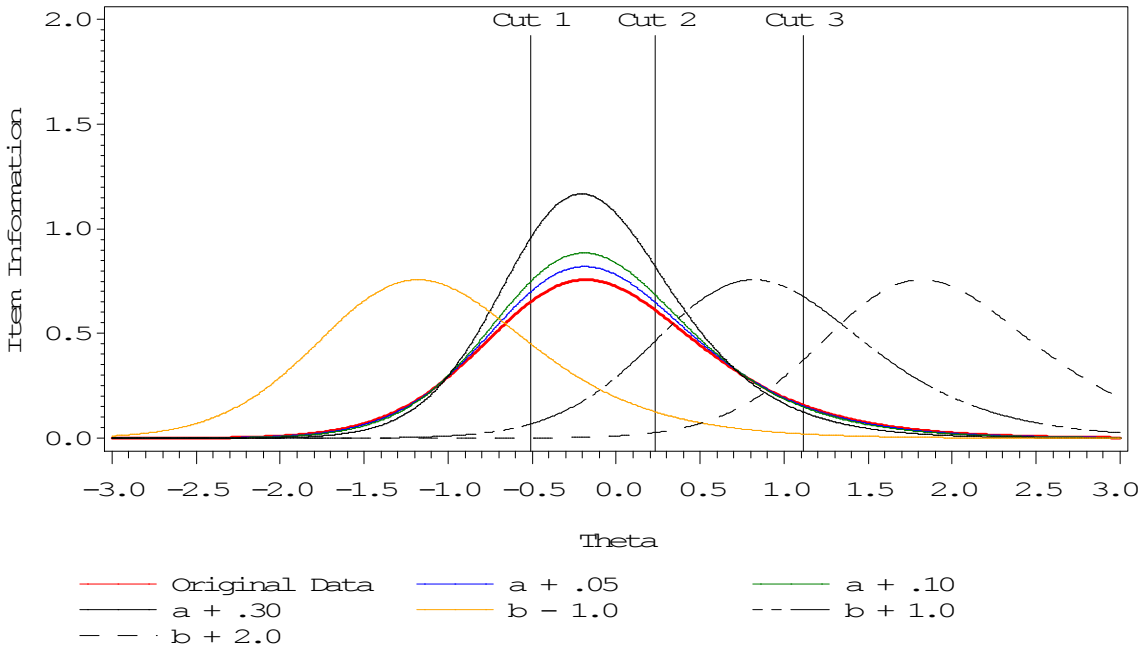
MC 25

Original parameters: a = .72, b = .32, c = .23



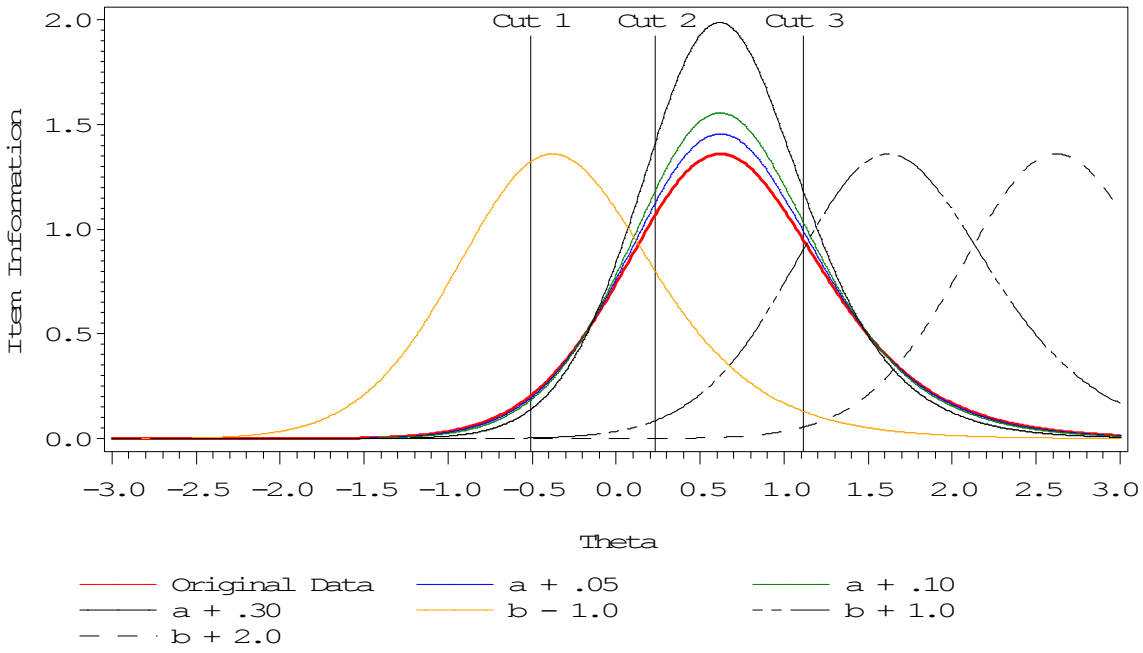
MC 26

Original parameters: a = 1.24, b = −.31, c = .20



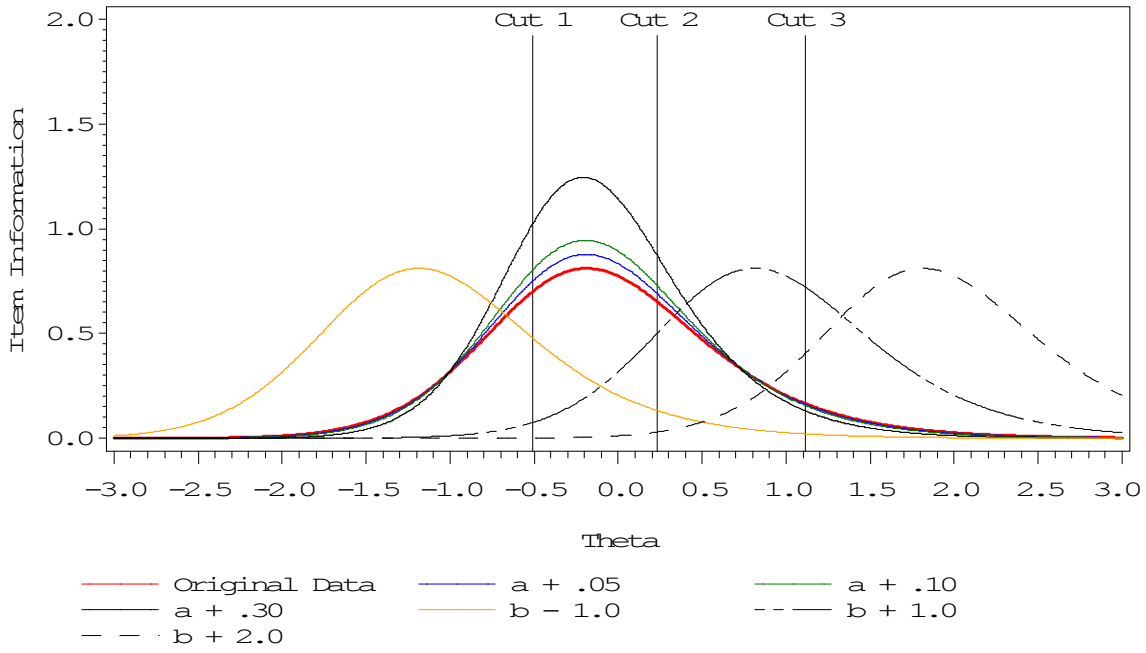
MC 27

Original parameters: a = 1.44, b = .59, c = .05



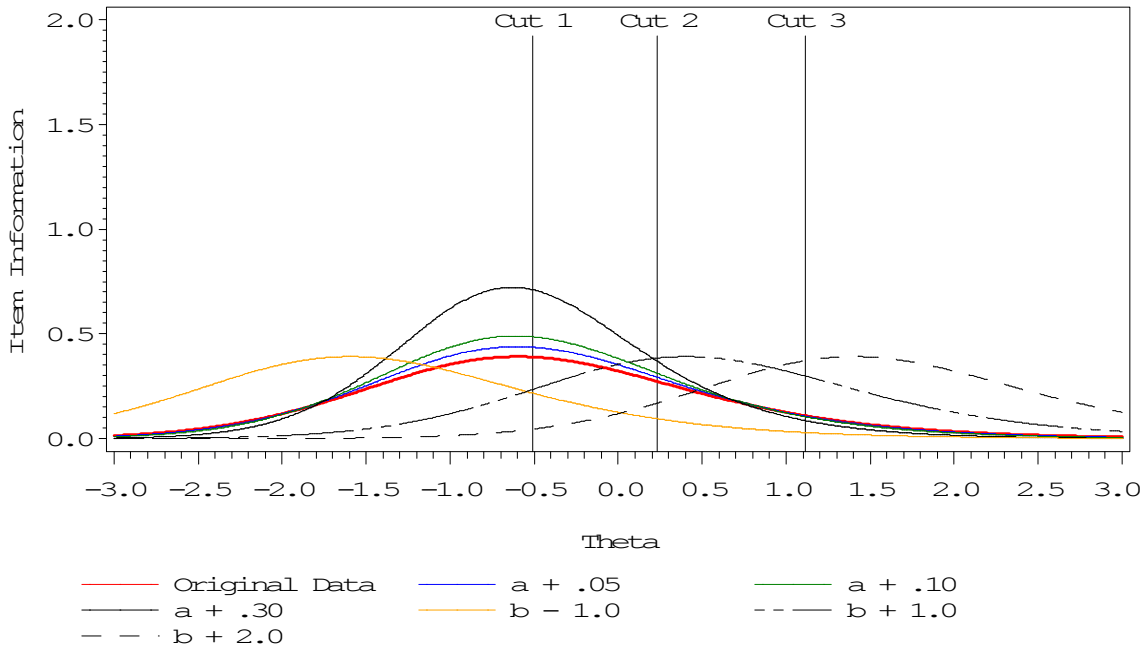
MC 28

Original parameters: a = 1.25, b = −.30, c = .17



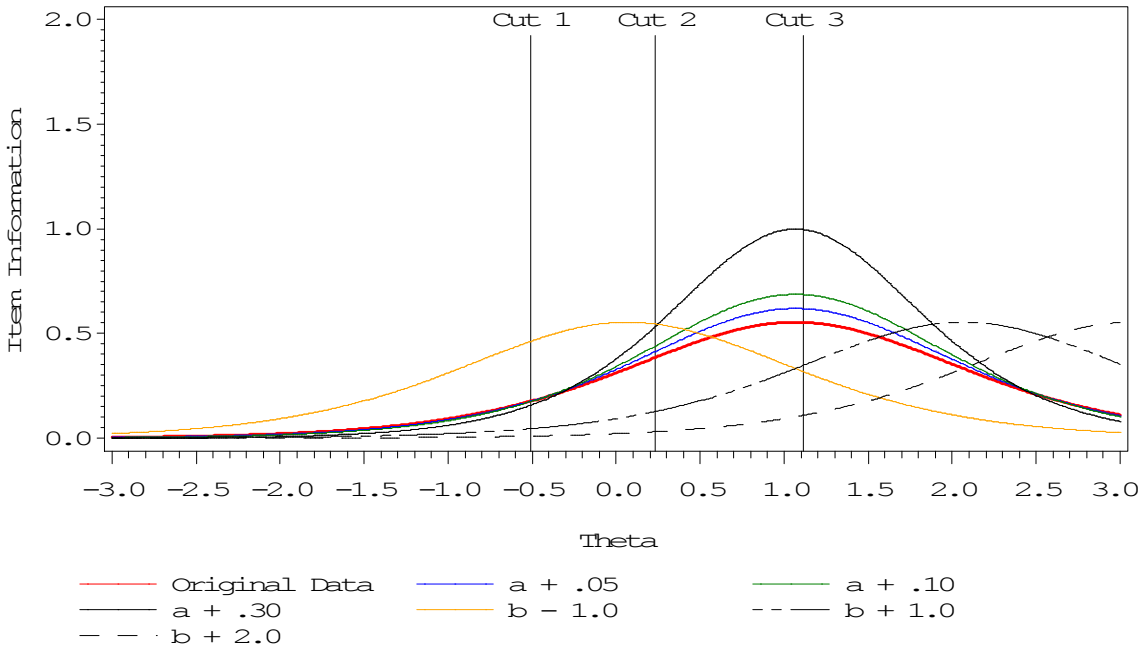
MC 29

Original parameters: a = .83, b = −.73, c = .13



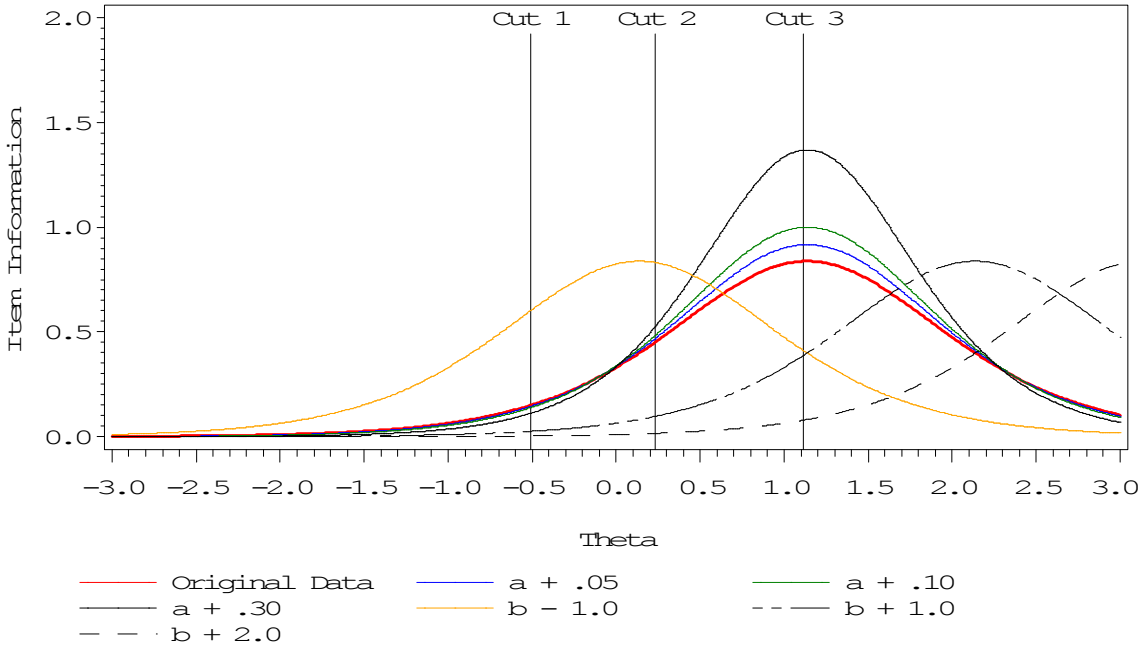
SA 1

Original parameters: a = .88, b = 1.07



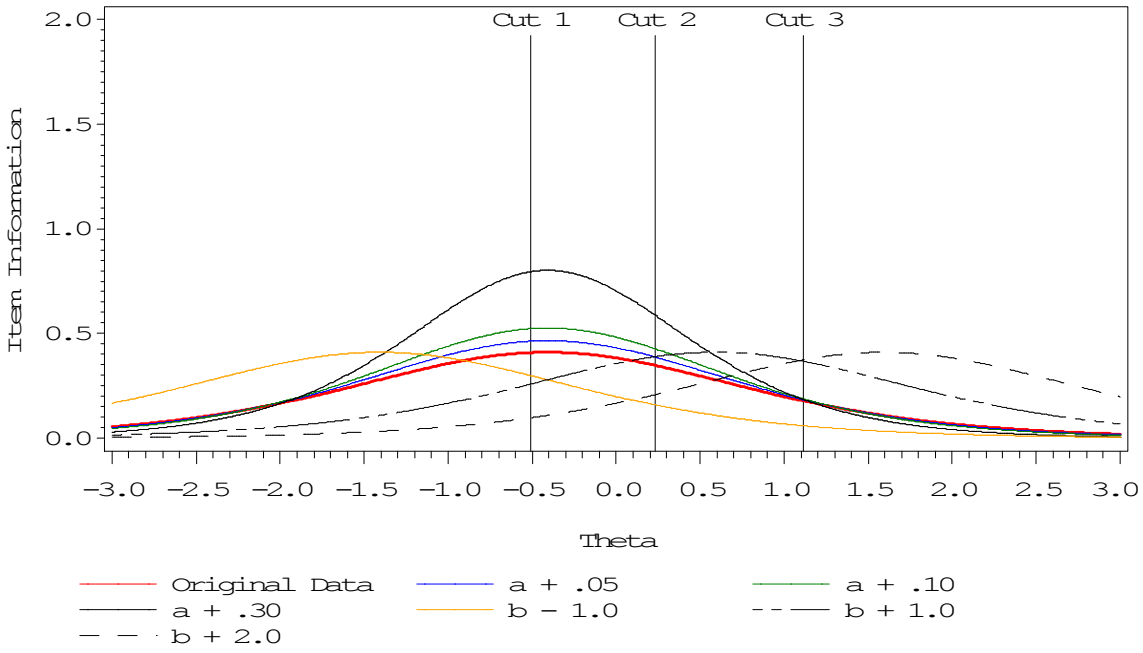
SA 2

Original parameters: a = 1.08, b = 1.14



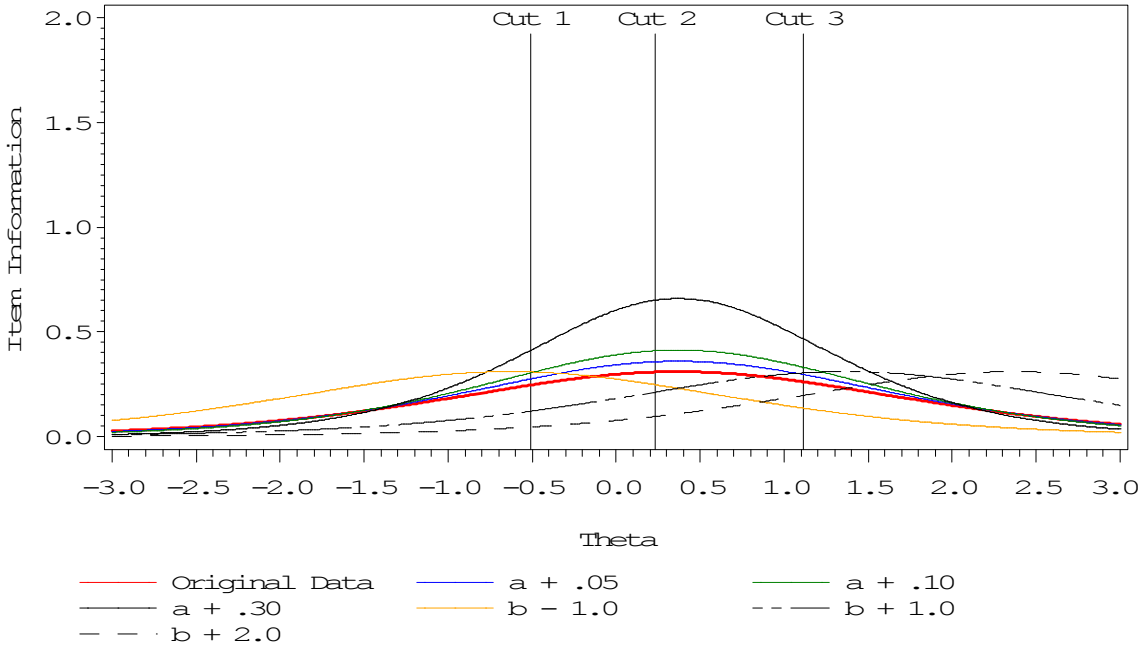
SA 3

Original parameters: a = .75, b = −.41



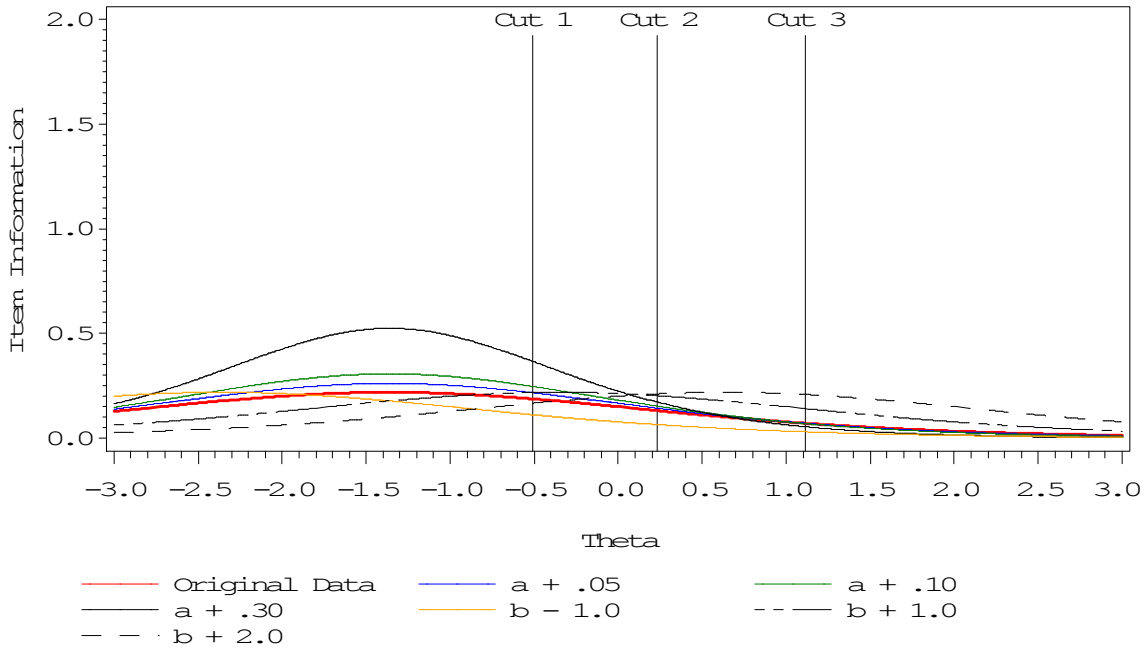
SA 4

Original parameters: a = .66, b = .37



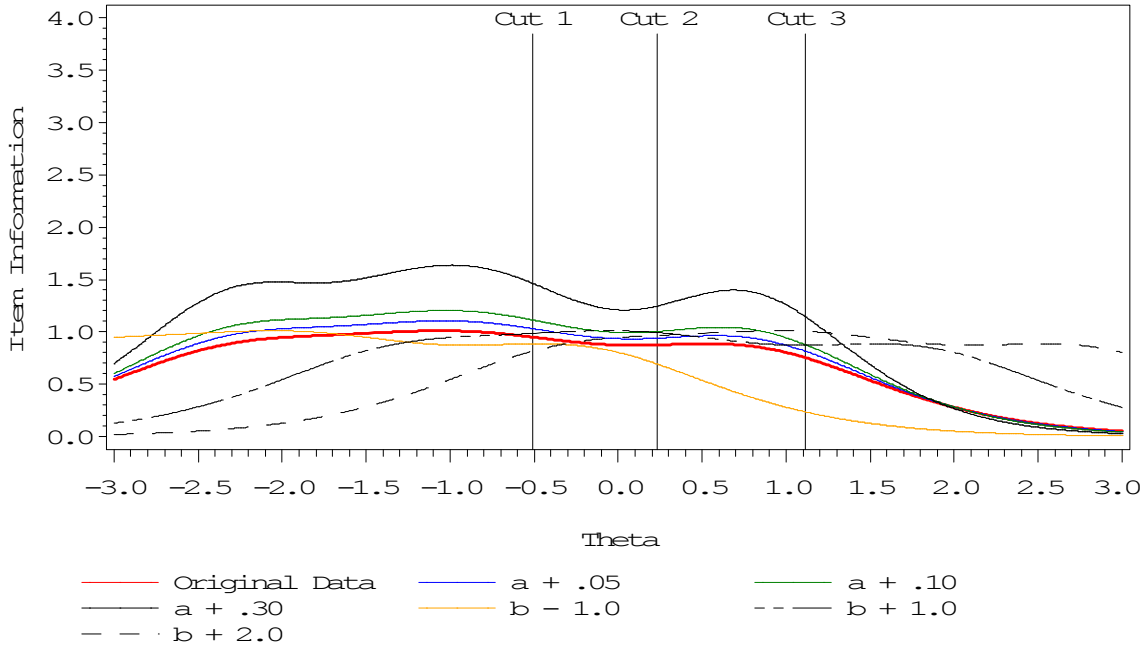
SA 5

Original parameters: a = .55, b = −1.36



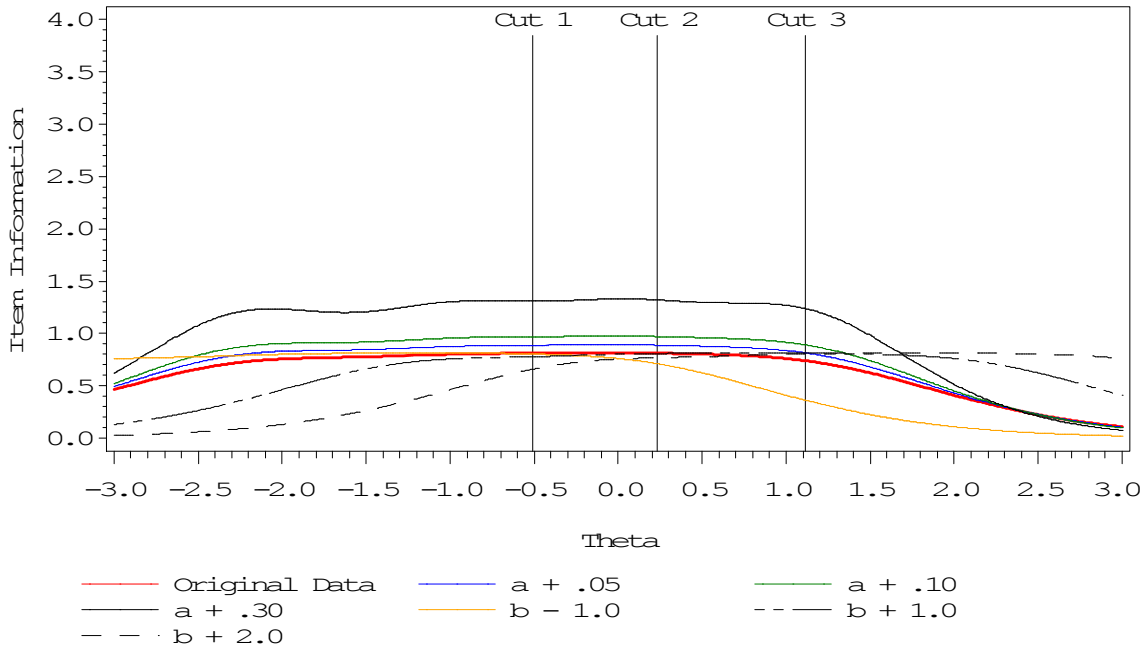
CR 1

Original parameters: a = 1.07, b1 = −2.26, b2 = −1.19, b3 = −.70, b4 = .75



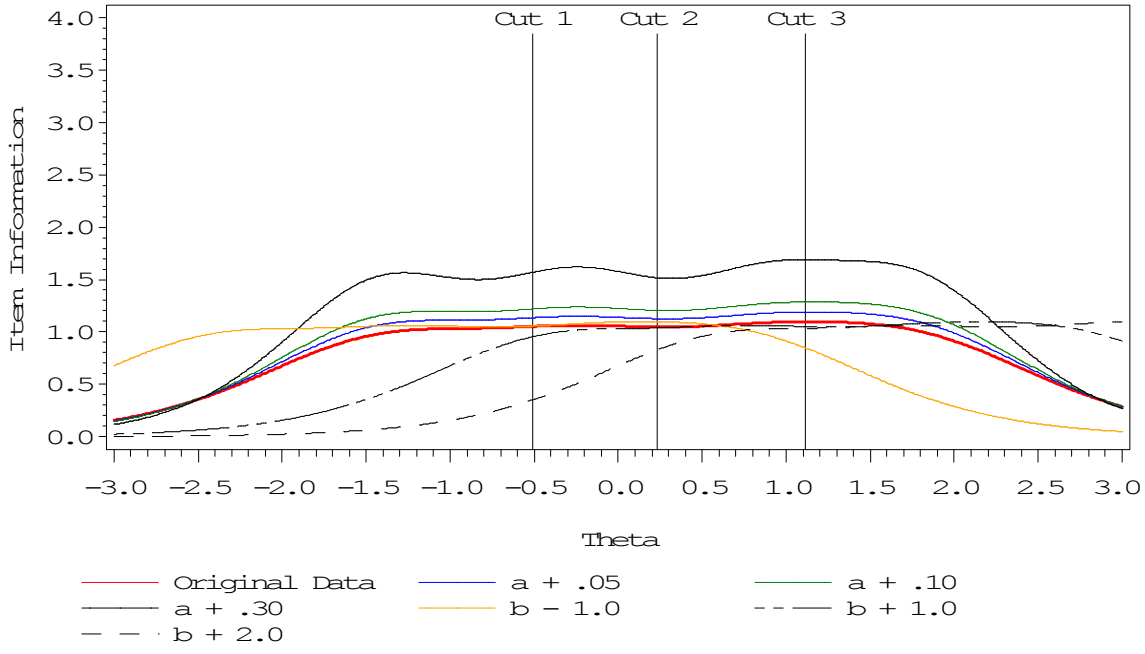
CR 2

Original parameters: a = .96, b1 = −2.23, b2 = −.97, b3 = .05, b4 = 1.10



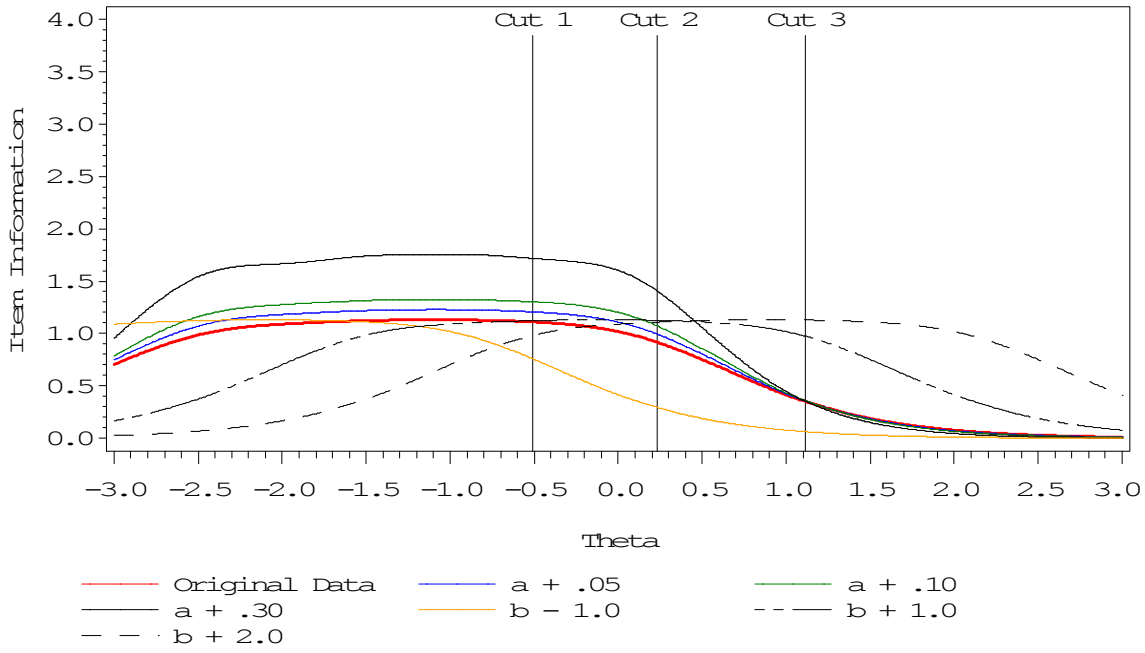
CR 3

Original parameters: a = 1.13, b1 = −1.40, b2 = −.25, b3 = .90, b4 = 1.76



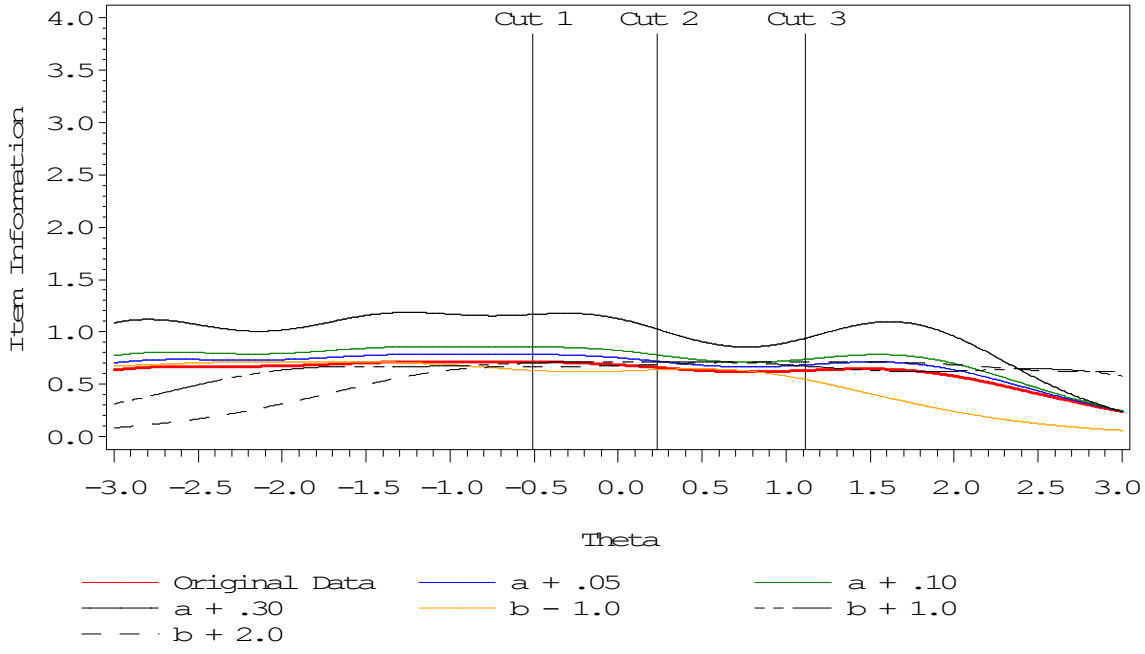
CR 4

Original parameters: a = 1.12, b1 = −2.44, b2 = −1.54, b3 = −.81, b4 = .01



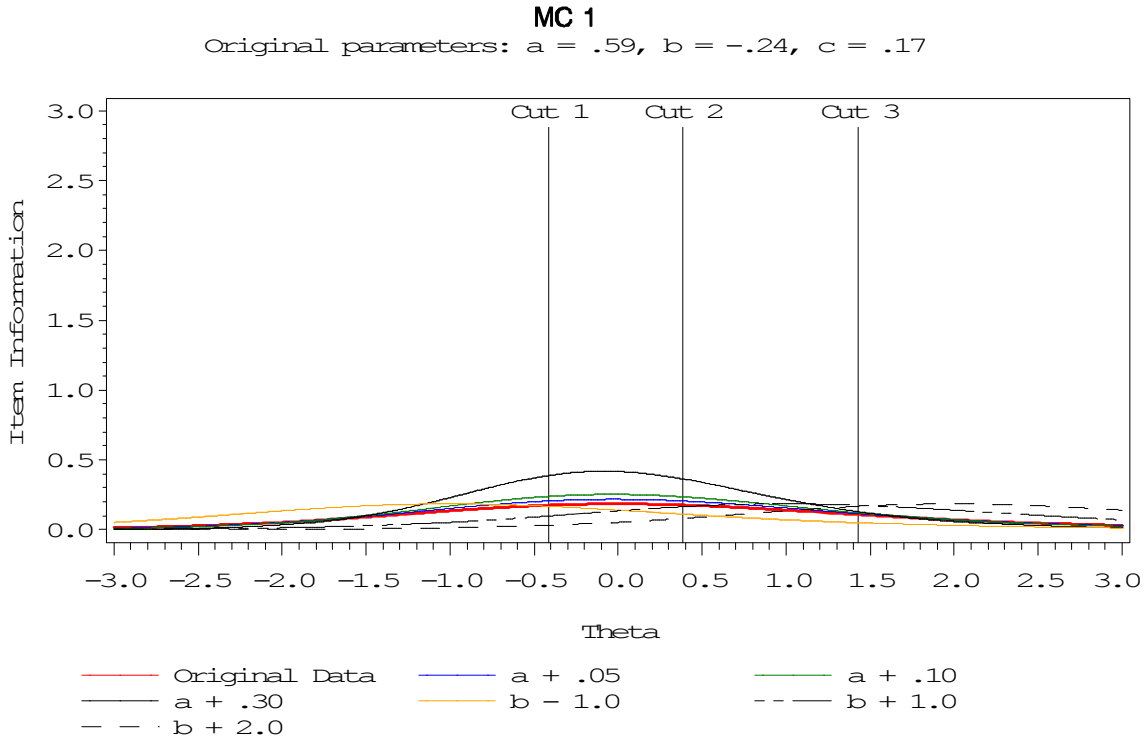
CR 5

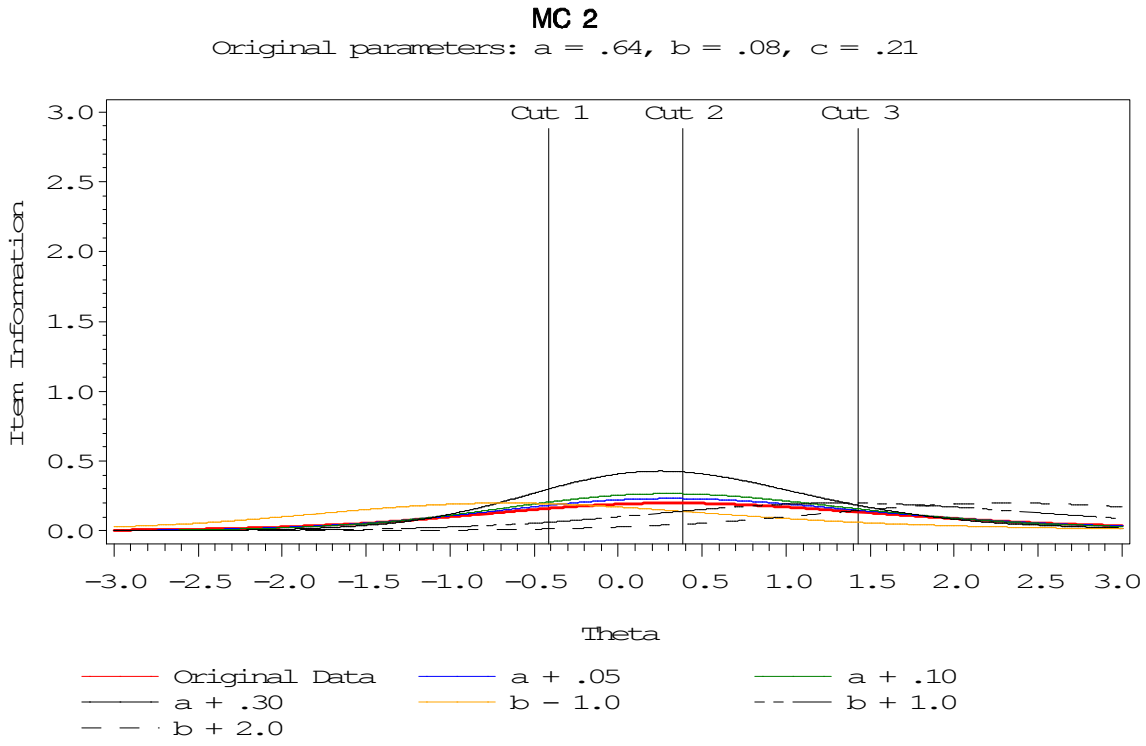
Original parameters: a = .92, b1 = −2.89, b2 = −1.36, b3 = −.15, b4 = 1.66

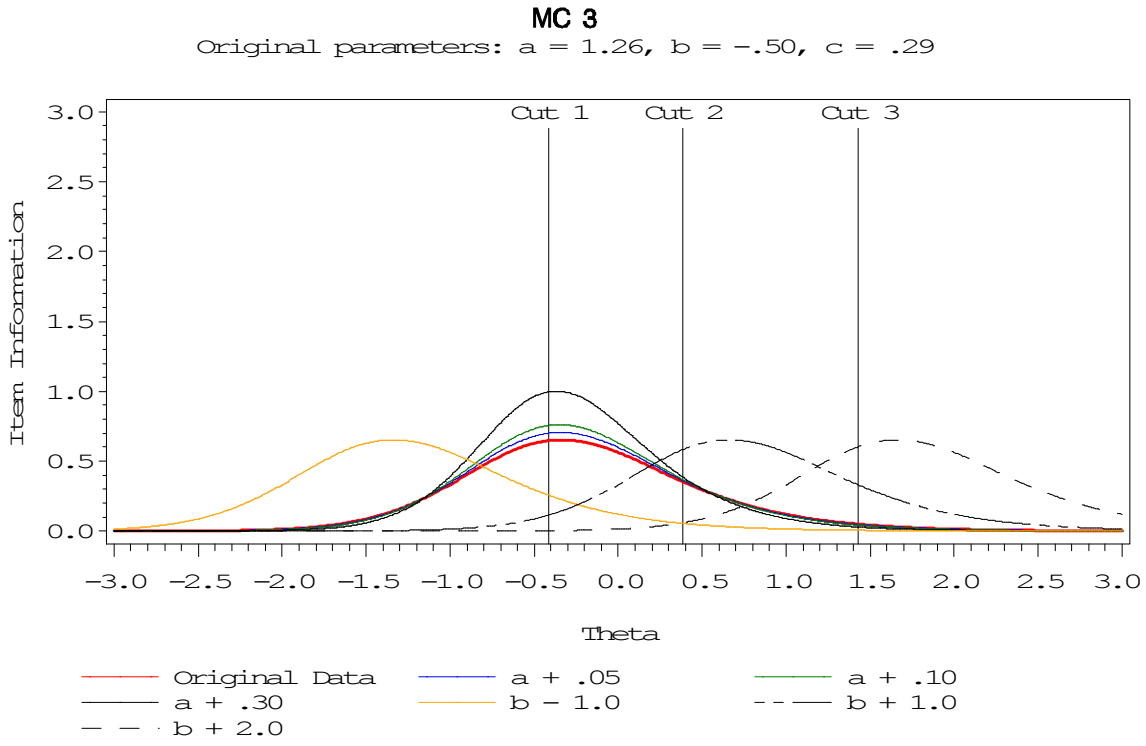


APPENDIX B.

ITEM INFORMATION FUNCTIONS FOR A HIGH SCHOOL ENGLISH LANGUAGE ARTS TEST

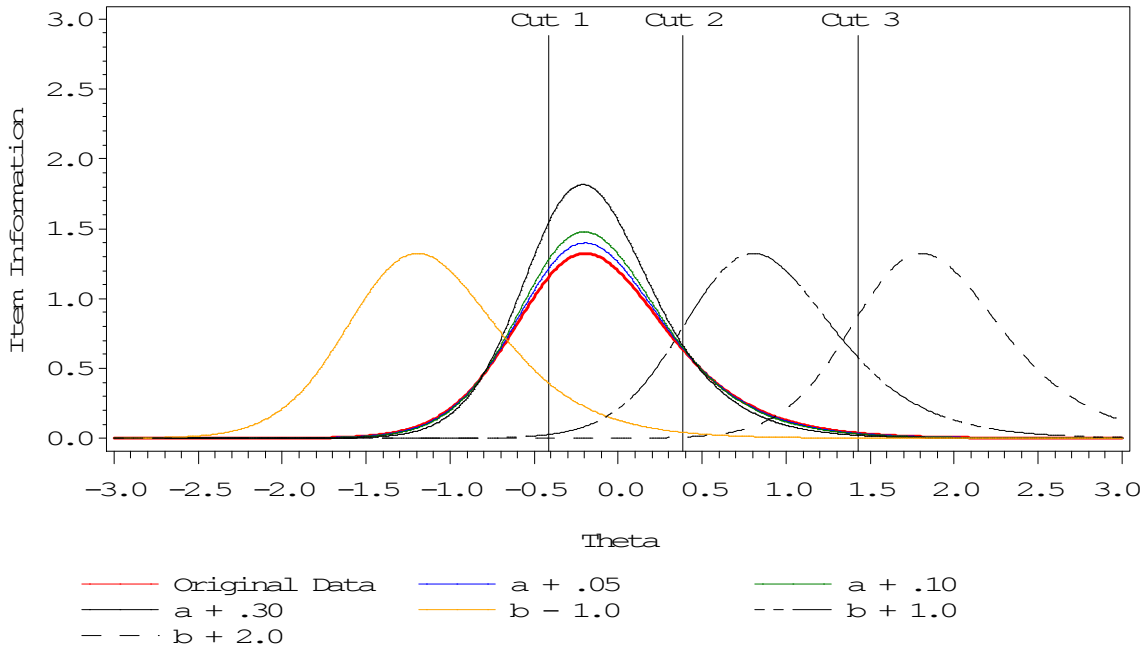



MC 2
Original parameters: a = .64, b = .08, c = .21
Cut 1
Cut 2
Cut 3
Item Information
Theta
Original Data
a + .05
a + .10
a + .30
b − 1.0
b + 1.0
b + 2.0


MC 3
Original parameters: a = 1.26, b = −.50, c = .29
Cut 1
Cut 2
Cut 3
Item Information
Theta
Original Data
a + .05
a + .10
a + .30
b − 1.0
b + 1.0
b + 2.0

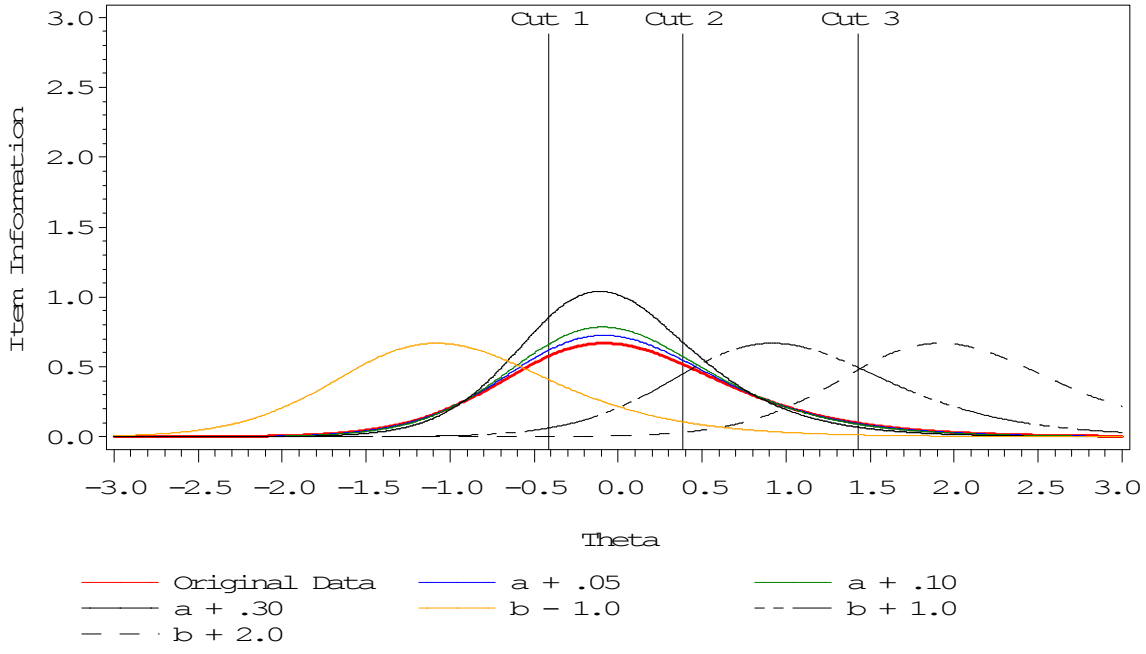
MC 4

Original parameters: a = 1.74, b = −.31, c = .26



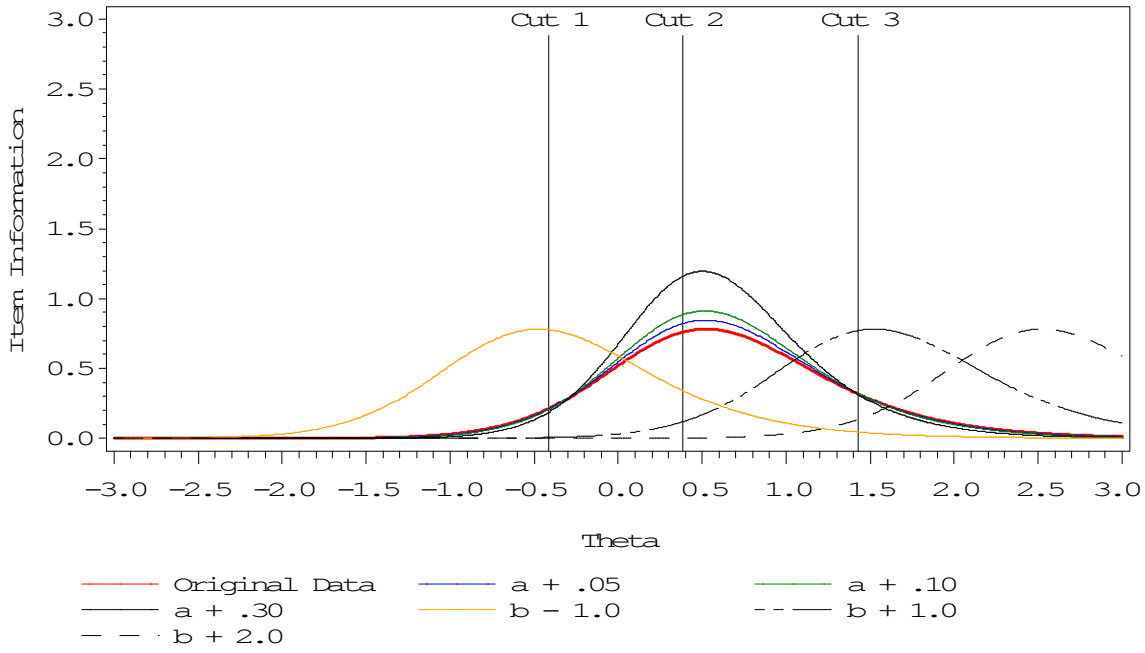
MC 5

Original parameters: a = 1.22, b = −.23, c = .24



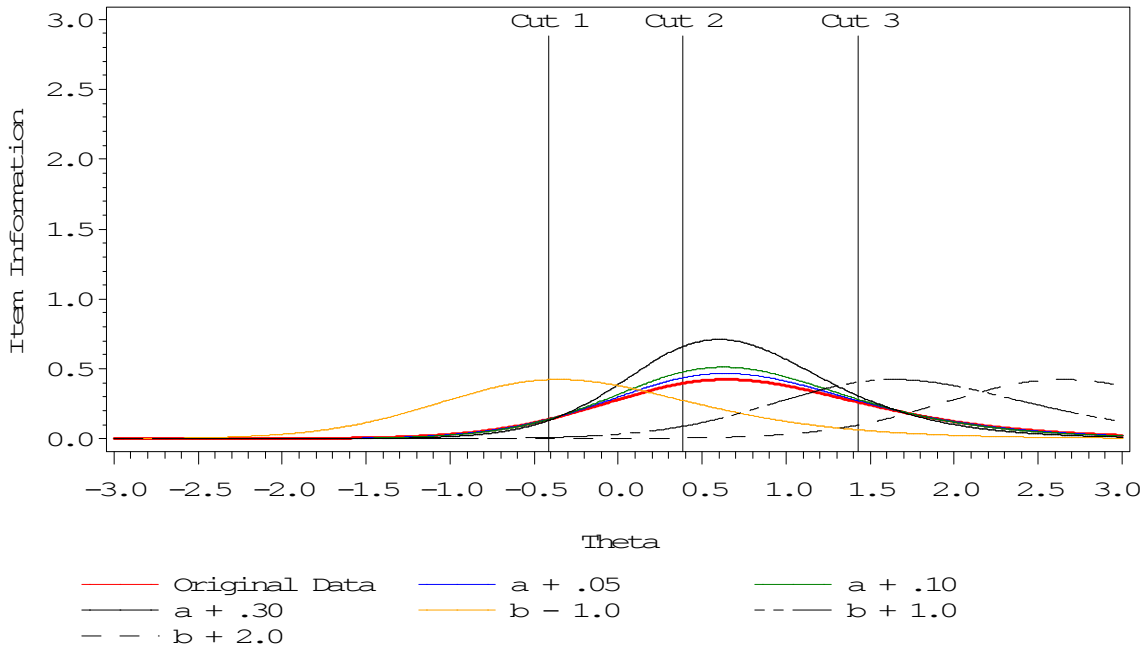
MC 6

Original parameters: a = 1.27, b = .40, c = .21



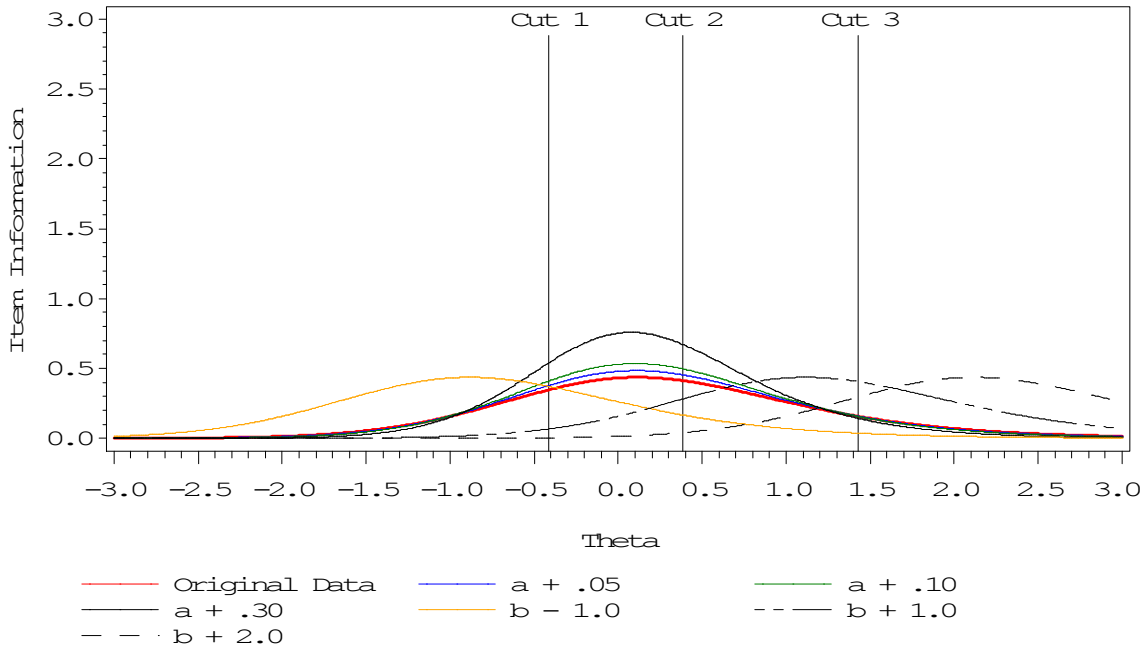
MC 7

Original parameters: a = 1.02, b = .44, c = .29



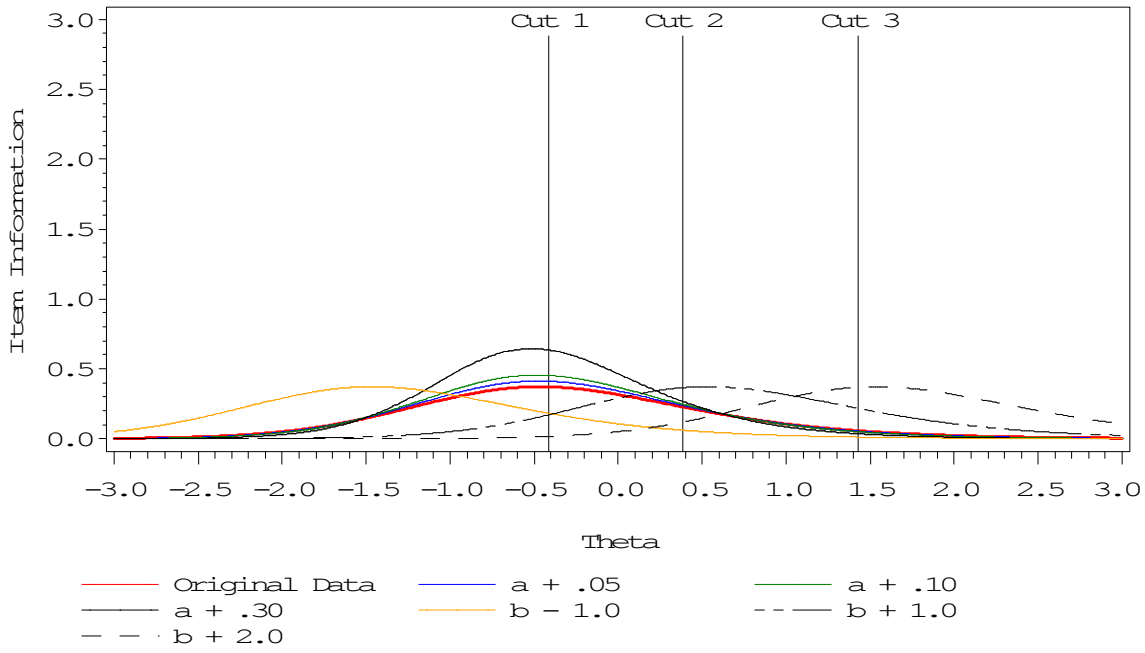
MC 8

Original parameters: a = .94, b = −.05, c = .20



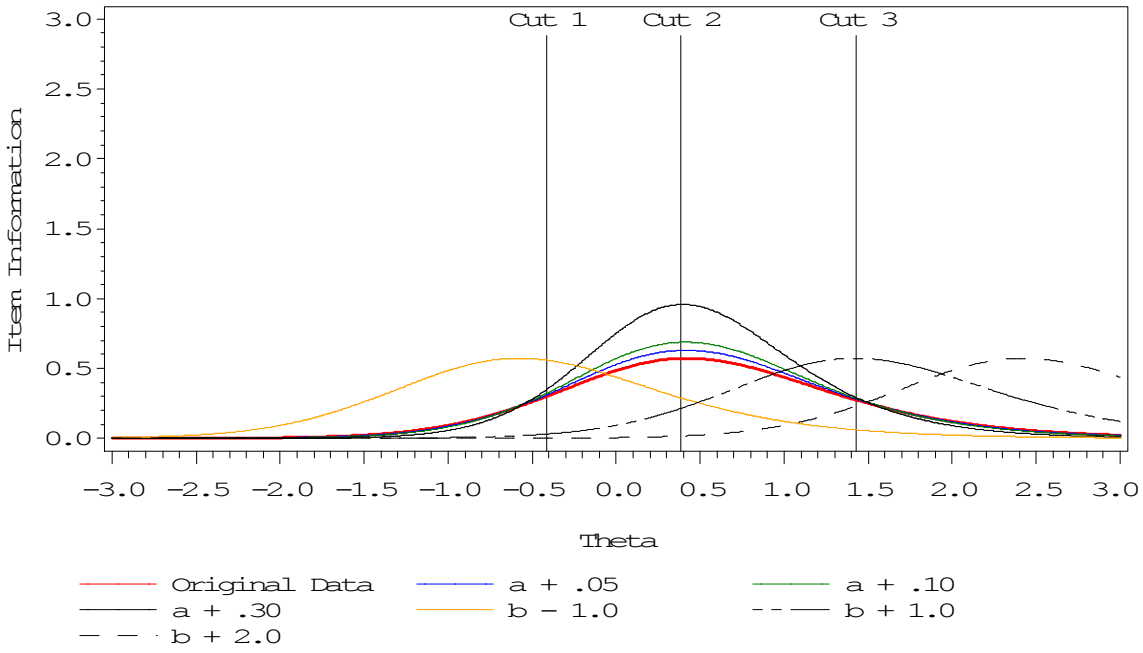
MC 9

Original parameters: a = .95, b = −.68, c = .29



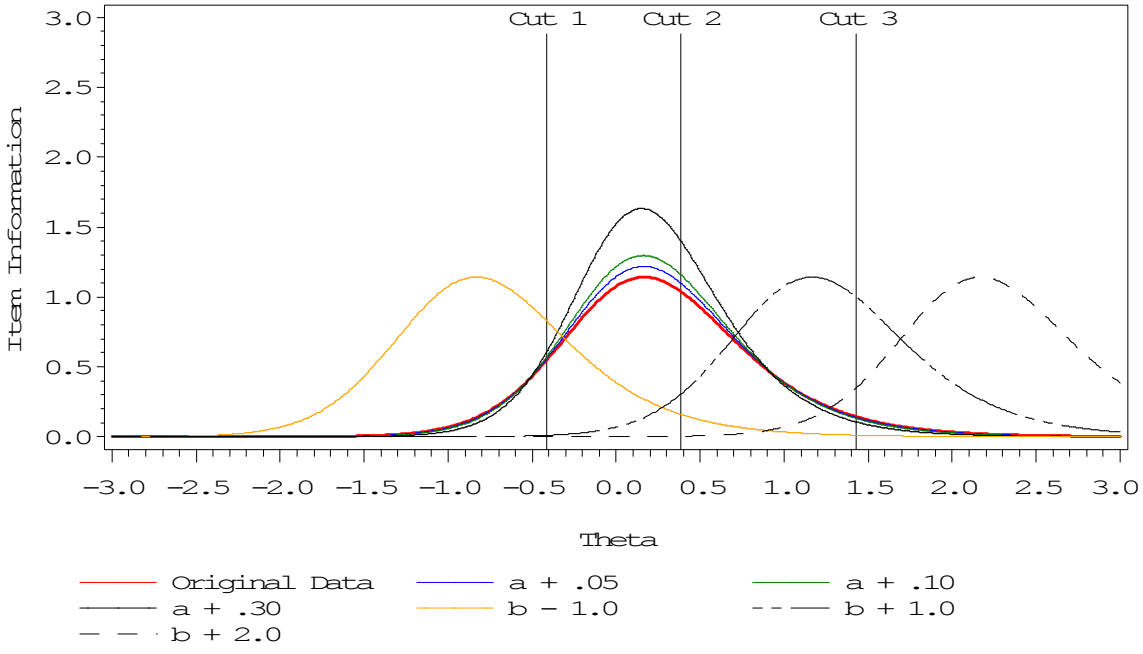
MC 10

Original parameters: a = 1.02, b = .30, c = .14



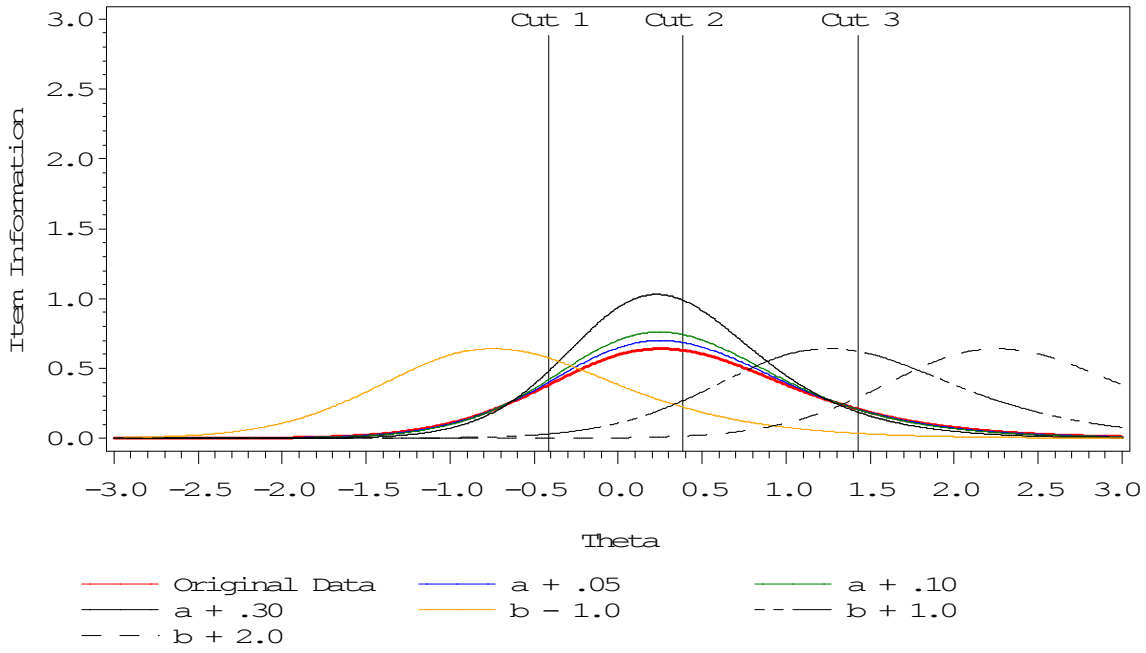
MC 11

Original parameters: a = 1.54, b = .06, c = .21



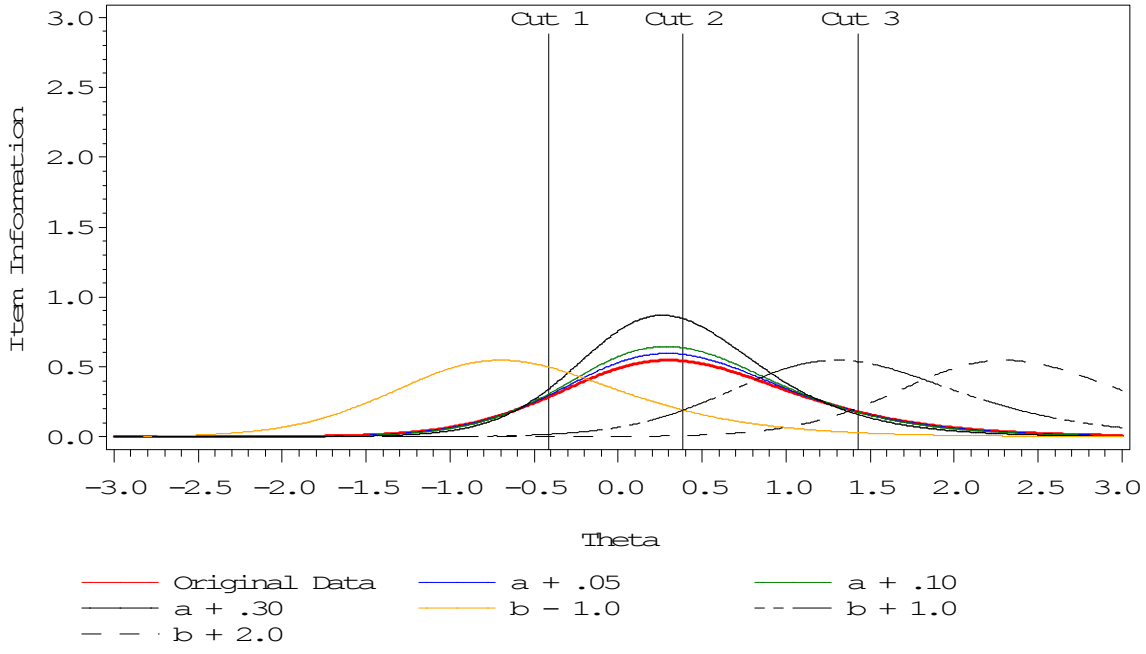
MC 12

Original parameters: a = 1.12, b = .12, c = .18



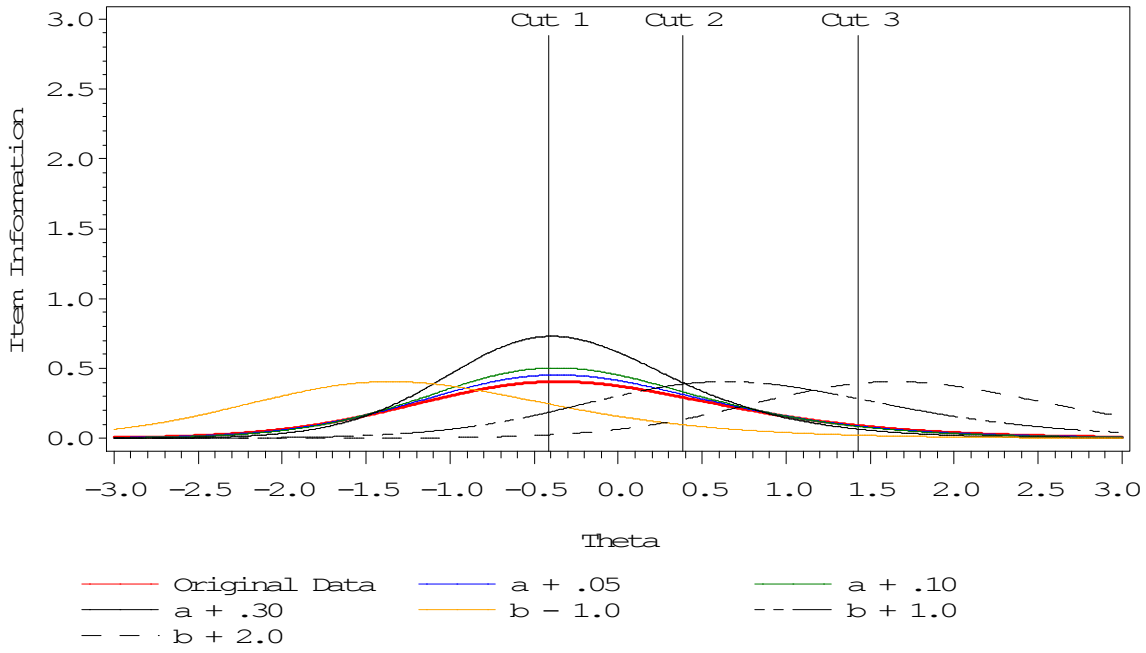
MC 13

Original parameters: a = 1.15, b = .12, c = .29



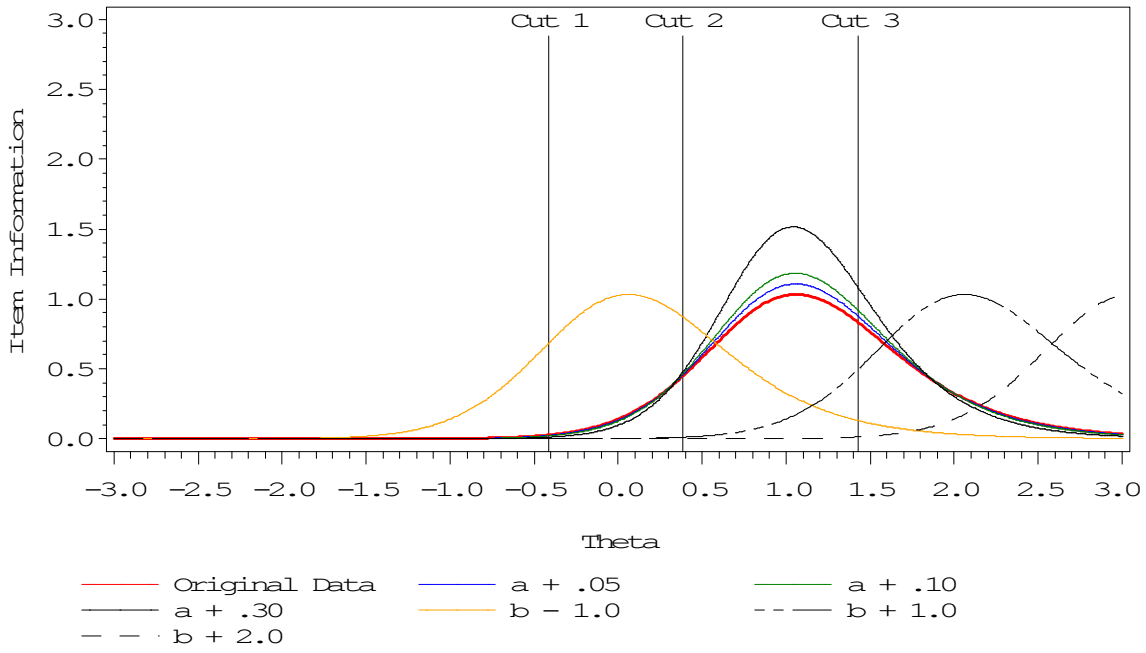
MC 14

Original parameters: a = .88, b = −.51, c = .17



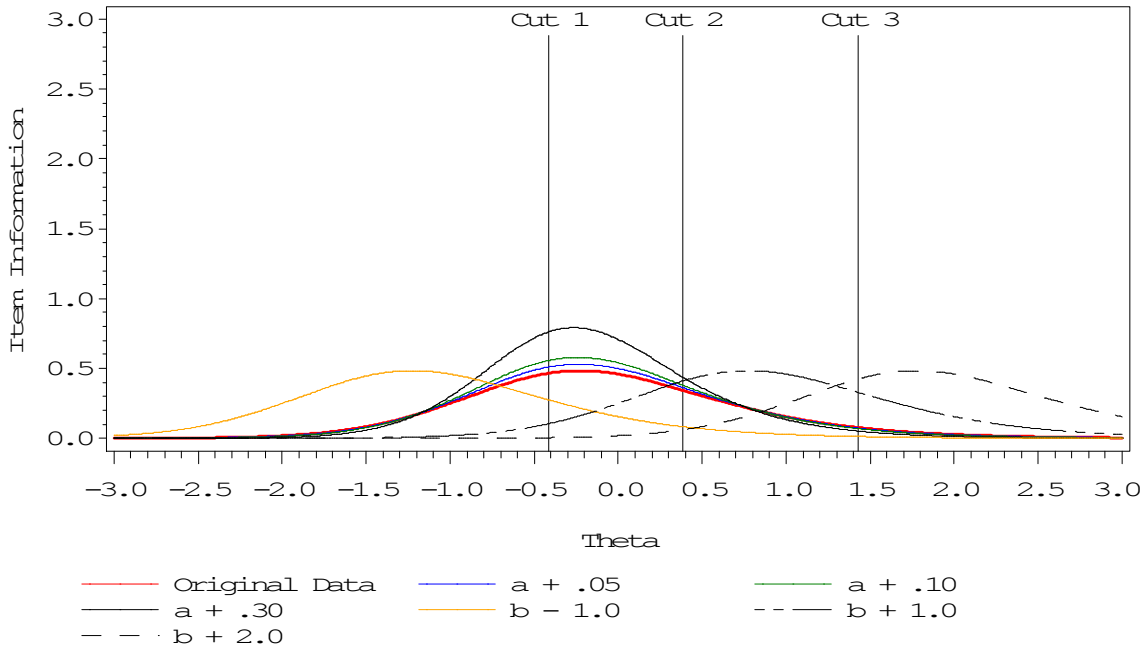
MC 15

Original parameters: a = 1.42, b = .96, c = .18



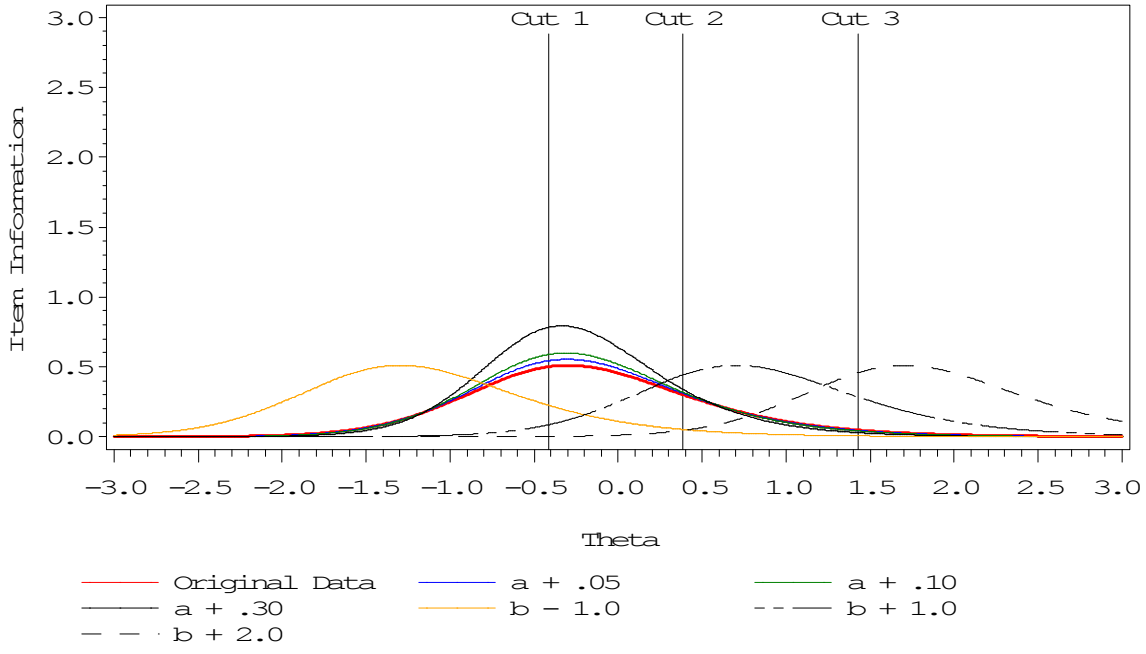
MC 16

Original parameters: a = 1.07, b = −.41, c = .28



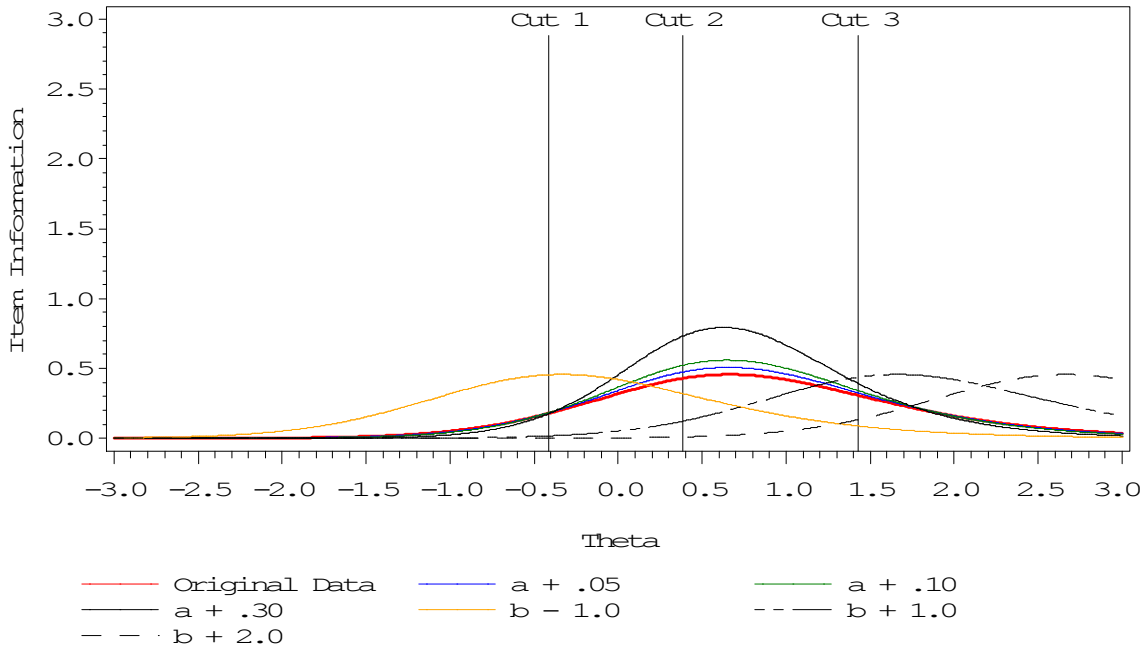
MC 17

Original parameters: a = 1.21, b = −.50, c = .38



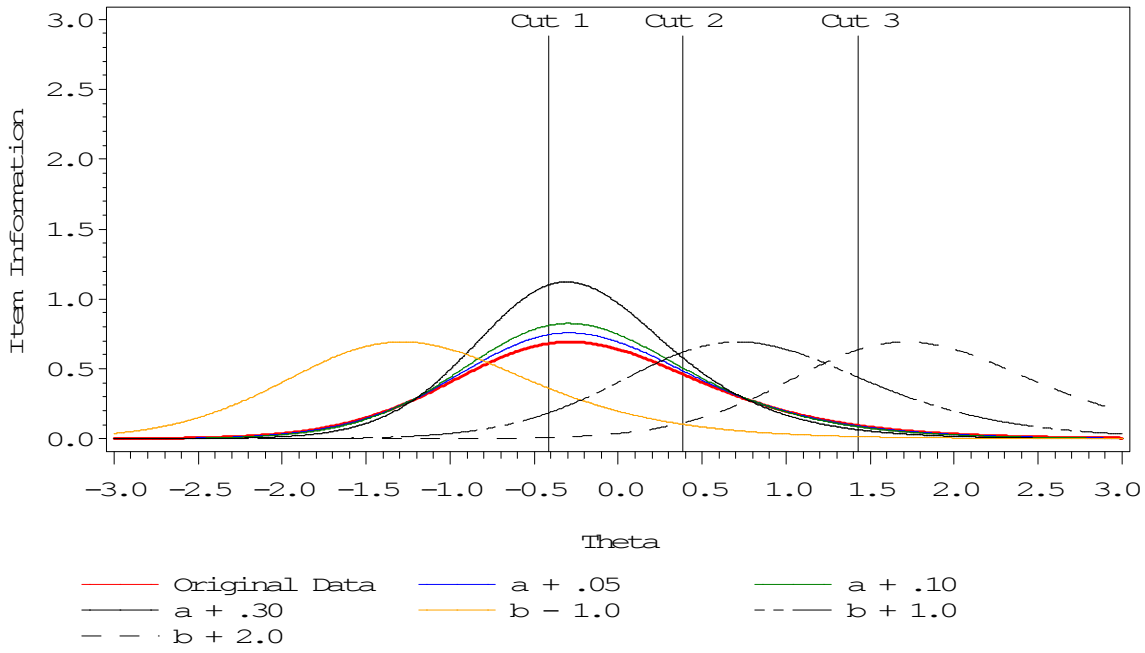
MC 18

Original parameters: a = .94, b = .51, c = .18



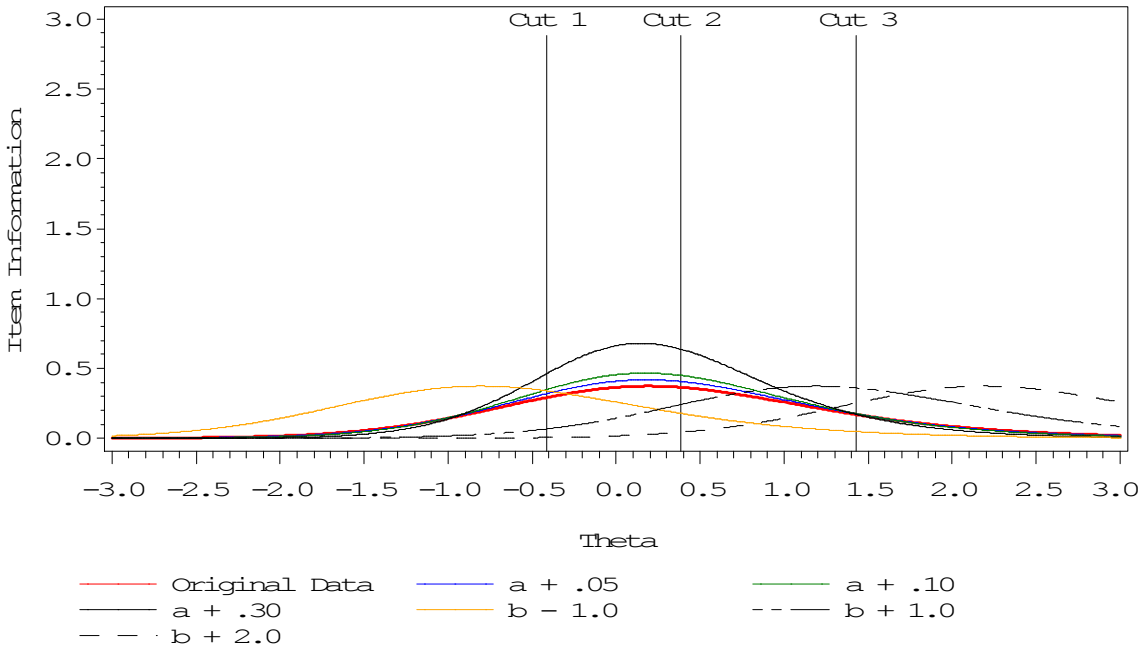
MC 19

Original parameters: a = 1.10, b = −.39, c = .12



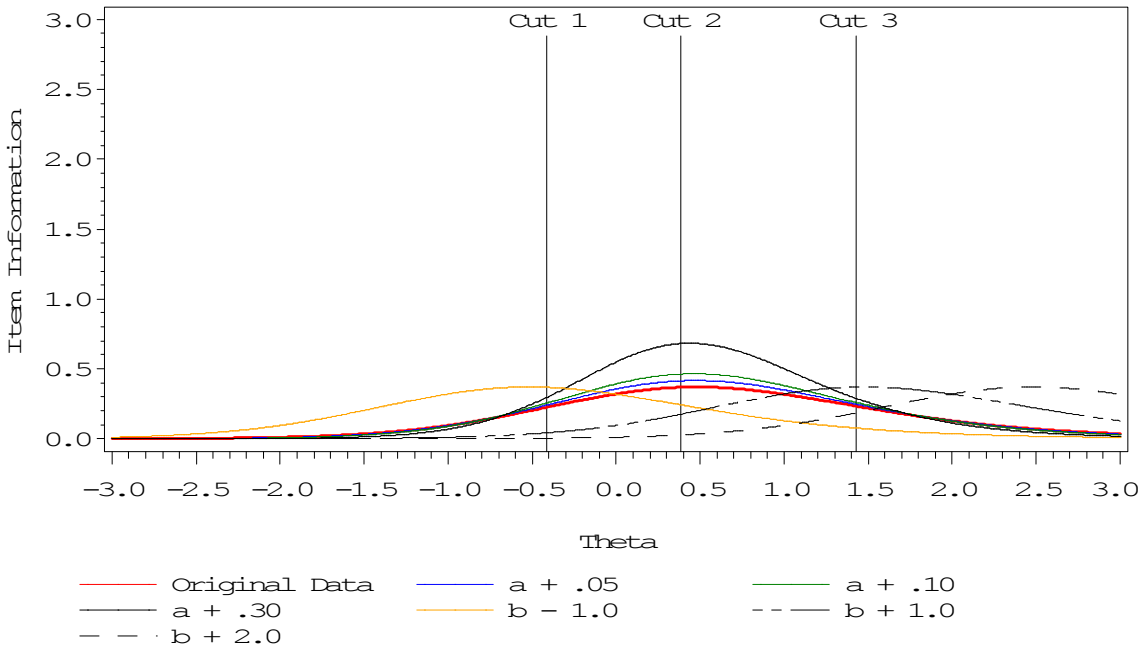
MC 20

Original parameters: a = .86, b = .02, c = .19



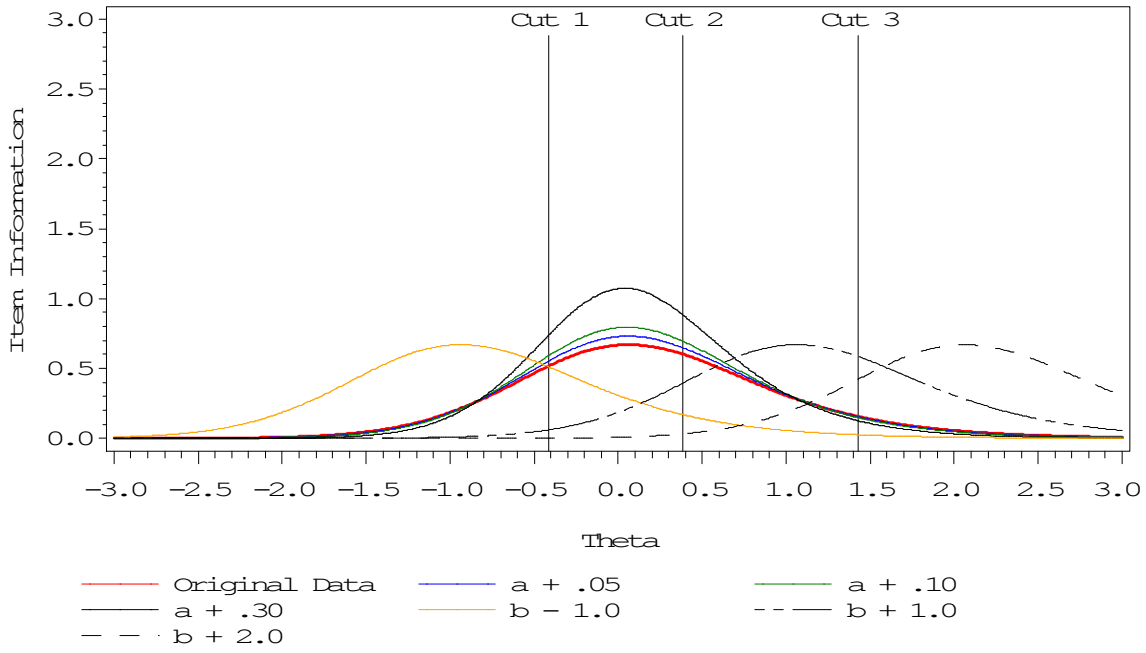
MC 21

Original parameters: a = .84, b = .32, c = .16



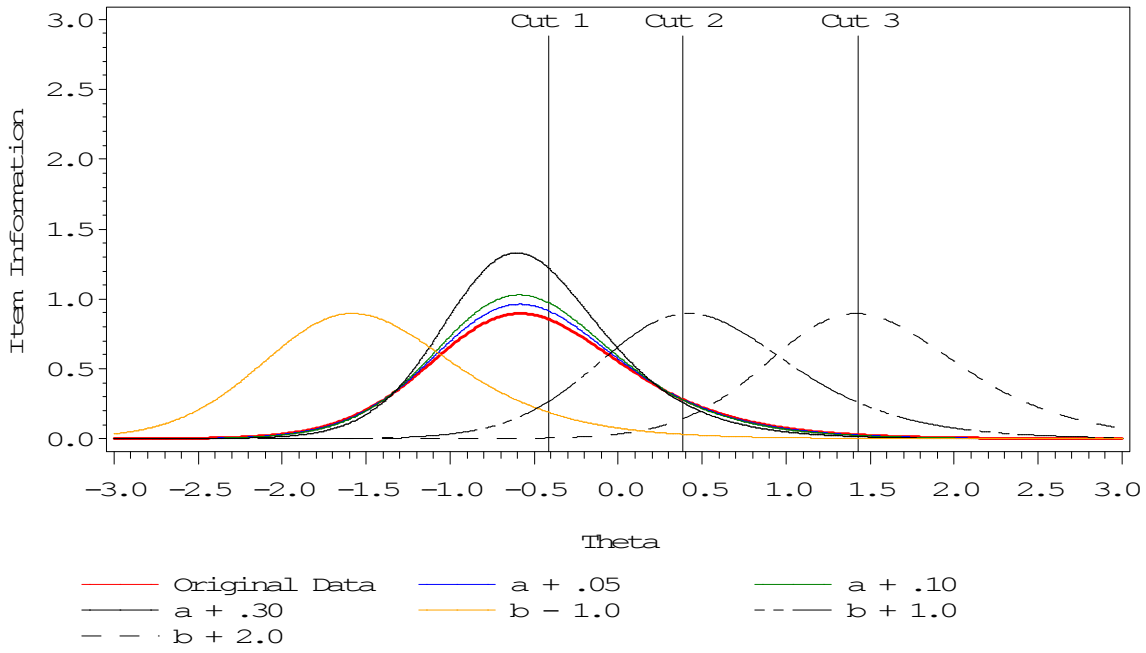
MC 22

Original parameters: a = 1.13, b = −.06, c = .17



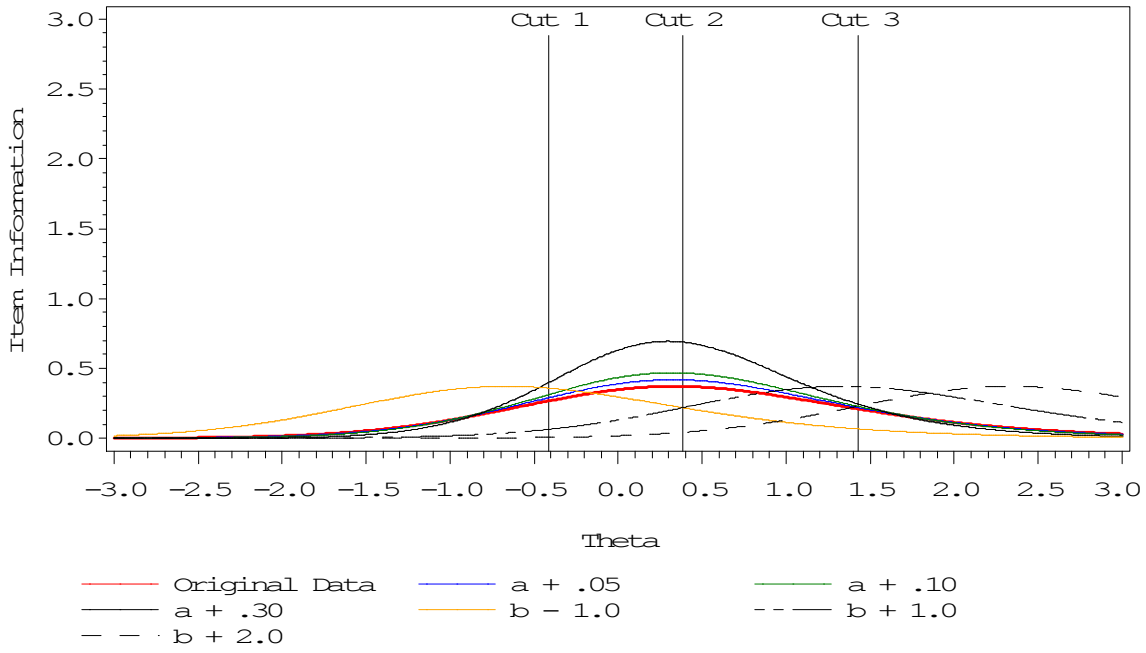
MC 23

Original parameters: a = 1.37, b = −.71, c = .22



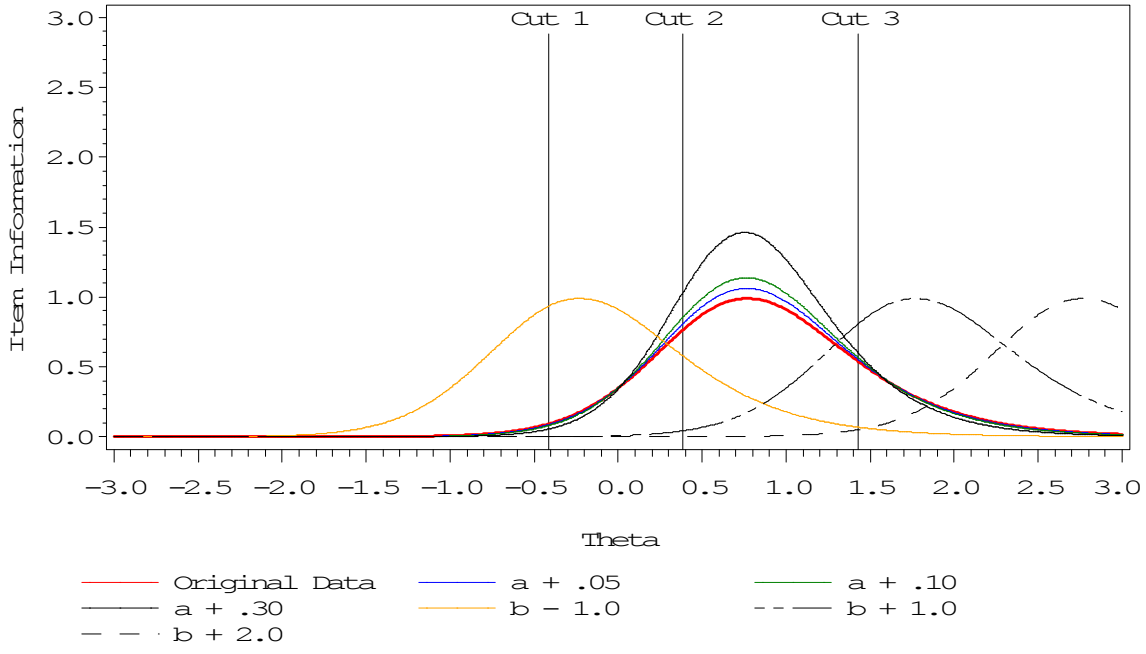
MC 24

Original parameters: a = .82, b = .19, c = .14



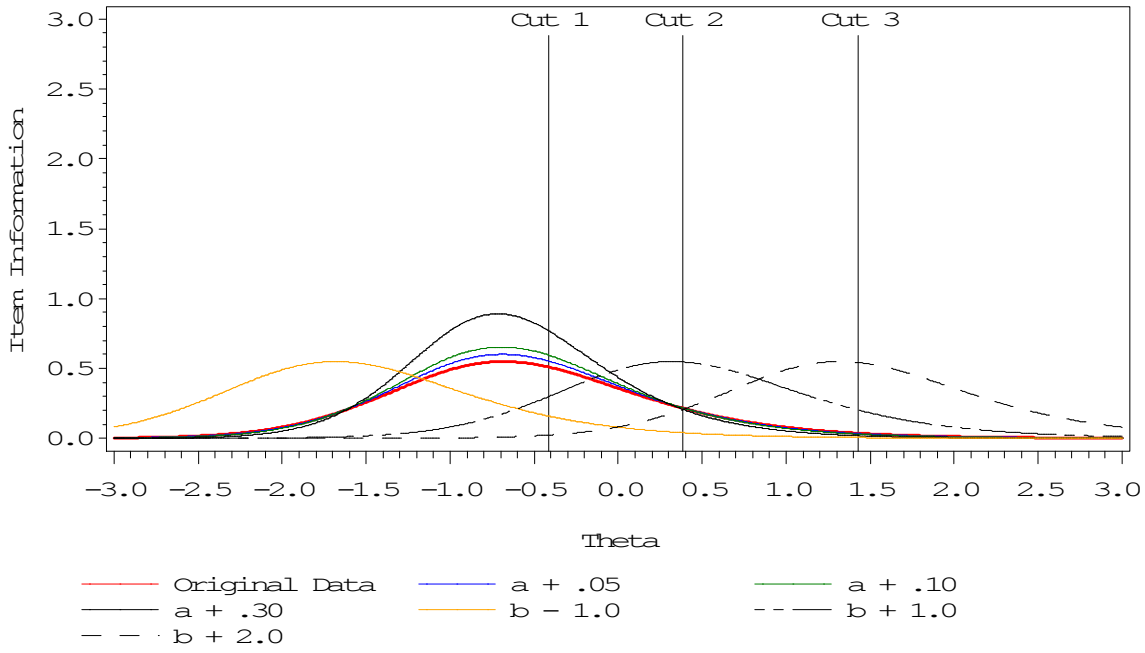
MC 25

Original parameters: a = 1.39, b = .66, c = .18



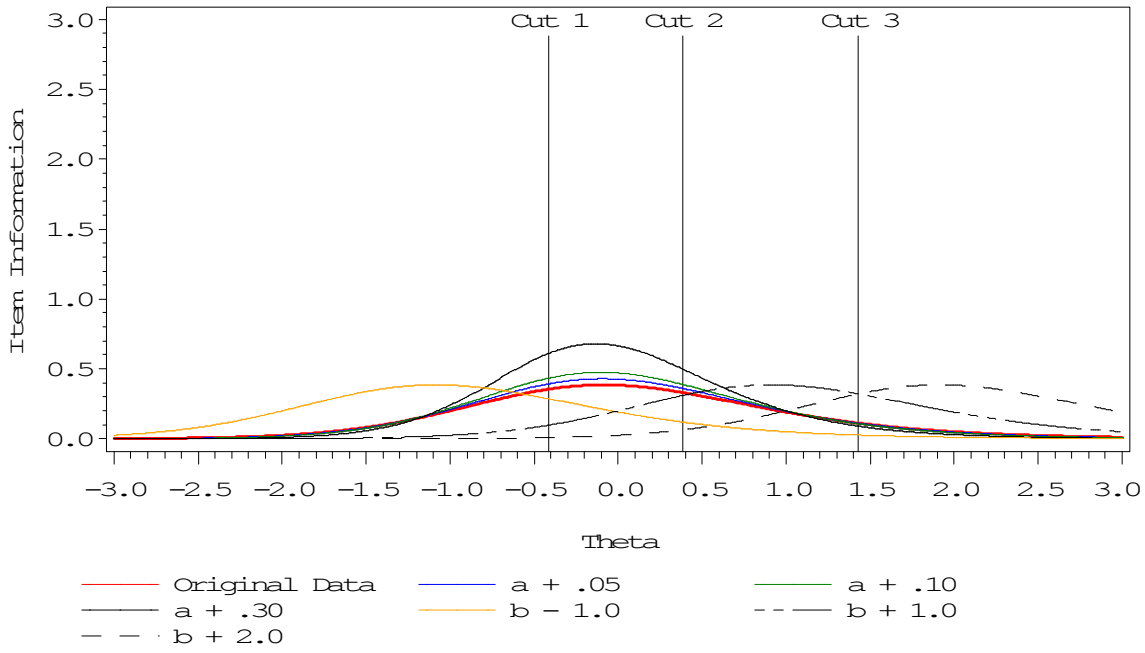
MC 26

Original parameters: a = 1.09, b = −.84, c = .24



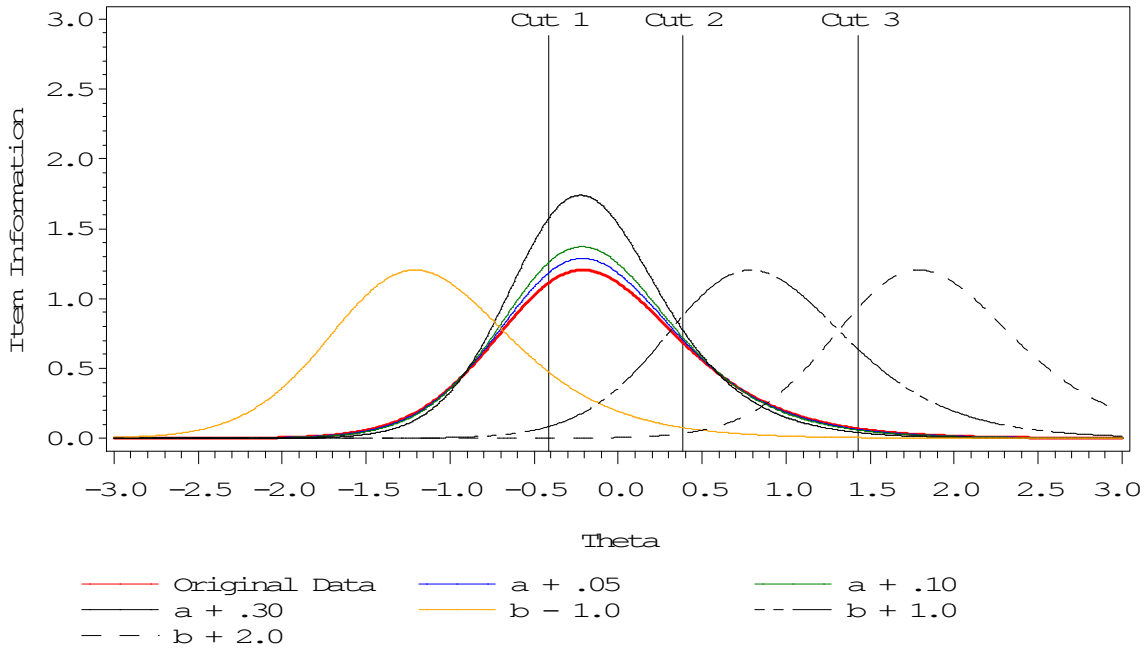
MC 27

Original parameters: a = .92, b = −.28, c = .24



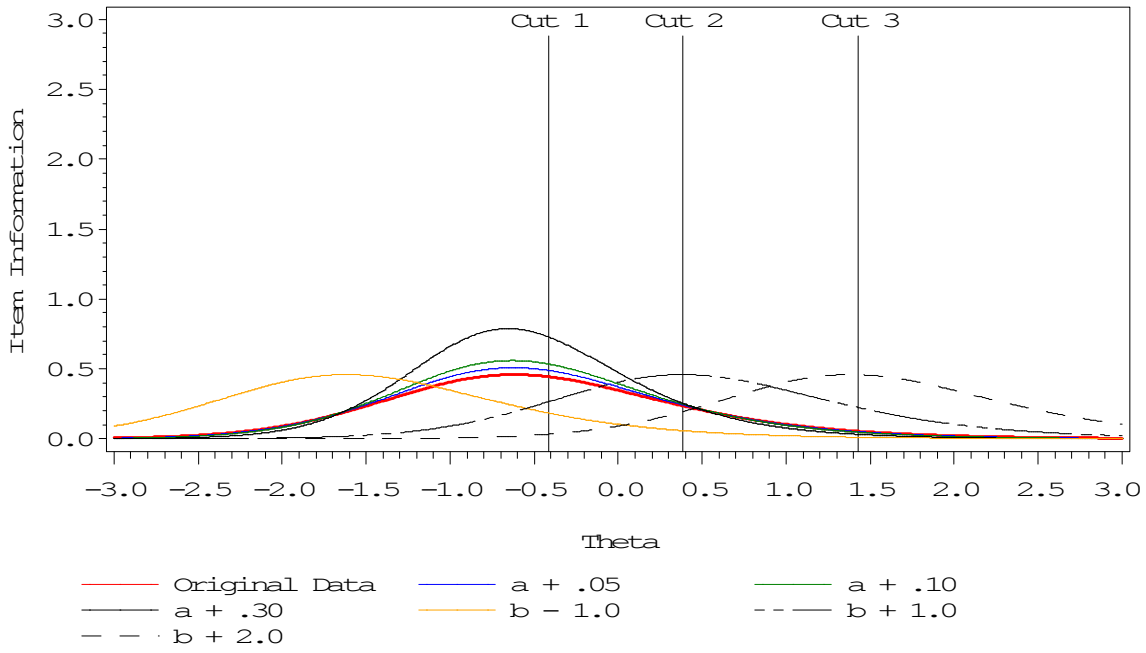
MC 28

Original parameters: a = 1.48, b = −.29, c = .14



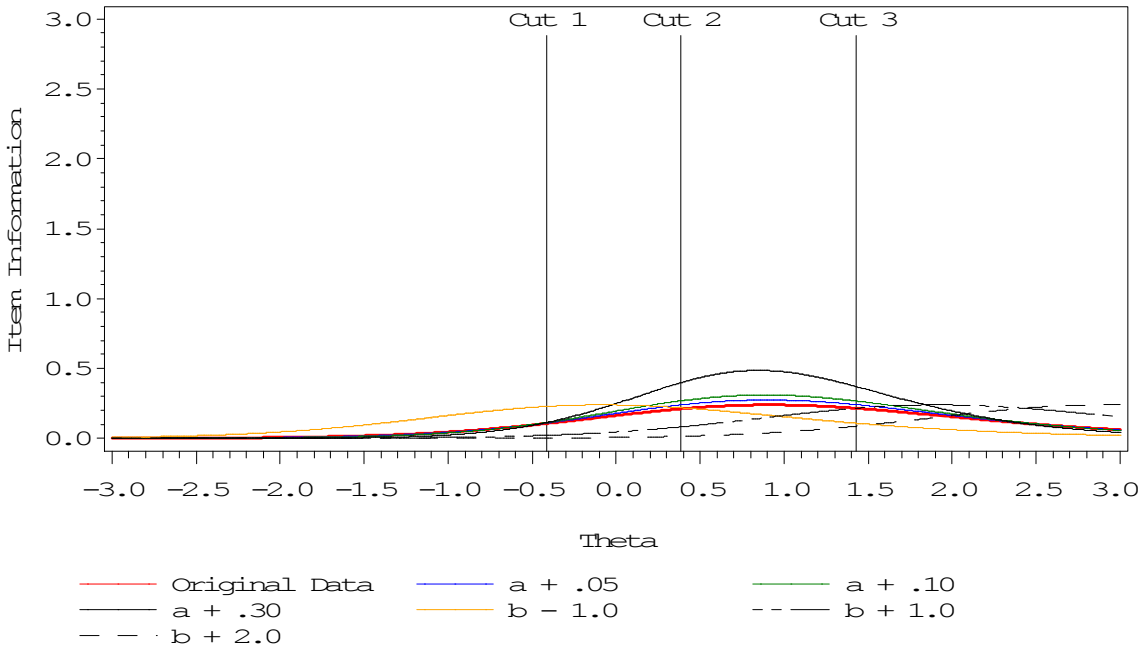
MC 29

Original parameters: a = .97, b = −.78, c = .20



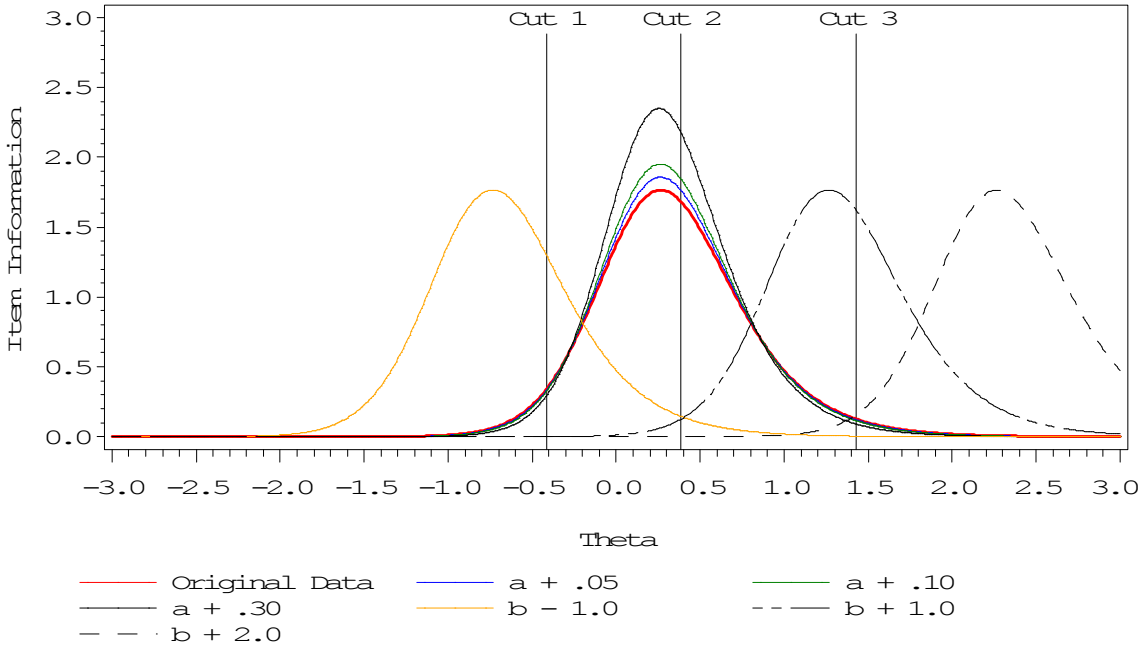
MC 30

Original parameters: a = .70, b = .68, c = .21



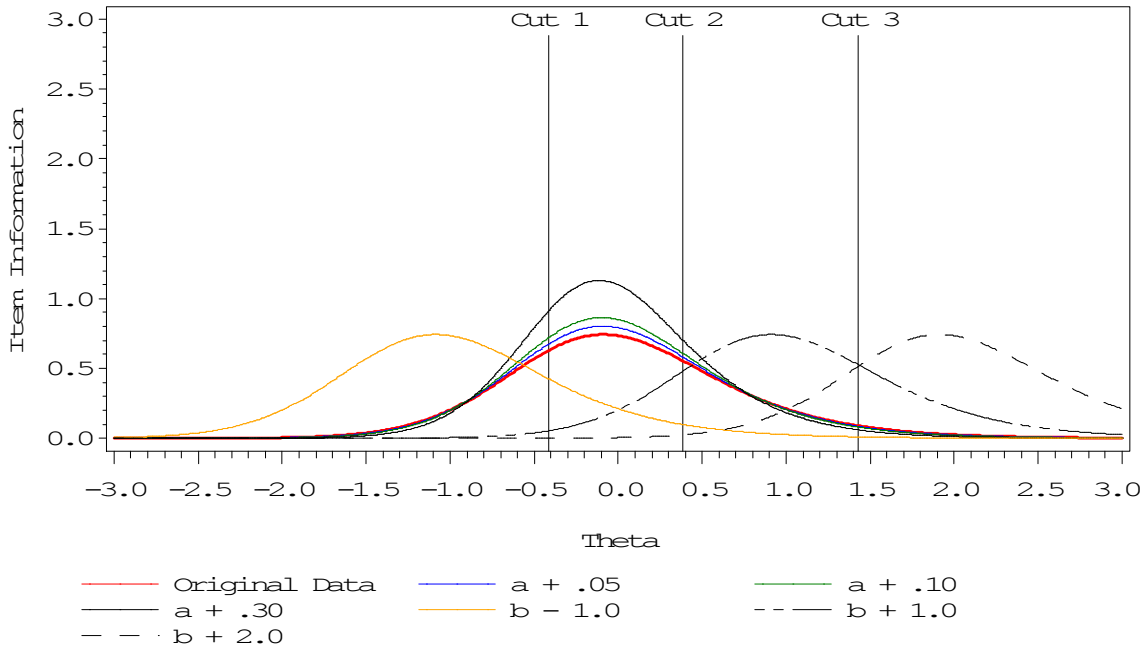
MC 31

Original parameters: a = 1.96, b = .18, c = .24



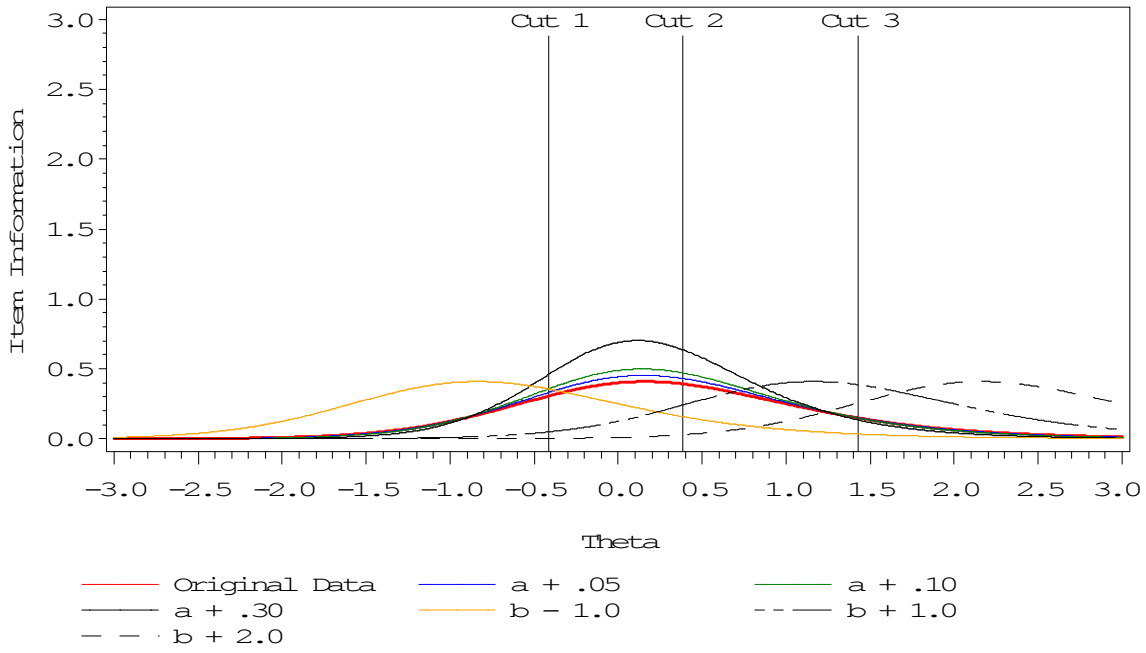
MC 32

Original parameters: a = 1.29, b = −.23, c = .25



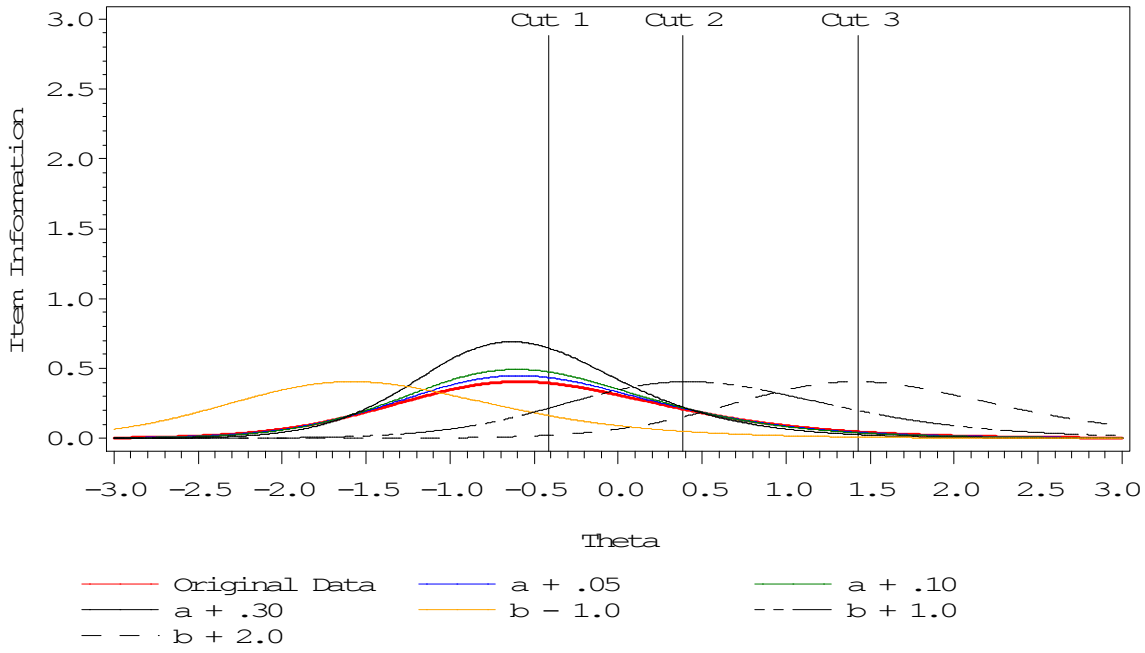
MC 33

Original parameters: a = .97, b = −.04, c = .26



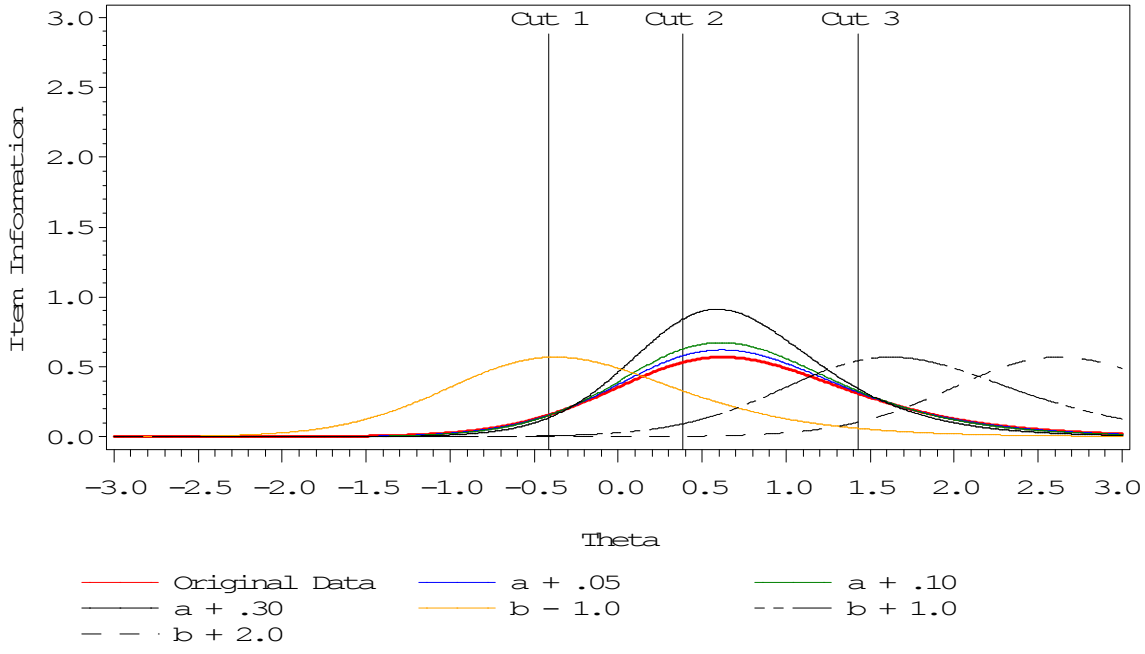
MC 34

Original parameters: a = .98, b = −.79, c = .28



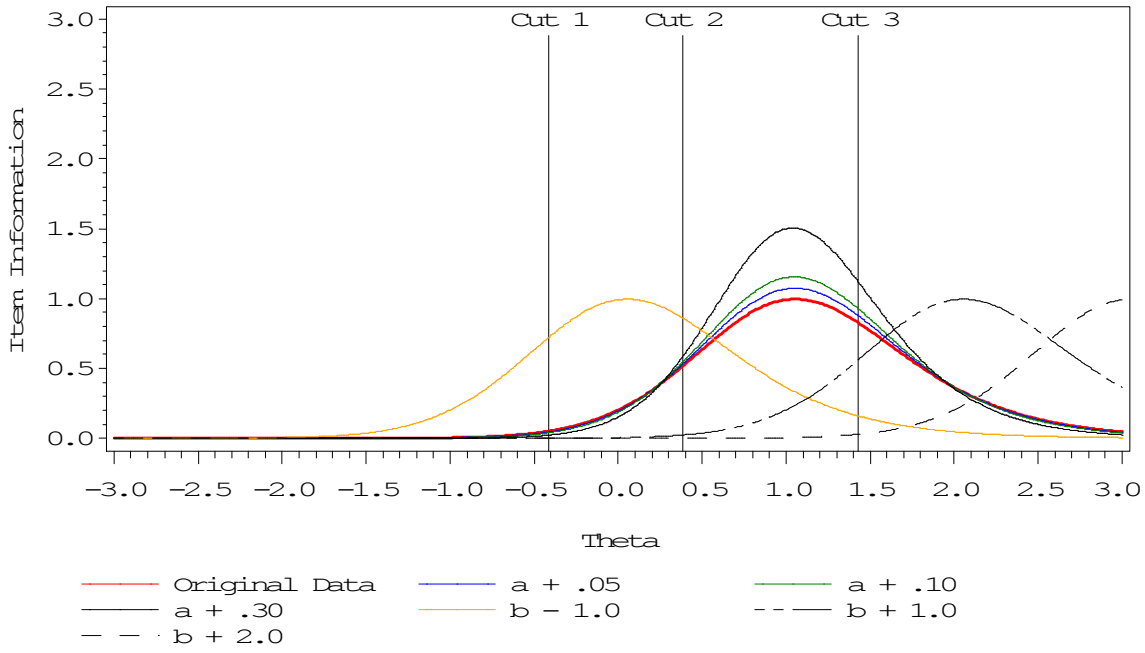
MC 35

Original parameters: a = 1.13, b = .46, c = .25



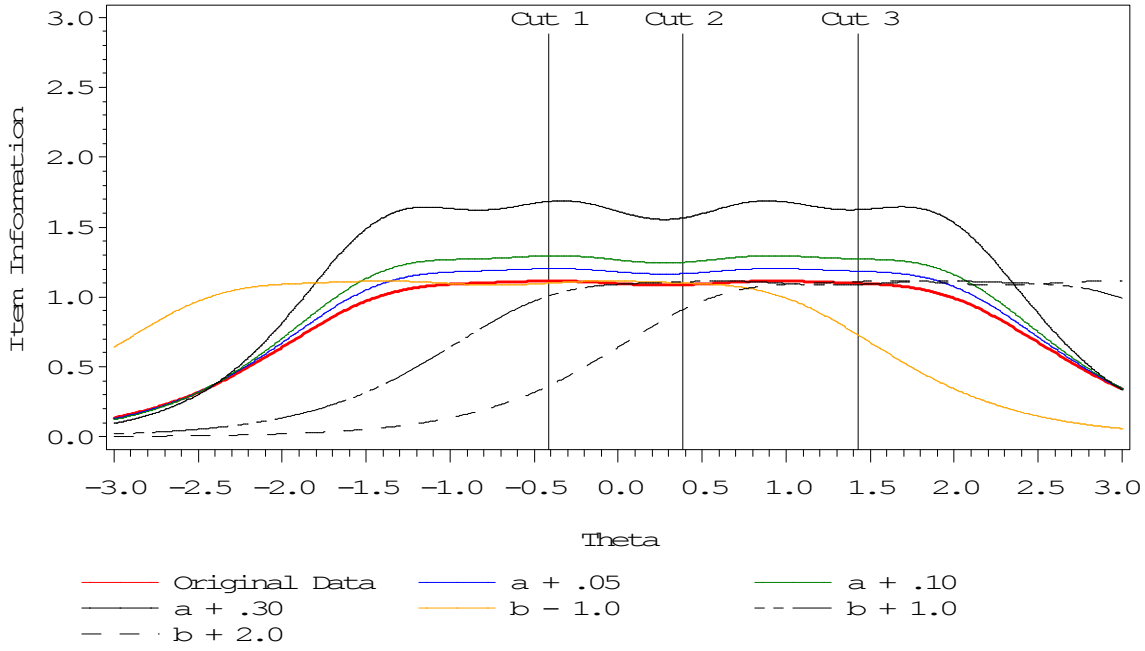
MC 36

Original parameters: a = 1.31, b = .97, c = .11



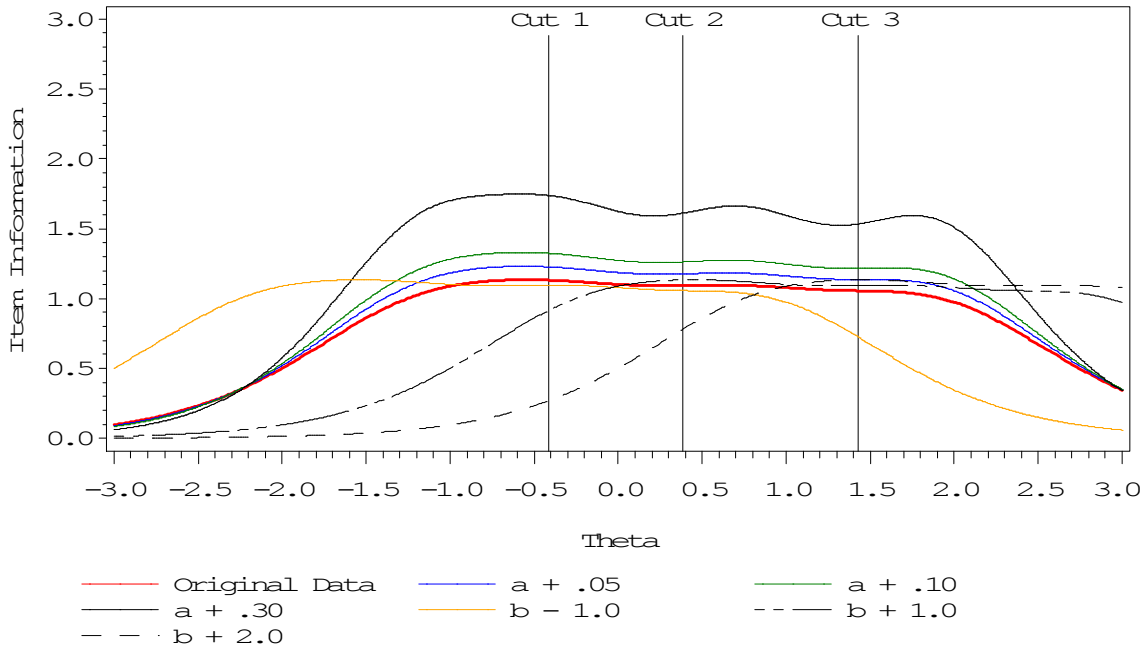
CR 1

Original parameters: a = 1.15, b1 = −1.33, b2 = −.29, b3 = .84, b4 = 1.87



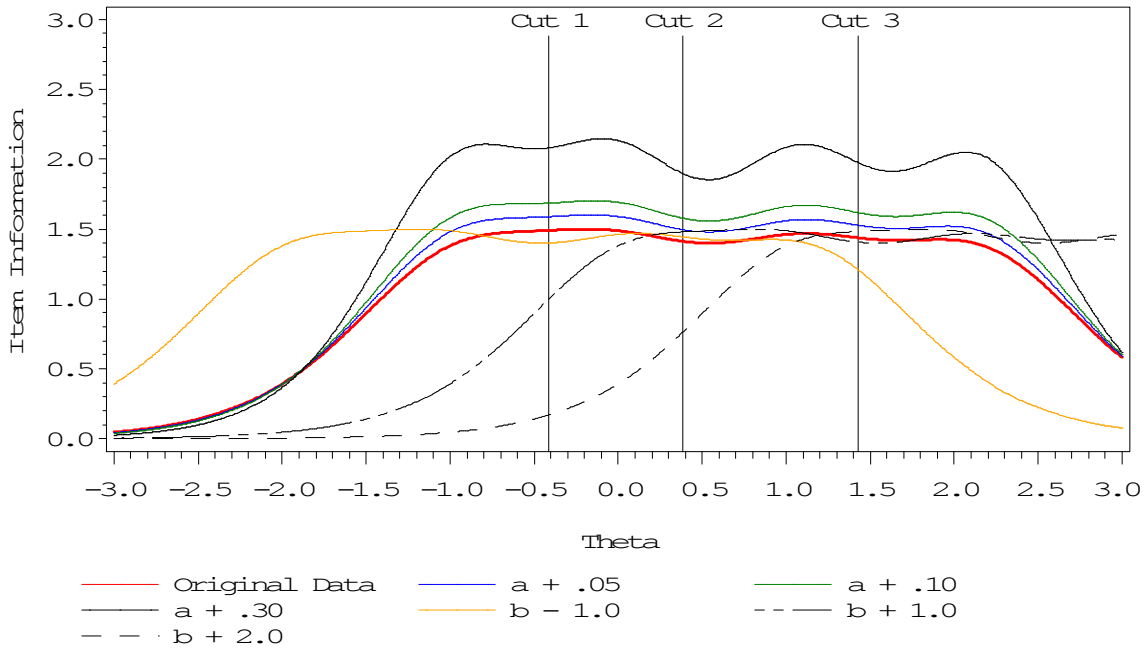
CR 2

Original parameters: a = 1.15, b1 = −1.14, b2 = −.34, b3 = .73, b4 = 1.88



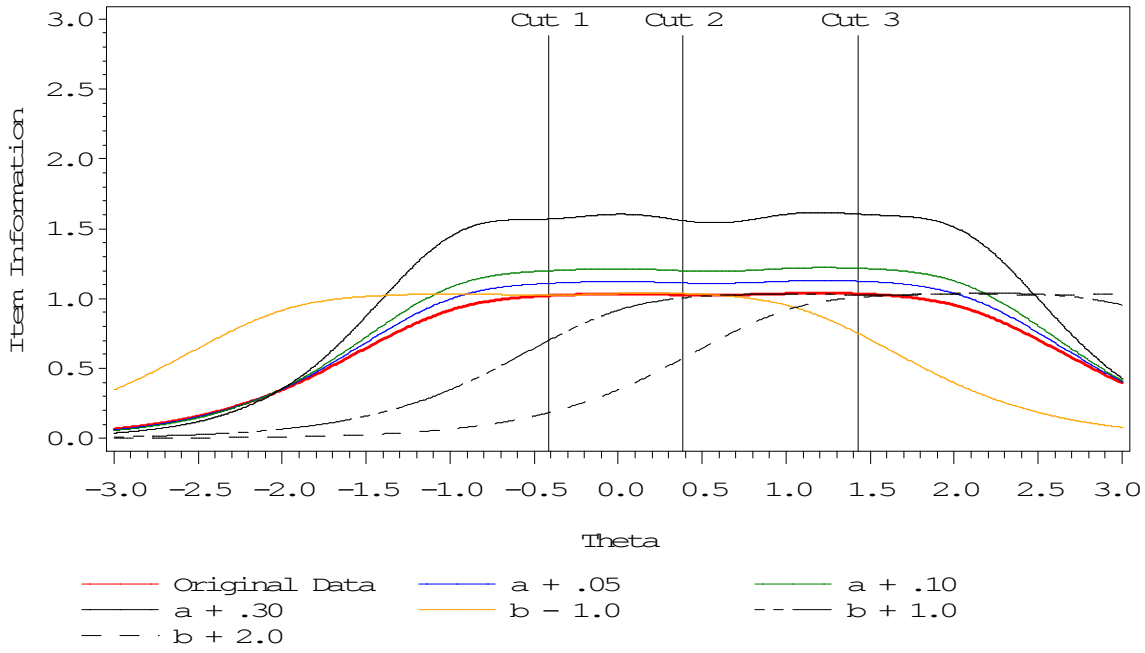
CR 3

Original parameters: a = 1.34, b1 = −.94, b2 = −.02, b3 = 1.09, b4 = 2.16



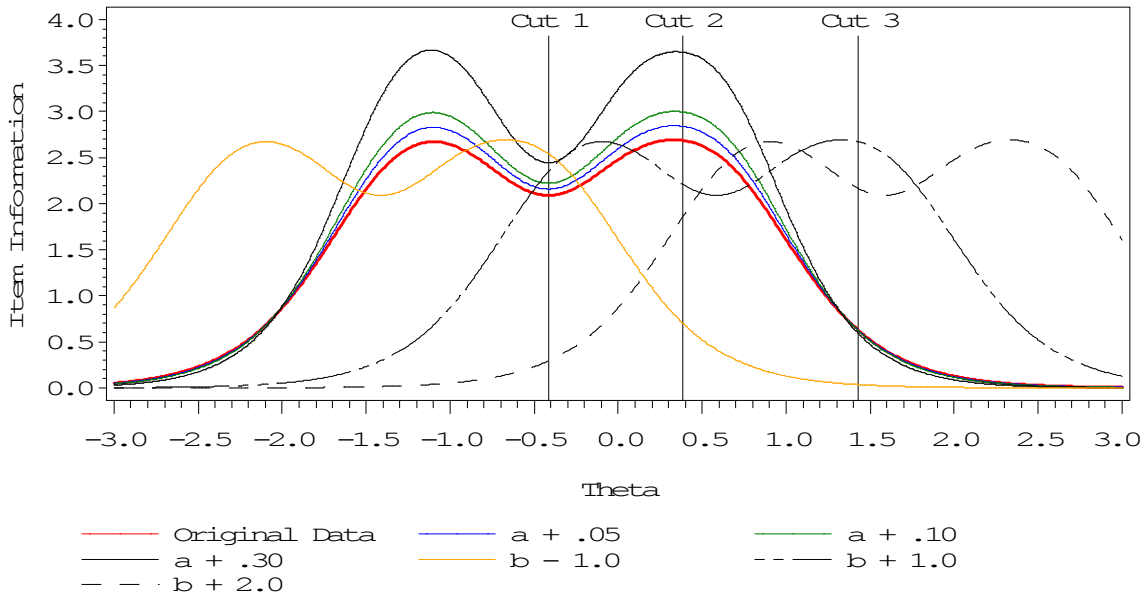
CR 4

Original parameters: a = 1.09, b1 = −.89, b2 = .06, b3 = 1.09, b4 = 1.99



EI 1

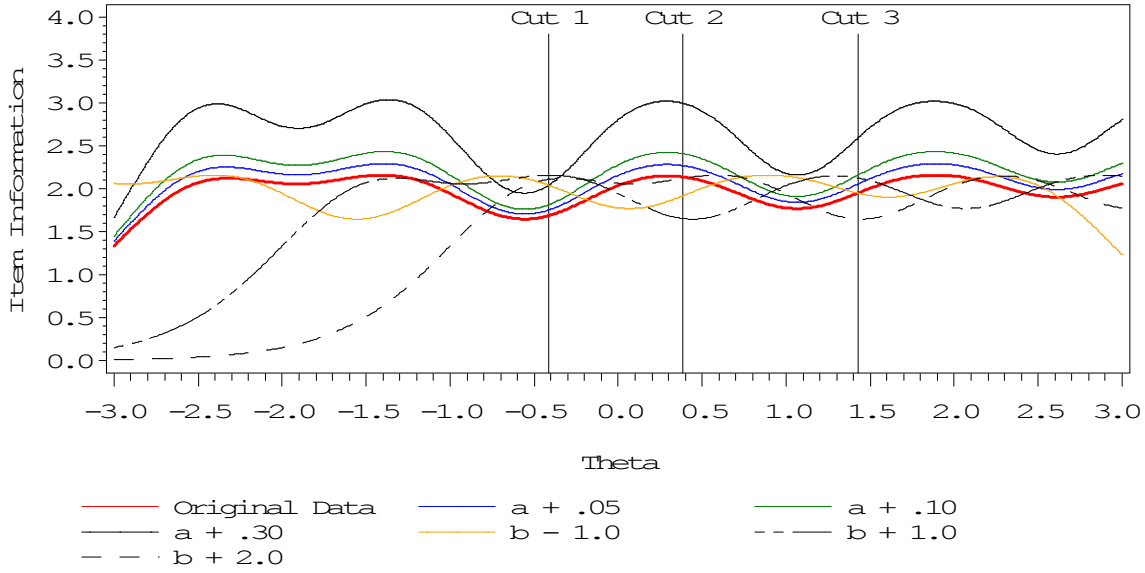
Original parameters: a = 1.77, b1 = −1.30, b2 = −.96, b3 = .12, b4 = .59



Collapse scores 0, 1 and 2 due to low n in each score categories

EI 2

Original parameters: a = 1.57, b1 = –2.32, b2 = –1.49, b3 = –1.17, b4 = .05, b5 = .51, b6 = 1.62, b7 = 2.12, b8 = 3.11, b9 = 3.52



Collapse scores 0 and 1 due to low n in each score categories

BIBLIOGRAPHY

Ackerman, T. A. (1989, March). *An alternative methodology for creating parallel test forms using the IRT information function.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

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

Bejar, I. I. (1977). An application of the continuous response level model to personality measurement. *Applied Psychological Measurement, 1*(4), 509-521.

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

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*(3), 249-253.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Orlando, FL: Harcourt Brace Jovanovich, Inc.

de Gruijter, D. N. M. (1986). Small *N* does not always justify the Rasch model. *Applied Psychological Measurement, 10*(2), 187-194.

Dodd, B. G., & Koch, W. R. (1985, April). *Item and scale information functions for the partial credit model.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Dodd, B. G., & Koch, W. R. (1987). Effects of variations in item step values on item and test information in the partial credit model. *Applied Psychological Measurement, 11*(4), 371-384.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement, 31*(4), 295-311.

Green, B. F. (1983). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A Festschrift for Frederic M. Lord* (pp. 69-80). Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, B. F., Yen, W., & Burket, G. R. (1989). Experiences in the application of item response theory in test construction. *Applied Measurement in Education, 2*(4), 297-312.

Hambleton, R. K. (2006). *Test developments with IRT models* [PowerPoint slides for Education 735]. Amherst, MA: University of Massachusetts.

Hambleton, R.K., & Cook, L. (1977). Latent trail models and their use in the analysis of educational test data. *Journal of Educational Measurement 14*(2), 75-96.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hambleton, R. K., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education, 7*(3), 171-186.

Hambleton, R. K., Jones, R. W., & Rogers, J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement, 30*(2), 143-155.

Hambleton, R. K., & Lam, W. (2009, April). *Redesign of state achievement tests based on a consideration of information functions*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10*(3), 159-170.

Hambleton, R. K., & Traub, R. E. (1971). Information curves and efficiency of three logistic test models. *The British Journal of Mathematical and Statistical Psychology, 24*, 273-281.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage Publications.

Han, K. T. (2006). WinGen3 (Version 3.0.0.414) [Computer software]. Available from http://www.umass.edu/remp/software/wingen/downloadsF.html.

Lin, T. H. (2007). Identifying optimal items in quality of life assessment. *Quality & Quantity, 41*(5), 661-672.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*(2), 179-197.

Loeving, L. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51*(5), 493-504.

Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph No. 7). Iowa City, IA: Psychometric Society.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*(2), 117-138.

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

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lou, Z.-S., Ouyang, X.-L., Qi, S.-Q., Dai, H.-Q., & Ding, S.-L., (2008). 项目反应理论等级反应模型项目信息量 [IRT information function of polytomously scored items under the graded response model]. *心理学报, 40*(11), 1212-1220.

Luecht, R. M. (2006). Designing tests for pass-fail decisions using item response theory. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 575-596). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Masters, G. N. (1988a). Measurement models for ordered response categories. In R. Langeheine & J. Rost (Eds.), *Latent traits and latent class models* (pp. 11-29). New York: Plenum.

Masters, G. N. (1988b). The analysis of partial credit scoring. *Applied Measurement in Education, 1*(4), 279-297.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

Muraki, E., & Bock, R. D. (2003). PARSCALE 4.1: IRT based test scoring and item analysis for graded items and rating scales [Computer software]. Chicago, IL: Scientific Software International.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response models*. Thousand Oaks, CA: SAGE Publications.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34* (Monograph Number 17).

Samejima, F. (1976). Graded response model of the latent trait theory and tailored testing. In C. K. Clark (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing* (pp. 5-17). Washington, DC: U.S. Government Printing Office.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1*(2), 233-247.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11*(4), 263-267.

Thissen, D. (1976). Information in wrong responses to Raven progressive matrices. *Journal of Educational Measurement, 13*(3), 201-214.

Thissen, D. (1991). MULTILOG: Multiple category item analysis and test scoring using item response theory [Computer software]. Chicago, IL: Scientific Software International.

Veerkamp, W. J., & Berger, M. (1999). Optimal item discrimination and maximum information for long IRT models. *Applied Psychological Measurement, 23*(1), 31-40.

Wiberg, M. (2003). An optimal design approach to criterion-referenced computerized testing. *Journal of Educational and Behavioral Statistics, 28*(2), 97-110.

Yamamoto, K., & Kulick, E. (1992, April). *An information-based approach to maintaining content validity and determining the relative value of polytomous and dichotomous items.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Chicago, IL: Scientific Software International.