The Dissertation Committee for Edgar Isaac Sanchez Certifies that this is the approved version of the following dissertation:

A COMPARISON OF ITEM EXPOSURE CONTROL PROCEDURES WITH THE GENERALIZED PARTIAL CREDIT MODEL

Committee:

Barbara G Dodd, Supervisor

Susan N Beretvas

Chris B Brownson

Keenan A Pituch

Tiffany A Whittaker

A COMPARISON OF ITEM EXPOSURE CONTROL PROCEDURES WITH THE GENERALIZED PARTIAL CREDIT MODEL

by

Edgar Isaac Sanchez, B.A., M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May, 2008

Dedication

I dedicate this work to the Lord, whose gifts have allowed me the ability to produce this work and to my beloved parents Rodolfo and Patricia Sanchez, the good Lord's greatest gift to me. It is my prayer that in this work others may see the light of the Lord in my life. It was only through prayer and devotion that I have achieved my goals thus far. My parent's steadfast prayers and support have so profoundly moved me that I am moved to dedicate this work and my life to my parents and my Lord.

Take, O Lord, and receive my entire liberty,
my memory, my understanding and my whole will.
All that I am and all that I possess You have given me.
I surrender it all to You to be disposed of according to Your will.
Give me only Your love and Your grace;
with these I will be rich enough,
and will desire nothing more.
-Saint Ignatius Loyola

Acknowledgements

I would like to first thank my advisor, Barbara Dodd. Through out my graduate career there is no doubt that she, more than any other person, provided me with the constant support that brought me through this process. The time I was gifted to spend learning from her was some of the most fun I have had in years. The support and encouragement she has shown me was invaluable. The faith she had in my abilities was the foundation from which I was able to find the courage, faith, and confidence in myself. At the times, when I had felt lost, her encouragement helped me to cultivate the strength to persevere. I hope that through it all, I have found a way to make her proud of the person her constant support has helped me become. The kindness and friendship she has shown me reminds me of the faith I have in all people.

Second only to my advisor, this dissertation and my graduation was immeasurably aided by the Graduate Coordinator for the Department of Educational Psychology, Virginia Stockwell. Time and again, if it were not for Virginia, I would have been lost in the labyrinth of forms and dates a graduate student must maneuver through. I have no doubt that I would not be graduating if it were not for her encouragement and support. Any time that I needed help she was always there for me. The most amazing this about Virginia is the lengths she will go to help graduate students. Any success I may enjoy in my future, I owe wholly and unequivocally to these two wonderful women. Without either of them, God only knows where I would be today.

I would also like to thank all the faculty and staff of the Educational Psychology department. The time I have spent at The University of Texas has changed me into the person I am proud I have become. The classes I was honored to take from the faculty has given me valuable knowledge and experience. In particular, I would like to thank Bill Koch. From the start of my graduate career, Bill was a wonderful font of encouragement and knowledge for me. I also thank every member of the staff who were always ready and more than able to help me along my journey.

Lastly, I wish to thank my parents Patricia and Rodolfo Sanchez. From childhood they always pushed me to never settle for second best. They helped instill in me the desire to never stop learning and most importantly, my faith. The sacrifices they have made for me over the years are staggering. Through the many highs and lows that life has brought, I always had their constant love and support. Their example has been a light shining brightly as the sun. I hope that my life is a testament to their love for me and their unending love in God.

A COMPARISON OF ITEM EXPOSURE CONTROL PROCEDURES WITH THE GENERALIZED PARTIAL CREDIT MODEL

Publication No._____________

Edgar Isaac Sanchez, Ph.D

The University of Texas at Austin, 2008

Supervisor: Barbara Dodd

To enhance test security of high stakes tests, it is vital to understand the way various exposure control strategies function under various IRT models. To that end the present dissertation focused on the performance of several exposure control strategies under the generalized partial credit model with an item pool of 100 and 200 items. These procedures are relatively easy to implement and have shown promise as an alternative to more complex exposure control strategies. Through unique algorithms these procedures select an item for administration from a subset of items in the item pool. The five procedures examined for efficacy were the modified within .10 logits, restricted modified within .10 logits, randomesque, restricted randomesque, and progressive restricted procedures. The modified within .10 logits, restricted modified within .10 logits, and randomesque, and restricted randomesque procedures select an item for administration from a subset of optimal items. To test the effect of the number of items available for selection in this subset, 3, 6, and 9 items were made available for selection in these procedures. Maximum information item selection was used as a base line, no exposure control, condition.

The progressive restricted, restricted randomesque, and restricted modified within .10 logits procedures were found to optimally protect test security while not significantly degrading measurement precision. The restricted forms of the randomesque and modified within .10 logits procedures proved superior to their base procedures, particularly in controlling average maximum exposure rate. The incrementation of item group size in the modified within .10 logits, restricted modified within .10 logits, and randomesque, and restricted randomesque procedures demonstrated that increasing the item group size provided better test security while not significantly degrading measurement precision. Additionally, in general, the increase of the item pool size from 100 to 200 improved measurement precision and test security. Implications towards practical application are discussed and directions for future research are suggested.

Table of Contents

List of Tables

List of Figures

Chapter I: Introduction

High stakes computerized adaptive tests (CATs) using item response theory (IRT) is an ever growing sector and the security of these tests is an issue that is of vital importance to testing companies. While the issuers of these tests are interested in test security, they are also interested in the initial and most important purpose of the test; that is, to accurately estimate the test-taker's ability level (θ) without having to invest an exorbitant amount of time and money creating an unnecessarily large item pool.

Parshall, Davey, and Nering (1998) cited the three, often, conflicting goals of item selection in CAT. This first goal is to maximize test efficiency by estimating examinee proficiency as quickly and accurately as possible. This goal focuses on maximizing measurement precision with the use of items that will provide the most information or posterior precision for the estimated ability level. The second goal is to protect the item bank from overusing popular items. This goal is focused on test security issues that arise when item exposure rates are left uncontrolled and is therefore referred to as exposure control. The third goal strives to ensure that the constellation of items given to examinees accurately and appropriately represents the content domains to be covered in the exam. As such, this goal calls for appropriate content balancing.

In addition to these three goals, Stocking and Swanson (1998) note that a fourth goal of CAT should be efficient pool utilization. For example, Parshall et al. (1998) found that in a simulated CAT using an item bank of 600 items, over 500 items were administered on only 1% of the simulated CATs. Not only will better pool use increase test security by allowing a uniform usage of all items as opposed to favoring a subset of items, better pool utilization makes good business sense. Developing an item pool is a serious investment for testing companies and it makes sense to get the most out of each dollar spent per item.

The simplest way to select an item for administration is to select the item which provides the most psychometric information at the estimated proficiency level of the examinee. This is done to maximize information provided by the item. As such, the item to be chosen is selected based upon its measurement properties alone. If we look at the maximum exposure rate, the number times an item has been administered divided by the number of tests previously administered, we can see an example of how detrimental it can be to select items based solely on their measurement properties. Focusing on the maximum exposure rate that has been reported in previous studies when no direct method of controlling exposure rate has been done and averaging those values we get an average maximum exposure rate of 0.951 (Davis, 2002; Chang & Ansley, 2003; Burt, Kim, Davis, & Dodd, 2003; and Pastor, Dodd, & Chang, 2002).

In a hypothetical situation, similar to that illustrated by Way (2005) it would be reasonable to expect that at least 200,000 students would take the CAT version of a graduation exam. Let us also assume that no direct item exposure controls are in place to limit the exposure rate of items in the item bank. Given this hypothetical situation, if some items in the item bank have an exposure rate of 0.951 we would expect that over 190,000 students would see these items. Given that these items would be used in subsequent administrations of this CAT graduation exam it would be relatively easy for individuals to organize an effort to memorize and share the items with each other. The negative effects of such an effort to gain item preknowledge is addressed in more detail shortly.

Parshall et al. (1998) noted that when items are selected only on the basis of their measurement properties we see that a subset of the item pool will almost always be administered. This results in a rather small percentage of the total item pool accounting for a rather large proportion of the items being administered. In other words, although the available item pool may be quite large the subset of item used for administration, the functional item pool, can be rather small. This can quickly result in a compromised item bank which no longer provides valid measurement due to over exposure of individual items and the administration of many CATs that use the overexposed items. Parshall et al. (1998) offer three guidelines for dealing with this phenomenon:

1. Use very large item banks of over 5,000 items which can be organized into subpools that can be rotated to avoid item overexposure within a given time period or geographical area.
2. Limit the availability of testing dates to certain periods of the year. This approach would provide greater flexibility in testing date availability than conventional testing but is not "testing on demand."
3. Utilize item selection procedures that directly control the exposure rates of items in the item bank.

They argue that neither the first nor second approaches will be able to ensure the security of the item pool. For this reason they believe that directly controlling the exposure rates will often be necessary. Given the current state of high stakes testing and the push for testing on demand it seems that directly controlling exposure rates when selecting items for administration will always be needed.

In regards to test security, Wainer (2000) noted that without direct control over exposure rates fifteen to twenty percent of an item pool commonly constitutes fifty percent of the test items administered. This subpool of administered items creates a functional item pool that can become a serious problem as “test security seems to increase logarithmically with item pool size” (Wainer, 2000, p. 126). This relationship therefore makes it impractical to depend only upon developing larger item pools to handle test security.

McLeod (1998) conducted simulation research to investigate the effects of item preknowledge on estimated ability level. One of the things she looked at was how various levels of compromising an item bank would affect proficiency estimation. She assumed that examinees would serve as sources who would take and correctly remember a large portion of the items they were administered. They would then report the remembered items to a beneficiary examinee, who would correctly remember a large portion of the items of which they were informed. It was shown that the inflation seen from item preknowledge depended on the number of sources available, the amount of potential gain available, and the concordance of items presented to the beneficiary at the time of testing with the items reported to the examinee by their sources. Not surprisingly it was found that the more sources a beneficiary had and the more potential for gain in estimated ability by the beneficiary the more estimates were inflated. For example, if an examinee has a low score, i.e. has the potential to increase their ability estimate by a large amount, and they have a lot of people providing previously administered items, i.e. has more sources, they will be able to significantly, yet artificially, increase their estimated ability level. It was also seen that the inflation in estimated ability level did not depend so much on the ability of the sources but on the number of sources available to a beneficiary.

McLeod (1998) looked at the effect of having 2, 4, and 8 informants on inflation of examinee ability estimation. As the number of sources increased, the amount of score inflation increased. On average when the beneficiary's true score was 10, examinees with no sources were estimated to have a test score of 10.18 with a standard deviation of 1.54. When the beneficiaries with the same true score had two sources they were estimated to have an average test score of 12.74 with a standard deviation of 6.96. When they had four sources the average test score increased to 22.03 with a standard deviation of 16.44. When the number of sources increased to eight the average test score jumped to 40.35 with a standard deviation of 20.35.

At the median true score of 35, average estimated test scores and standard deviations with no, two, four, and eight sources were 35.06(3.69), 40.71(5.97), 47.12(6.80), and 54.48(4.74) respectively. At a high true score of 55, average estimated test scores and standard deviations with no, two, four, and eight sources were 54.98(1.69), 56.37(1.51), 57.32(1.36), and 58.44(1.17) respectively. We can see that on average the test score was inflated from 3 to 30 points when true score was 10, 6 to 20 points when true score was 35, and 1 to 3 points when true score was 55.

Bearing in mind the negative effects of selecting items based solely on their psychometric properties, as noted by Parshall et al. (1998), the futility of attempting to use larger item pools to handle security issues noted by Wainer (2000), and the potentially dramatic effects McLeod (1998) demonstrated on trait estimation from item preknowledge, this dissertation investigates the utility of using various exposure control procedures with the generalized partial credit model. Each procedure is evaluated on the basis of its ability to control item exposure and test overlap rates. They are also evaluated for their effect upon measurement precision and their ability to efficiently utilize the two investigated item banks. The exposure control procedures investigated are three variants of the modified within .10 logits procedure (Davis & Dodd, 2001), three variants of the modified within .10 logits procedure with and exposure rate constraint, 3 variants of the randomesque procedure (Kingsbury & Zara, 1989), three variants of the randomesque procedure with an exposure rate constraint, and the progressive restricted procedure (Revuelta & Ponsoda, 1998).

Chapter II: Literature Review

The following literature review will attempt to provide for the reader a brief exegesis of computerized adaptive testing as applied with item response theory (IRT). The purpose of this section being the laying of the foundational understanding for the present research study, the reader is first introduced to the fundamental concepts and precepts of IRT. This shall include the introduction of several IRT based measurement models. Following this the implementation of these IRT concepts is described as utilized in computerized adaptive test (CAT) systems. Models of CAT systems are briefly discussed along with relevant considerations for those systems. The emphasis is then focused onto the issue of exposure control methodology and several key strategies that are used for this purpose. This attempted exegesis will culminate in a review of current research with exposure control methodology for the purposes of supporting the present research.

Item Response Theory Measurement Models

Item Response Theory utilizes the set of responses to test items to estimate a person's trait or proficiency level. The trait level is estimated from the responses based upon a given IRT model. This model is a mathematical model which relates observed behaviors to their latent trait and explicates how the item responses are combined numerically to predict the latent proficiency level. IRT models are considered to be strong models due to the necessity of the strong assumptions that must be made and that must be met. These assumptions involve the assumption of a given form for the item characteristic curves (ICCs), local independence, and appropriate dimensionality. The discussion that follows focuses upon the fundamental dichotomous models where items are scored as either correct or incorrect with expansions to the polytomous case detailed in a later section.

Assumptions of Item Response Theory Models

A common assumption of many IRT models is that of unidimensionality. Historically, models have been designed to measure a single dimension at a time. While this can be overcome with models specifically designed to model multidimensionality (McKinley & Reckase, 1982), the use of such models is not presently widespread. Ensuring appropriate dimensionality is a fundamental assumption to properly model performance. Multiple methods have been used to test for dimensionality including Stout's Test of Essential Unidimensionality (DIMTEST) (Stout, 1987) and item-level nonlinear factor analysis (NOHARM) (Fraser, 1988, as cited in Childs & Oppler, 2000). Evaluation of the dimensionality of the data helps to ensure that the IRT model being used accurately models performance.

The second assumption of IRT models is the assumption of local independence. Local independence can be defined in one of two ways. The two definitions vary in the strictness of independence required between observed responses. In the strong definition of local independence the association that is observed between the responses is created by one or more latent traits and when those latent traits are held constant the observed responses are independent. Formally written this can be expressed as:

$$P\left[Y_1, Y_2, \ldots, Y_K\right] = P\left[Y_1 \middle| \eta\right] P\left[Y_2 \middle| \eta\right] \ldots P\left[Y_K \middle| \eta\right],$$

where $Y_1, Y_2, ..., Y_K$ are random observed responses, η is a vector of latent factors, $P\left[Y_1, Y_2, ..., Y_K\right]$ is the joint probability of the observed responses, and $P\left[Y_1|\eta\right]P\left[Y_2|\eta\right]...P\left[Y_K|\eta\right]$ are the conditional probabilities of the observed responses (Bollen, 2002). If the latent factor is responsible for the dependencies among responses then when the latent factor is controlled for in the conditional probabilities the product will equal the joint probability.

The weaker form of local independence requires only that holding the latent factor constant the linear association of responses is zero. Formally written this weaker form of local independence can be defined as $\rho_{Y_iY_j\bullet\eta} = 0$ for all i, j where $i \neq j$ and $\rho_{Y_iY_j\bullet\eta}$ is the partial correlation between responses controlling for the latent factor (Bollen, 2002). Controlling for the latent factor this partial correlation will be zero. If an association remains, it is likely that an incomplete set of factors was used. The difference between the two definitions revolves around the nature of the relationship allowed.

Using the strong definition, it is being stated that once the effects of the latent factor, the proficiency level (θ), have been controlled for, there exists no dependence of any nature among responses. The weaker form requires only that once the effects of the latent factor have been controlled for, there be no linear dependence among the responses. Implied through this is that once the effect of the latent factor has been controlled for, errors of measurement are uncorrelated and therefore unrelated, observed responses have no direct or indirect effects upon each other, and the observed responses do not directly affect the latent factor (Bollen, 2002).

The third assumption of IRT models is that data responses can be accurately represented by the mathematical model chosen. The mathematical model chosen dictates the nature of the responses via an item characteristic curve (ICC). In the case of dichotomously scored items the ICC regresses the probability of item success on the ability level. The relationship is a monotonically increasing one where little change in probability of success is expected at the extreme values of the ability level while rapid change in the probability of success is expected at the point of inflection. The point of inflection is the point on the ICC where the slope is steepest. The point of inflection can be found with the formula $(1-c)/2$ where *c* is the pseudoguessing parameter. In models that do not allow the pseudoguessing parameter to vary *c* will equal zero. This will result in the point of inflection equaling the point on the ICC corresponding to a probability of 0.50.

Functions modeled by an ICC vary according to their lower asymptote, slope and location. Each of these three points corresponds to a statistic being modeled. The lower asymptote is a representation of the pseudoguessing parameter. This pseudoguessing parameter can raise the lower asymptote to represent that even at very low ability levels the probability of a correct response is never zero because it can be correctly guessed.

In dichotomous IRT models the slope of the ICC is multiplied by a constant (0.425) to attain the item discrimination parameter, which describes the relative change in probability around the point of inflection. In highly discriminating items we would expect a greater change in the probability of a correct response on either side of the point of inflection. The location of the item across the difficulty and ability scale depends upon the difficulty of the item. The location of the item is defined by the point on the trait continuum where the point of inflection lies. Being that ability levels and difficulty are on the same scale it can be inferred that for a highly discriminating item a smaller difference between the item difficulty and the trait estimate will result in greater change in the probability of answering the item correctly.

Dichotomous Item Response Theory Models

The manner in which one chooses to analyze responses primarily determines the type of model chosen. When one chooses to score items as correct or incorrect the subset of applicable models are called dichotomous models. This is commonly designated as 0 for an incorrect response and 1 as correct. For this reason these have also been called 0/1 scoring or binary IRT models. The three primary dichotomous models are discussed in detail in the following sections. All models are subject to the assumptions and properties previously described.

Most commonly used IRT models are based upon one of three fundamental models. These models are the one-parameter logistic (1-PL) model (Wright, 1968; Rasch, 1960), the two-parameter logistic (2-PL) model (Birnbaum 1968), or the three-parameter logistic (3-PL) model (Birnbaum 1968). The models each progress towards a more complex modeling of performance. Within this group of models, parameters are fixed or allowed to vary based upon theory. In these models both the difficulty (b) and trait estimate $(\hat{\theta})$ parameter are on the same scale.

The concept of information as used in item response theory provides an index of the amount of psychometric (Fisher) information an item provides across the ability scale. An item information curve (IIC) is a transformation of an item's item-response curve (ICC). A general computational formula for item information is:

$$I(\theta) = \frac{\left[P_i^{'}(\theta)\right]^2}{\left[P_i(\theta)\right]\left[Q_i\right]},$$

where $P_i^{'}(\theta)$ is the first derivative of the item response curve at a given theta value, $P_i(\theta)$ is the probability of getting item *i* correct at a given theta value, and Q_i is equal to $1 - P_i(\theta)$ which is the probability of getting item *i* incorrect at a given theta value.

Embretson and Reise (2000) note two key principles that can be inferred from this formula. The first pertains to where item information is maximized on the difficulty scale. For the 1-PL and 2-PL models an item will provide the most information when items are matched in difficulty to the examinee's ability. For the 3-PL model the most information is provided at an ability level slightly above the item's difficulty parameter as the pseudoguessing parameter lowers the amount of information an item provides (Birnbaum, 1968). The second principle noted is that the higher the discrimination parameter is, for the 2-PL and 3-PL models, the more information the item will provide.

When items have been calibrated on a common ability scale, item information curves (IICs), which plot the amount of psychometric information across the theta scale, are additive. Due to this characteristic of IICs test information may be calculated by summing the item information values of a test. The standard error of measurement of any given examinee's ability estimate is then the reciprocal of the square root of the test information.

The One-Parameter Logistic Model (1-PL)

In the one parameter logistic model the probability of a correct response ($x = 1$) is modeled as follows:

$$P_{ij}\left(x_i = 1 \middle| \theta_j\right) = \frac{\exp\left(\theta_j - b_i\right)}{1 + \exp\left(\theta_j - b_i\right)},$$

where θ_j is the estimated ability level for examinee *j* and b_i is the difficulty or location parameter for item *i*. The point at which the probability of responding correctly is the same as the probability of answering incorrectly is known as the point of inflection. The point on the ability scale that the point of inflection corresponds to is also the difficulty parameter for the item in dichotomous models. It is at this point that the most useful information can be attained. In this model the simple sum of correct responses is a sufficient statistic for estimating ability and therefore every examinee with the same number of correct responses will receive the same ability estimate. Simply stated, the IRT model compares the persons estimated proficiency level $\left(\hat{\theta}\right)$ to an item's difficulty to determine the probability of a correct response. If the examinee's trait level is well above the difficulty of the item then the probability of answering that item correctly is high and when their trait level is well below the difficulty of that item then the probability of a correct response is low.

The Two-Parameter Logistic Model (2-PL)

In the two-parameter logistic model the probability of success ($x = 1$) is defined as:

$$P_{ij}\left(x_i = 1 \middle| \theta_j\right) = \frac{\exp\left[a_i\left(\theta_j - b_i\right)\right]}{1+\exp\left[a_i\left(\theta_j - b_i\right)\right]},$$

where θ_j represents the ability level for examinee *j*, b_i is the difficulty parameter of item *i*, and a_i is the discrimination parameter for item *i*. In this model the discrimination is used as a multiplier for the difference between trait level and item difficulty. The item discrimination is related to the biserial correlation of the performance on the item and the total score. It is because of this that the impact of the difference between the ability level and difficulty is attenuated by the discrimination parameter. Therefore the items that are more discriminating will demonstrate a lesser attenuation on the difference of the ability level and difficulty level.

This also implies that when item discrimination is introduced in the model, the trait level estimate depends upon the response string of the examinee since items will have a differential effect upon the difference of the ability level and difficulty (Embretson & Reise, 2000). As a result, the sum of correct responses is no longer a sufficient statistic for determining ability. In this model the sufficient statistic for calculating proficiency level is the sum of the item discrimination times the raw response for every item (Birnbaum, 1968).

The Three-Parameter Logistic Model (3-PL)

The three-parameter logistic model defines the probability of success (x = 1) for item *i* as:

$$P_{ij}\left(x_i = 1 \middle| \theta_j\right) = c_i + \left(1 - c_i\right)\left\{\frac{\exp\left[a_i\left(\theta_j - b_i\right)\right]}{1 + \exp\left[a_i\left(\theta_j - b_i\right)\right]}\right\},$$

where θ_j is the ability level for examinee *j*, a_i is the item's discrimination parameter, b_i is the item's difficulty parameter, and c_i is the pseudoguessing parameter for item *i*. The pseudoguessing parameter is the lower asymptote of the ICC and represents the probability that examinees with very low ability will correctly guess the answer to the item. In this formula the probability of a correct response is, in essence, the sum of the probability of a correct response due to chance and the probability of getting the item correct according to the two-parameter logistic model attenuated by the probability of getting the item correct due to ability. In this model, the point of inflection which corresponds to the *b* parameter moves as dictated by the formula $\left(1 - c\right)/2$ or simply half the distance of the pseudo-guessing parameter to the upper asymptote (1.0). For the three-parameter logistic model there exists no sufficient statistic.

Polytomous Item Response Theory Models

When one chooses to score items polytomously their answer is not recorded as right or wrong, as in the case of the models previously described, but rather their response is assessed using a gradient scale. That response is used to assess additional information about the examinees trait level. In the case of polytomously scored items, category response curves (CRCs) regress the probability of response in each category on the ability level. In this section several polytomous IRT models are introduced. Each provides a unique model with which to handle various types of data. Thissen and

Steinberg (1986) classified polytomous IRT models as members of either divide-by-total models, difference models, or left-side added divide by-total models.

In difference models such as the graded response model (GRM, Samejima, 1969) the probability of responding in any particular category is obtained through subtraction. The GRM is appropriate for Likert type items and items that have ordered categorical responses. In divide-by-total models such as the partial credit model (PCM, Masters, 1982), the generalized partial credit model (GPCM, Muraki, 1992), and the rating scale model (RSM, Andrich, 1978) “the probability of responding in a given category is obtained by dividing the numerator by the sum of all category probability numerators so that the probabilities conditional on θ sum to unity” (Dodd, De Ayala, and Koch, 1995). The PCM was originally developed for the analysis of data where multiple ordered steps are required to produce the correct answer. In addition this model allows for the modeling of attitudinal data using a multi-point scale. The GPCM is, as the name suggests, a generalization of the partial credit model where the slope parameter is allowed to vary between items. The RSM is similar to the partial credit model in its handling of attitudinal data however it makes use of a fixed set of rating points and makes no assumption in regards to the spread of category intersection parameters.

Samejima (1969) extended the information function for dichotomous IRT models to the polytomous case. This extension of item information was defined as:

$$I_i(\theta) = \sum_{x=1}^{m_i} \frac{\left[P_{ix}^{'}(\theta)\right]^2}{P_{ix}(\theta)},$$

where *m* is the number of categories for item *i*, $P_{ix}^{'}(\theta)$ is the first derivative of $P_{ix}(\theta)$ with respect to theta, and $P_{ix}(\theta)$ is equal to the probability of responding in category *x* on item *i*. To calculate the test information, item information is summed across all items. This extension of the dichotomous calculation of information can be used with all polytomous models.

Graded Response Model (GRM)

The graded response model (Samejima, 1969) is a difference model which calculates the probability of a response by subtraction. This model is used when there are more than two categories of items that can be ordered to correspond to varying degrees of the trait measured by the item. The graded response model is used when ordered categorical responses are used as in the case of Likert scales or partial credit scoring. This model also allows the number of categories in the items to vary. Item parameters can vary in discrimination (a_i) and difficulty (b_{ix}) using the homogenous case of this model.

This difference model requires a two step process to calculate the cumulative probability of responding in a particular category. This process requires the response categories to be artificially dichotomized at every point and then subtracted from consecutive ordered categories to attain the probability of an examinee responding in a given category or higher. The number of ways to dichotomize the item $(m+1)$ is derived from the number of categories m . The graded response model defines the probability function of scoring in category x or higher in item i given the examinee's trait level (θ) as:

$$P_{ix}^{*}(\theta)=\frac{\exp\left[a_i(\theta-b_{ix})\right]}{1+\exp\left[a_i(\theta-b_{ix})\right]},$$

where $x=1,\dots,m_i$, a_i is the item discrimination, and b_{ix} is the category boundary defined as the theta level corresponding to the 0.50 probability level of the P_{ix}^{*} function. For each item a single discrimination parameter is assumed across all levels of an item's categories.

To calculate the probability of responding in a given category, adjacent cumulative $P_{ix}^{*}(\theta)$ functions are subtracted as follows:

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

In the lowest and highest categories ($x=0$ and $x=m+1$), the cumulative probabilities will always result in $P_{i0}^{*}(\theta)=1.0$ and $P_{i,m+1}^{*}(\theta)=0$ respectively. What this means is that the probability of responding in the lowest category or higher is 1.0, and the probability of responding above the highest category is always 0.0.

Partial Credit Model (PCM)

The partial credit model was developed by Masters (1982) as an extension of the Rasch model extended to handle analysis of polytomously scored items. In this divide-by-total model the probability function of responding in category $x=j$ on item *i* given the examinee's trait level θ is defined as:

$$P_{ix}(\theta)=\frac{\exp\left[\sum_{j=0}^{x}(\theta-b_{ij})\right]}{\sum_{r=0}^{m_i}\left[\exp\sum_{j=0}^{r}(\theta-b_{ij})\right]},$$

where $\sum_{j=0}^{0}\left(\theta-b_{ij}\right)\equiv 0$, m_i is the response category, and b_{ik} is the step difficulty parameter for category *j*. As can be seen from the probability formula this model does not allow for the discrimination parameters between items to vary. This model allows for the use of steps that are not ordered in terms of difficulty but are ordered in terms of steps of completion.

Generalized Partial Credit Model (GPCM)

In the present study, the generalized partial credit model (Muraki, 1992) will be used. This model is appropriate for items in which multiple steps are necessary for successful completion or in which items may be awarded partial credit. This model was developed as an extension of the partial credit model, which is itself an extension of the two parameter logistic model. In this model the slope and step difficulty parameters are allowed to vary.

Examinee responses are categorized in (*m* + 1) ordered categories. In this way a higher category score indicates a higher amount of the latent trait and a lower category score indicates a lower amount of the latent trait. As with the partial credit model the ordering of steps is necessary in regards to steps of completion and not difficulty. The probability of responding within a given category under this model is defined as:

$$P_{ix}\left(\theta\right)=\frac{\exp\sum_{j=0}^{x}a_i\left(\theta-b_{ij}\right)}{\sum_{r=0}^{M}\left[\exp\sum_{j=0}^{r}a_i\left(\theta-b_{ij}\right)\right]},$$

where $\sum_{j=0}^{0} a\left(\theta - b_{ij}\right) \equiv 0$, m_i is the number of categories minus one and b_{ij} is the step difficulty parameter for category x in item i.

Step difficulty parameters are the point at which two adjacent curves intersect. Given that an examinee has completed previous steps this is the point at which the probability of adjacent category curves is equal. The slope parameter is a measure of the variance of categorical responses as θ changes. Category response curves (CRCs) are flatter for slope values below 1.0 and are more peaked as the slope increases to 1.0.

Rating Scale Model (RSM)

The rating scale model (Andrich, 1978) is a polytomous model expanded from the Rasch model. In this model, items with the same response format have a single scale location parameter (b_i) which reflects a relative measure of difficulty for the item. In addition, each $J = K - 1$ category thresholds, across all items, is described by a single category threshold parameter (t_i). Since a single response format is used, all categories are assumed to have the same set of thresholds. For example, in a rating scale with anchors of 1 = disagree, 2 = no opinion, and 3 = agree, across items, it is assumed that the psychological distance between the three categories is constant. With that in mind it is reasonable to assume a constant set of t_j parameters across items and a single b_i location parameter per item. In addition, the model assumes that the discrimination parameter is constant. In this model the probability of response in a given category is defined as:

$$P_x\left(\theta\right)=\frac{\exp\left\{\sum_{j=0}^{x}\left[\theta-\left(b_i+t_j\right)\right]\right\}}{\sum_{x=0}^{M}\exp\left\{\sum_{j=0}^{x}\left[\theta-\left(b_i+t_j\right)\right]\right\}}$$

where $\sum_{j=0}^{0}\left[\theta-\left(b_i+t_j\right)\right]=0$.

Computerized Adaptive Testing

Each of the aforementioned IRT based model provides test developers with unique methods to operationalize an examinee's ability. One a test developer has decided which model best serves the theoretical and practical demands of their situation they must next decide how to implement a CAT for their unique situation. Prior to exploring the manifestations of operational CAT systems a discussion of the justification for implementing a CAT as opposed to the more traditional paper and pencil test is needed.

It has been suggested that testing using a CAT offers many benefits over traditional paper and pencil testing. Wainer (1990) noted that among the benefits of using a CAT are increased testing efficiency, improved test security, a reduction in effects of speededness for some examinees, reduction of examinee boredom and frustration, the elimination of separate answer documents, immediate scoring and feedback to examinees, easy removal of faulty items, simplified pre-testing of new items, and the ability to add new and innovative item types.

Item selection in CAT provides test creators with a method of exposing examinees to items that are appropriate for their individual ability levels. Since item selection is dependent upon the ability estimate at the time of item selection, items that minimize the difference between the ability estimate and item difficulty are selected as possibilities for administration. In addition, IRT allows for estimation of the standard error of measurement throughout the test as opposed to a single index for the whole test, as in classical test theory. Therefore measurement precision of the test is increased. This also means that across the ability scale test developers can examine the standard error of measurement at different ability levels. Implicit in the item selection procedure is that inappropriate items with large differences between the current ability estimate and the item's difficulty parameter will not be considered for administration.

Increased test security results from both the lack of physical documents as well as the adaptive nature of the item selection procedure used throughout the test. The lack of physical testing documents eliminates the concern of theft of test materials. In addition, the nature of the item selection precludes many items in the item bank from being administered to, and therefore seen by many examinees. This restricted item pool for each individual theoretically decreases the likelihood of artificially inflating trait estimation as the examinee would need to learn a large portion of the item pool to affect their trait estimate and therefore their test score. As illustrated previously, however, in practice this advantage may not be realized. The benefits seen through this can also be attenuated by the distribution of the item pool. The nature of the item pool distribution shall be described in the subsequent section on components of a CAT.

The pace at which the examinee takes the test is largely left to the individual examinee. The examinee is usually presented with a single item at a time with the subsequent items being administered once a response is given. This means that within practical time limits the examinee may use as much or as little time as they may need. In addition, the actual time that an examinee takes to provide a response can itself become a time variable to be potentially used in assessing proficiency level or for diagnostic purposes.

As noted previously, the item selection procedure used implies that examines only see items tailored to their particular ability level. This means that they should never receive items that are too easy or too difficult for their ability level. In traditional paper and pencil tests, versions of the test are created and items included are fixed regardless of ability level. This means that they will likely see items too difficult and too easy for them. As a result items that are too easy for the examinee can result in boredom and overly difficult items can frustrate the examinee, both of which can negatively affect performance.

The elimination of a separate answer document for examinees also reduces the likelihood of errors in marking responses to items. Large scale multiple choice testing in the traditional paper and pencil method requires examinees to formulate an answer and then find the appropriate place to mark their response on a separate document which is usually a sheet of bubbles for their responses. In this case, if an examinee marks one item in the wrong place the likelihood of marking all subsequent items in the incorrect place increases. The result is that their ability estimate is incorrectly estimated and their resulting test score is negatively impacted. In CAT, responses are elicited and stored when each item is presented and therefore this potential problem is eliminated.

Theoretically, examinees have as much time as they need to complete the exam. Realistically, however, the cost of testing time necessitates some time limits. Care must be taken in how time limits may effect a test as was seen in 1997 by ETS when they chose to allow the scoring of tests that were at least eighty percent complete, and retesting those who did not complete eighty percent. Later research revealed that an average examinees score was increased if only their first eighty percent of responses were used for trait estimation. For that reason they were forced to cease the use of that practice and subsequently scored all unanswered items as incorrect (Wainer, 2000). This demonstrates that while some traditional problems associated with testing may be solved, new problems may arise. Wainer also alludes to the potential for technical problems that may be experienced by the computers used for administration. This could include computer or network failure and electronic security breaches.

Models of Administration

CAT systems can by utilized in a number of different ways and in a number of different configurations. CAT systems can be designed to be administered and scored by a stand alone computer or a network can be designed to allow computers to remotely administer and score tests. Stand-alone systems must contain all needed software and the item pool to be used. The system must be able to record all needed data and be able to output the data in a useful medium. The system can either score the test and provide a result to the examinee or not. It can also report the score to a proctor or central server where scores can then be reported to examinees. This may be useful for CAT systems regardless of its configuration as a stand-alone or network system.

If a networked approach is desired the system can be designed in several configurations. Depending upon the approach taken some functions can be assigned to the terminal used for administration while other functions can be assigned to a proctor's computer or other networked server. Several alternatives are also available for handling the item bank. When an examinee begins the test their terminal can download the items pool to be used from the proctor's computer or some central server if they do not reside on the examinee's computer. Intricate pool rotation can also be used to maximize pool utilization and exposure control.

Components of a Computerized Adaptive Test

Of the many considerations in producing a CAT one must keep in mind the design of the item pool, the item selection procedure, the trait estimation method, the stopping rule, content balancing, and the exposure control strategy used. Each has the potential to bring complications to a testing system if improperly handled. Each of these considerations is therefore now discussed.

Item Pool

The goal of the design of an item bank for use in a CAT is high quality, highly discriminating, items at all levels of the proficiency scale. In traditional paper and pencil tests used for norm-referenced tests, item banks are created to get the best measurement at the average ability level. This results in item pools whose highly discriminating items are normally distributed. In paper and pencil tests used for criterion-referenced tests, item banks are constructed to get the best measurement possible to assess mastery of some content. In either case, the items on every printed test form are fixed and cannot be changed regardless of their selection from a larger item bank. In CAT the uniform distribution of high quality items throughout the proficiency level ensures that the subset of items selected for administration is appropriate for any examinee's proficiency level (Flaugher, 2000).

In addition to this requirement, the item pool must satisfy the specific requirements of the IRT model chosen, the item selection algorithm, and content balancing restrictions. The appropriate size of an item pool is dependent on many variables; item types (dichotomous, polytomous) (Dodd et al., 1989; Davis, 2002), item selection method (Kingsbury & Zara, 1989; Davis, 2002), and content balancing (Kingsbury & Zara, 1989; Davis, 2002; Dodd et al., 1995) are a few of these variables. Kingsbury & Zara (1989) report that when dealing with content balancing, exposure control, and high information assurance across all levels of θ, an item pool of only 100 dichotomously scores items will not suffice.

Flaugher (2000) suggested the following general plan for item pool development:

1. Given the test specifications it is necessary to create enough items to satisfy each content category as well as the item selection procedure.
2. Evaluate the quality of the written items being sure that they are presented to test and sensitivity specialists.
3. One should pretest newly written items to be used in the item pool. Wainer (1990) notes that, if need be, items pre-tested via traditional paper and pencil methods can be used for CAT purposes.
4. Use the item-analysis statistics from the pretest to determine items to be included in the item bank. Using traditional indices such as the proportion correct and biserial correlation between the item and the total test score are used in addition to IRT indices. Ability regression plots can also be used

to examine how well a model fits the data (Hambleton & Swaminathan, 1985)

5. One should ensure that the content balancing needs of the test are met and it is recommendable to run simulations at various proficiency levels to ensure appropriate item functioning

Item Selection Procedure

In CAT systems, item selection is adaptive to the needs of the examinee. Upon administration and response to an item the examinee's proficiency level is re-estimated and used as criteria for selection of the subsequent item such that a given mathematical formula is maximized. The two fundamental item selection procedures are the maximum information procedure and the Owen's Bayesian procedure (Owen, 1969). In the maximum information procedure items are selected such that items maximize the information attained at the examinees proficiency level. Upon response to an item by the examinee item information is recalculated for all unadministered items and the item with the greatest information value is selected for administration.

Owen's Bayesian procedure seeks to minimize the posterior variance of the trait estimate. Upon the submission of an examinee's response to an item the posterior distribution of the trait level is re-estimated. The item that will minimize the variance in the distribution is selected for subsequent administration. Thissen and Mislevy (2000) noted that Owen's Bayesian procedure tends to result in a proficiency estimate that fluctuates given the item order. Samejima (1980) also noted that individuals who receive the same items and respond in the same way may result in different test scores.

Trait Estimation

Under item response theory (IRT), trait estimation is a function of an examinees item response pattern, particular item parameters, and the IRT model used. In real-world situations it is usually the case that no information about examinees ability level is known. For this reason it is common to start the CAT assuming the examinee has an ability level equal to the mean of the examinee population. However, when prior information is available it is also possible to change this initial ability level. Trait estimation occurs throughout the CAT. Upon the administration and receipt of a response the CAT system reestimates the ability level of the examinee and uses this provisional trait estimate in selecting subsequent items for administration. This process continues throughout the CAT until some termination criteria has been met. The last ability estimate is at this point reported as the examinees estimated ability level given the full set of responses to the items administered. This estimation of ability level is commonly conducted through either maximum likelihood estimation or Bayesian estimation. The following two sections detail maximum likelihood estimation (MLE) and expected a posteriori (EAP) estimation.

Maximum Likelihood Estimation. Maximum likelihood estimation (MLE) is commonly used as the estimator of choice in computerized adaptive tests (CATs). When using the maximum likelihood estimator the formula for the conditional probability of the vector of item responses ($\boldsymbol{x}$) for examinee *j* to items 1 to *i* given θ and the matrix of item parameters (β) is as follows:

$$P\left(\boldsymbol{x}\middle|\theta_j,\beta\right)=\prod_i P_i\left(\theta\right)^{x_{ji}} Q_i\left(\theta_j\right)^{1-x_{ji}}.$$

In this equation the term $\left[P_i\left(\theta\right)^{x_{ji}}\right]$ represents the ICC for correct responses for person *j* on items 1 to *i* while the term $\left[Q_i\left(\theta_j\right)^{1-x_{ji}}\right]$ represents the ICC for incorrect responses for person *j* on items 1 to *i*. In order for this formula to hold true it is essential that the model hold true for the observed data and that each item is conditionally independent. MLE of θ provides the value $\left(\hat{\theta}\right)$ which maximizes the likelihood of an item response pattern or equivalently its log likelihood ($\ln \mathrm{L}\left(\boldsymbol{x}\middle|\theta\right),$ where $\boldsymbol{x}$ is the vector of item responses). To solve for the trait estimate which maximizes the log likelihood function the first derivative of $\ln \mathrm{L}\left(\boldsymbol{x}\middle|\theta\right)$ with respect to θ is set to zero such that (Wang, 1999):

$$\frac{\partial}{\partial}\sum_{i=1}^{n}\sum_{k=0}^{m_j} x_{ik} \ln P_{ik}\left(\theta\right)=\sum_{i=1}^{n}\sum_{k=0}^{m_j}\left(\frac{x_{ik}}{P_{ik}}\right)\left(\frac{\partial P_{ik}}{\partial\theta}\right)=0,$$

where $P_{jk}\left(\theta\right)$ is the probability of an examinee responding to category *k* on item *i* based on a polytomous IRT model with ordered responses.

As this equation cannot be solved directly it is necessary to use the Newton-Raphson procedure to solve for θ. This iterative process utilizes the ratio of the first derivative to the second derivative $\left(\varepsilon\right)$ to adjust a previous $\hat{\theta}$ by subtracting ε to achieve a new $\hat{\theta}$ value. This iterative process of adjusting old estimates to achieve new estimates is repeated until the ratio reaches some prespecified value, typically 0.001 (Embretson & Reise, 2000). One limitation of the likelihood function is its inability to estimate the trait level when responses to CAT items are either all correct or all incorrect, in dichotomous models, or when category responses are all in the same response category, for polytomous models.

For this reason it is necessary to employ variable step-size estimation until there is at least one correct and one incorrect response, for dichotomous models, or two different response categories, for polytomous models. This variable step-size estimation procedure will set the theta estimate to half the distance from the current ability estimate to the upper or lower maximum or minimum step difficulty in the item pool depending upon whether the item responses have been in the upper or lower half of the response scale (Koch & Dodd, 1989).

Expected a Posteriori Estimation. In expected a posteriori (EAP) estimation it is possible to incorporate prior knowledge into the estimation process. The EAP estimator is derived by finding the mean of the posterior distribution (Embretson & Reise, 2000). For a CAT a set of probability weights $\left(W\left(Q_r\right)\right)$ are calculated at a given number of θ quadrature points $\left(Q_r\right)$ that typically take the form of the standard normal distribution. These weights are transformed such that they equal 1.0 and represent a discrete prior distribution. Using these quadrature points and weights the EAP estimator is calculated using the following formula (Bock & Mislevy, 1989, as cited in Embretson & Reise, 2000):

$$\theta = \frac{\sum_{r=1}^{q}\left\{\left[Q_r\right]\left[L\left(Q_r\right)\right]\left[W\left(Q_r\right)\right]\right\}}{\sum_{r=1}^{q}\left\{\left[L\left(Q_r\right)\right]\left[W\left(Q_r\right)\right]\right\}},$$

where $L(Q_r)$ is the exponent of the log likelihood for q quadrature points. This noniterative estimation procedure provides estimates even when responses to CAT items are either all correct or all incorrect, in dichotomous models, or when category responses are all in the same category, for polytomous models.

Stopping Rule

The stopping rule can be designed to terminate the test when a fixed number of items have been administered, a target measurement precision is reached, or after some specified time limit. Using a fixed number of items is the easiest stopping rule to implement and exposure rates can be more easily predicted. Unfortunately, this will result in a variable measurement precision with the worst measurement precision being associated with more extreme trait levels. When a target measurement precision is used test administrators can be assured that a desired measurement precision is always attained.

This implies that the test length will be variable, however, simulations can be conducted to observe likely test lengths at various proficiency levels. In practice, however, it may prove more difficult to convince stakeholders of test fairness with variable length tests. The final option in setting a stopping rule is to specify a maximum time limit. In certain situations it may be desirable to terminate the test after a fixed time limit. In situations where a test taker uses an inordinate amount of time it may be necessary to terminate the test.

Content Balancing

Content Balancing allows test designers to ensure that each CAT taken represents the appropriate domains to be covered. The nature of CATs dictates that each examinee will receive a unique constellation of items. In essence a new form is created for every examinee. There are several methods to handle multiple content domains should the test require that. One option is to divide the test so that each content domain constitutes its own form. In this way each domain is estimated alone. Alternatively one could make use of a multidimensional model

One of the more frequently used method for controlling content balancing is through the use of a constrained CAT (CCAT) (Kingsbury & Zara, 1989). After the administration of an item this method calculates the percentage of items that have been administered in each content domain and compares those percentages with the desired target percentages for the test. The content domain with the largest discrepancy is then identified. From within this identified content domain the item which best satisfies the item selection procedure is administered.

Exposure Control

Exposure control must be considered as overexposure can commonly be seen with maximum information item selection. In maximum information item selection examinees with the same trait estimate will be administered the same item, as that item will always offer the highest information. This is particularly threatening at the beginning of a CAT. At this time estimated ability for many examinees will be similar. The concern is for test security due to overexposure of popular items. The availability of test dates can also play into exposure control and test security. If the test is a nationally offered test and test dates are frequent or if testing on demand is used it is possible for examinees to join together in hopes of attaining enough of the functional item pool to artificially increase their test scores as McLeod (1998) illustrated. Specific exposure control strategies are now introduced.

Exposure Control Strategies

In attempting to control the exposure of the items in the item bank two broad types of strategies have grown in prominence. These two strategies are the randomization and conditional strategies (Way, 1998). Rather than choosing the single most informative item for selection, the randomization strategies select the item for administration from a group of items that are virtually optimal. These types of strategies are relatively easy to implement, as they do not require additional simulations as is needed in the conditional strategies. While the randomization strategies are easy to implement, most have no way to guarantee that item exposure rates will remain at acceptable levels. In all conditional strategies "the probability that a selected item will be administered is conditioned on the frequency with which the item is selected within a particular targeted population" (Way, 1998).

Based upon the principles of the randomization and conditional strategies a third type of strategy, which is subsequently referred to as combinatorial procedures, utilizes the principle of conditioning item administration on exposure rate and subsequently administering an item based on randomization principles. A fourth exposure control strategy discussed in the literature is called stratification procedures. These procedures incorporate a method of stratifying the item bank based upon a single or multiple item parameters to increase item usage while reducing exposure rate throughout the test.

Way (1998) suggested that there were at least two key factors in determining how much control of item exposure is needed. The first factor to consider is whether the stakes of the outcome of the exam are high or low. High stakes exams would include admissions tests and licensure exams. Low stakes exams would include placement exams and educational testing. The second factor is the control over the examinee population. When there is low control over the examinee population there is little to no control over who can take the exam and conversely when there is high control over the examinee population there is strict control over who can take the exam.

These factors interact to produce four "quadrants" which Way (1998) labels high stakes/low control, high stakes/high control, low stakes/low control, and low stakes/ high control. Way offers several suggestions for controlling item exposure in computerized adaptive testing based upon CATs using dichotomous models. He suggests that for high stakes admissions exams the ratio of pool size to test length should be 12:1. The average item exposure control should range from 0.08 to 0.12. The overall average percent overlap should range from 10-15%. Finally he suggests that the maximum average percent overlap conditional on ability, examinees with similar abilities, should be 30%. One must bear in mind, however, that these recommendations are for dichotomous models and may not be the best guidelines for polytomous models. The following sections provide a detailed discussion of exposure control strategies as they apply to polytomous item response theory models.

Randomization Procedures

The randomization strategies select items for administration from a set of nearly optimal items and are variations of the maximum information procedure. These are referred to as randomization strategies because they incorporate a random component in deciding item administration. These methods include the 5-4-3-2-1 procedure, randomesque method, the modified within .10 logits procedure, the restricted modified within .10 logits model, and the progressive procedure.

5-4-3-2-1 Procedure

In this procedure the first five items to be selected for administration are not selected solely based upon maximum information. In this procedure the first item to be selected for administration is selected randomly from among the five most informative items. The second item is then randomly selected from the four most informative items. This continues until the fifth, and all subsequent, item(s) where the procedure selects the most informative item.

Randomesque Procedure

The randomesque procedure for selecting items was developed to deter item over-exposure (Kingsbury & Zara, 1989). In randomesque item selection the CAT randomly selects an item from among a subset of most informative items for the estimated ability. The CAT programmer specifies the number of items chosen to select the administered item from. Kingsbury and Zara suggest that the randomesque procedure results in a secure testing procedure because examinees with the same ability level will most likely not see the same items. Even if examinees share information about the items they saw, beneficiaries' ability to predict a set of items from an extremely large item pool becomes almost unattainable (Kingsbury & Zara, 1989).

Modified within .10 logits Procedure

The within .10 logits procedure, originally proposed for the dichotomous models, randomly selects the next item for administration from a set of items within .10 logits of the desired item difficulty (Lunz & Stahl, 1998). While we would expect this to add some variation to the selection algorithm, the number of items available for the random selection is dependant upon the distribution of the item difficulties in the item bank. If content areas are also implemented in the testing algorithm there is the potential for the item bank to lack the spread of difficulties needed to adequately perform the algorithm. If such a case arises the next item selected for administration is the item whose difficulty is closest to the desired difficulty.

The within .10 logits procedure for item exposure control does not use the information function to select the items to be administered. Instead, this method selects items within .10 logits of the item difficulty needed to match the current θ estimate, and the administered item is randomly chosen from these items (Lunz & Stahl, 1998). Lunz and Stahl used this exposure control procedure with five item pool sizes (from 183 to 823 items) that were content balanced. The .10 logits procedure was successful in administering a greater number of items across testing sessions and controlled test overlap, especially with larger pool sizes (Lunz & Stahl, 1998).

Davis and Dodd (2001) expanded this procedure to work for the polytomous case. This modified within .10 logits procedure selects items for administration via maximum information selection at three points on the difficulty scale. This alteration is needed for polytomous models because in these models there is no single difficulty parameter but rather multiple step values or thresholds are used to describe the probability of responding in a given category. For this reason three points are used. These three points are the estimated trait level, estimated trait level minus .10, and estimated trait level plus .10. One of the items is then randomly selected for administration.

By doing this, a measure of variability is introduced into the selection of items during the test. This means that persons with the same trait estimate may not receive the exact same item. Davis and Dodd (2001) found good control over item exposure and overlap rates with a slight decrease in measurement accuracy using the modified within 0.10 logits procedure. By using this method the percent of the item pool not used was reduced by 44% - 45% in comparison to maximum information item selection.

Progressive Procedure

The progressive method of exposure control, first proposed by Revuelta in 1995 (Revuelta & Ponsoda, 1998), adds a random component to the maximum information method. A weight is computed for every item available for administration based upon the mathematical formula:

$$W_i = (1-s)R_i + sI_i\,,$$

where s is the serial position of the item in the test, I is the item information at the estimated trait level, and R is a random number from the uniform distribution. The item with the largest weight is then selected for administration. This will result in item information being weighted less in the initial stages of the test, and more in the latter stages of the test.

In this way during the beginning of the test where examinees are likely to have similar estimated trait levels, the item selected will not always be the most informative item, but an item from among the more informative items. In the latter stages of the test, information plays a greater part in the weight and therefore the item selected will be from among the most informative items. Revuelta and Ponsoda reported a decrease in the number of items that were not administered over 2000 simulated testing sessions, but this method did not control item over-exposure very well (maximum exposure rate = .641). The progressive method did not show a considerable decrease in test precision (Revuelta & Ponsoda, 1998).

Conditional Procedures

In the conditional strategies the probability that an item will be selected for administration is conditioned upon a given criterion. Since there is a dependence upon the given criterion it can be assured that through these strategies item exposure can be controlled. However, before operational use can begin it is necessary to conduct simulations in order to attain the exposure control parameters. Additionally these simulations may need to be periodically rerun as test conditions change. Although none of these procedures are included in the present research they are included here for the sake of providing a complete picture of exposure control strategies.

Sympson-Hetter Procedure

The Sympson-Hetter procedure controls exposure rates by ensuring that the exposure rates for items never go beyond a prespecified maximum (r). In order to do this, exposure rates for all items must be estimated using simulated CATs. The exposure control parameter (k_i) is always between 0 and 1.0. This parameter is then used in live testing to constrain the probability that an item will be selected for administration. When an item is selected for administration k_i is compared to random number from the uniform distribution. The item is only administered if k_i is greater than the random number. If it is not, that item is no longer available for administration in that CAT. If rejected, the next most informative item is selected and compared to a random number from the uniform distribution.

In order to use this procedure one needs, beforehand, to conduct simulations to set the k_i parameters. To do this all k_i values are first set to 1.0. This value indicates that should an item be selected it will also be administered. A simulated CAT is conducted with simulees with a known trait level. As each item is selected it is compared to a random uniform number and its k_i value. To attain the proportion of times that an item has been selected and administered the following formulas are used:

$$P(S) = NS/NE,$$
$$P(A) = NA/NE,$$

where *NS* is the number of times the item has been selected, *NA* is the number of times the item has been administered, and *NE* is the total number of simulees. The probability of selection is compared to the target exposure rate (r) and the following rules are used to make adjustments:

$$\begin{aligned}&\text{If } P(S) > r, \text{ then } \mathrm{k}_i = r / P(S),\\ &\text{If } P(S) \leq r, \text{ then } \mathrm{k}_i = 1.0.\end{aligned}$$

This procedure is repeated until the maximum probability of an item being administered is slightly above r . Each time the process is repeated the k_i values from the previous iteration are used.

Conditional Sympson-Hetter Procedure

The conditional Sympson-Hetter procedure is conducted in much the same manner as the Sympson-Hetter is. The alteration in this procedure is that the frequency of item administration is tallied separately for each trait level. The simulations to set k_i are performed individually at each trait level. One then is able to construct a matrix of items by theta values. The intersection of column and row gives the conditional exposure control parameters.

Stocking and Lewis Multinomial Procedure

This procedure also requires the establishment of k_i parameters. Item selection takes place based upon a distribution of multinomial probabilities. The multinomial probability is the probability that an item is selected and administered and that all previous items were rejected. The multinomial probabilities are calculated using the following formulas:

$$\begin{aligned}
&k_1 = P_1(A/S),\\
&k_2 = \left[1 - P_1(A/S)\right] * P_2(A/S),\\
&k_3 = \left[1 - P_1(A/S)\right] * \left[1 - P_2(A/S)\right] * P_3(A/S), etc.,
\end{aligned}$$

where k_1 is the probability that the first item is selected and administered, k_2 is the probability that the first item was rejected and the second item has been selected and administered, and k_3 is the probability that the first and second items were rejected and that the third is selected and administered. By summing all $P_i(A\,|\,S)$ values a multinomial distribution is created. This cumulative distribution of probabilities is compared to a random uniform number. Item administration is then decided upon the location in the cumulative distribution which corresponds to the random uniform number, with all rejected items being blocked from administration.

Davey-Parshall Procedure

The Davey-Parshall procedure (1995) establishes an exposure control parameter for pairs or groups of items that are often administered together. Unlike the Sympson-Hetter procedure which conditions upon ability, the Davey-Parshall procedure conditions the probability of exposure on the items that have been previously administered in the test. Therefore if one item in a paring is administered during a test the other item in the pairing is barred from selection for administration. This procedure uses an *n x* n table of exposure parameters where *n* is the number of items in the item pool. In this matrix the diagonal elements are a measure of the popularity of a single item similar to those used in the Sympson-Hetter procedure. The off-diagonal elements are a measure of the popularity of two items appearing together. This table is used during live testing to calculate the conditional exposure parameters. High values for either element near 1 indicate that the item is less attractive and therefore less susceptible to over exposure while values closer to zero indicate the pair of items are frequently administered together.

To establish the exposure table all elements of the table are initially set to1. This indicates that if an item is selected, it will be administered. Simulations are then conducted to determine the number of times that any item appears alone and as part of a pair. After each set of simulated CATs is done, adjustments are made to both the diagonal and off-diagonal elements. For diagonal elements any item that appears more frequently than an established maximum exposure rate is adjusted lower and those appearing less frequently than the target exposure rate is adjusted higher (to the maximum value of 1). Davey and Parshall (1995) suggest this being done by multiplying the elements by .95 or 1.04 depending on the needed adjustment. For off-diagonal elements, a modified chi-squared test is performed to determine if the pair of items is appearing more often then would be expected by chance (see Davey and Parshall (1995) for details on the modified chi-squared test). If the pair is appearing more often than would be expected by chance the element is adjusted down, and if not it is adjusted upwards.

This procedure begins by selecting an item based upon the item selection procedure being used. The diagonal element from the table corresponding to this item is extracted along with the off-diagonal elements for any items that have already been administered. The conditional probability of administering the item given it is selected can then be computed as the product of the diagonal element and the average of the off-diagonal elements of previously administered items.

Combinatorial Procedures

The combinatorial procedures described in this section are the synthesis of randomization strategies and conditional strategies. Each of the following exposure control procedures incorporate elements of conditional strategies and marry them to randomization strategies. This is done in an effort to incorporate the beneficial aspects of both strategies, while avoiding the complexities involved with conditional procedures. These combinatorial procedures include: the restricted maximum information, progressive restricted, restricted modified within .10 logits, and the restricted randomesque procedures. As will be seen in the subsequent sections each of these three procedures conditions the administration of an item on a given maximum exposure rate. Each will also, through differing methods, incorporate the use of a random component to their item selection algorithm.

Restricted Maximum Information Procedure

In this procedure a maximum exposure rate for items is defined (*k*) so that items with an exposure rate above this maximum will not be considered for administration (Revuelta and Ponsoda, 1998). If we define the number of times an item has been administered in previous tests as a and the number of previous tests administered as t then the exposure rate (*k*) for item i will equal *a/t.* The subset of items available for selection and administration in the subsequent test will then consist only of items whose exposure rate is below the predefined limit (*k*).

In this way, the subset of items available for selection and administration will vary from test to test. As more tests are administered the exposure rate for items above k

will decrease making the items available once again for selection and administration. At every item administration the CAT will examine all items in the item pool and include those whose exposure rate is less than or equal to *k* into a subset for consideration in administration. Once an item's exposure rate exceeds *k* it is no longer available for selection into the subset of items. The longer the item is not available for selection the lower that item's exposure rate becomes. Once it drops below *k* it will be available for selection into the item pool subset. As more tests are administered the exposure rate will decrease.

Progressive Restricted Procedure

In the progressive restricted procedure (Revuelta & Ponsoda, 1998) items are selected for administration via the progressive method described earlier. As in the restricted maximum information method exposure rate is used to define a subset of items available for selection and administration in a given test. Exposure rates for each item are calculated using the same method as described previously. Items whose exposure rate exceeds the predetermined maximum exposure rate are not available for selection in the subset of items for the subsequent test. As more tests are administered the exposure rate will drop below the predetermined maximum exposure rate thus making it once again available for selection and administration.

Restricted Modified within .10 logits Procedure

The modified within .10 logits procedure has shown great promise in controlling item exposure and overlap rates with greater pool usage. There is, however, no guarantee that the strategy would constrain the item exposure to desirable levels. For that reason a modification, suggested by Dr. Barbara Dodd (personal communication, November 1, 2007), will be made to the procedure in the present research that will exclude items above a predetermined maximum from being considered for administration. Using the same logic utilized by the restricted maximum information procedure items are selected for administration using the modified within .10 logits procedure, however, the exposure rates for items will be used to create a subset of items from the item bank to be used for selection and administration in the same manner as the restricted maximum information and progressive restricted procedures.

Restricted Randomesque Procedure

Utilizing the same logic of extending the modified within .10 logits procedure the same extension is proposed for the randomesque procedure. Here again, the impetus for this extension is the capturing of the benefits that have been demonstrated by the randomesque procedure in previous research while implementing a constraint to limit the maximum exposure rate produced by the randomesque procedure. In this procedure, item selection is conditioned upon a predetermined maximum exposure rate. Items that meet this condition are then selected for administration via the randomesque item selection procedure.

Stratification Procedures

As Chang and Ying (1999) note, the idea of stratification in item response theory has been used to assess differential item functioning. In the context of item selection, stratification procedures were proposed as a way to control overexposure and ensure good utilization of the item pool. In these methods the entire item pool is stratified based upon a given variable. This variable can be the discrimination, difficulty, or even content area. The correlation observed between discrimination and information means that with maximum information item selection more discriminating items with the most information are going to be administered the most. Stratification procedures were designed with the expressed goal of countering this phenomenon. Various stratification procedures are here discussed.

a-stratified Procedure

Chang and Ying (1999) developed this procedure based upon the observations that, for the 3-PL model, items with larger *a* and smaller *c* values are very likely to be administered with maximum information item selection and if discrimination is considered separately from Fisher information more efficient use of items with high *a* values can be seen. They also note that while higher information can be attained from items with higher discrimination values the benefits are attenuated by the fact that estimates of true ability are used in selecting items rather then true, unknown, ability. They argue that by stratifying the item pool by the discrimination parameters and selecting items for administration from within these strata more even exposure rates will be seen because the frequency in which items with various *a* values will be selected is equal. This results in raising the exposure rates of items that would be underutilized and lowering the exposure rates of items that would be over exposed using maximum information item selection.

This procedure starts off by dividing the item bank into *K* strata by the discrimination parameters. The first stratum would contain the items with the lowest discrimination parameters and the last stratum should contain the items with the highest discrimination parameters. The number of strata may be small for item pools with similar *a* parameters or many if the range of the *a* parameters is large. The size of the strata should be approximately equal to the quotient of the test length divided by the number of strata. The strata should contain an equal number of items with the exception of the first which may be large to ensure proper theta estimation. The test is then partitioned into *K* stages. Within each k^{th} stage n_k items are administered based upon the proximity of the *b* values to estimated theta. This process is repeated until all stages of the test have been administered.

a-stratified Procedure with b-blocking

This procedure was designed as a refinement of the *a*-stratified design intended to deal with the procedures potential pitfalls when the items in the strata can not match *b* values with estimated theta. This procedure attempts to ensure an adequate distribution of difficulty parameters in each stratum by blocking on *b* parameters. As noted by its developers, Chang, Qian, and Ying (2001), this distribution is important because CATs strive to provide a good match between the estimated trait level and an item's difficulty parameter. The a-stratified design assumes that the distribution of *b* values is not affected by the stratification of the item pool. Chang et al. note that this is rarely the case. In this procedure item pools are sorted in ascending fashion and subsequently divided into M blocks based on their b parameters. The blocks are sorted such that the first block contains the items with lowest b values and the last has the items with the largest values. Chang et al. reported that this procedure might improve estimation at extreme values of theta and provide more efficient pool utilization.

Chang et al. (2001) note that the number of items in each block should be equal or differ by at most one item if the number of items in the item pool is not evenly divisible. Following this, the *M* blocks are partitioned into *K* strata according to the item's discrimination values. This means that each *b*-block stratum is partitioned such that the first *K* stratum contains items with the lowest *a* values and the last stratum contains items with the highest *a* values. After this, all items within the discrimination strata (K_i) of all of the *M* blocks are combined. Once this has been done an item pool of *K* strata has been produced each containing the items from all *M* blocks merged by their *K* strata assignment. The test is now divided into *K* stages. Item selection during testing is conducted in *K* stages. The test proceeds from the first to last stage using the item that minimizes the discrepancy between estimated theta level and item difficulty. The appropriate number of items is administered within each stage of the test before proceeding to subsequent stages. The number of items per strata and the number of items administered per strata are determined as they are in the *a*-stratified procedure.

a-stratified Procedure with Content Blocking

Ying and Chang (2003) proposed the *a*-stratified method with content blocking as an extension of the *a*-stratified design with *b*-blocking item selection procedure. In this method one first decides how many strata are to be used (K). This decision is reached in the same way as is done in the *a*-stratified procedure. Following this the item pool is divided into G groups based upon the content domains. The items are sorted into each group based upon their difficulty parameters from smallest to largest. The items in group g are partitioned into P_g based upon their difficulty parameters. This is done by placing items with the lowest b values into the first block and those with the highest b values in the last block. One proceeds to sort the items in every P_g block in ascending order of the discrimination parameters. Following this the sorted items in the blocks are placed in the K^{th} strata based upon their discrimination parameters. Through this the item with the highest discrimination parameter is placed into the first stratum, the item with the second highest a parameter is assigned to the second stratum. This is done until the item with the lowest a parameter is assigned to the K^{th} stratum. Following this all items that were assigned to stratum one are grouped together, all items assigned to stratum two are grouped, etc. Finally the test is partitioned into K stages. In the K^{th} stage, n_k items are selected based upon the similarity of the item difficulty and the estimated trait level of the examinee.

Enhanced a-stratified Procedure

In this procedure two procedures are merged to address a problem with the *a*-stratified design. As Leung, Chang, & Hau (2002) note the *a*-stratified design can not guarantee a maximum exposure rate, especially when the ratio of item pool size to test length is small. This procedure requires the setting of a target maximum exposure rate (*r*). A common value for this exposure rate is 0.20. The item pool is subsequently partitioned into *k* subpools with each test divided into *k* stages as called for by the *a*-stratified design.

Simulations are conducted as called for with the Sympson-Hetter procedure to set the exposure control parameters for all items in the item pool. The process for administering an item for administration is based upon how close the difficulty parameter is to the estimated theta. The item that minimizes this discrepancy is selected and its exposure control parameter is compared to a random number from the uniform distribution.

If the exposure control parameter for the selected item is larger than the random number that item is administered. If, however, the exposure control parameter for the item is smaller than the random number from the uniform distribution the item with the next smallest discrepancy is selected and its exposure control parameter is compared to a new random number from the uniform distribution and evaluated for administration. If need be this process is repeated until an item with an exposure control parameter greater than the random number from the uniform distribution is found. This process is continued until the prespecified number of items for that block has been administered. At this point, the subsequent block is administered following the same item selection rules. This continues until all blocks have been administered.

Exposure Control Research with Polytomous IRT Models

In practice each of the exposure control strategies introduced in the previous sections will produce various desirable and undesirable results. As such, it is necessary to examine how each of the procedures functions when implemented. In this section the reader is treated to a brief survey of relevant research pertaining to exposure control strategies as they have been applied to the partial credit model and the generalized partial credit model. Research involving the partial credit model is addressed first followed by the research involving the generalized partial credit model.

Davis, Pastor, Dodd, Chiang, and Fitzpatrick (2003) investigated the use of the Sympson-Hetter procedure, rotated content balancing, test length, item pool size, and dimensionality in the partial credit model. The Sympson-Hetter procedure was compared to the maximum information, no exposure control, procedure. Rotated content balancing was investigated to see the effect of implementing content constraints as opposed to not using content balancing. To investigate test length stopping rules of 15 items or a standard error of 0.30 were implemented. Item pools of 60, 120, and 240 items were investigated. To investigate the effects of dimensionality to datasets were utilized. Each dataset's first dimension was a general mathematics dimension with either 80% or 72% common variance.

It was determined that none of the variable investigated affected the distribution of estimated theta when compared to the known theta distribution. Correlation of known to estimated theta, bias, and root mean squared error were not affected by any of the variables included. As such the use of the Sympson-Hetter procedure as a method of exposure control, content balancing, variable or fixed length tests, multiple item pools or variances in dimensionality did not affect measurement precision.

An effect on average exposure rate was observed for the variable length tests when item pool increased. As item pool increased, average exposure rate decreased for the variable length test conditions. Average exposure rate was not however effected by the implementation of an exposure control procedure or the use of content balancing. The standard deviation of exposure rates did decrease when the Sympson-Hetter was implemented and also when item pool size increased. The percent of the item pool that was not administered increased as item pool size increased and when variable length tests were used while the implementation of the Sympson-Hetter procedure and the use of content balancing reduced this percentage.

In the larger 120 and 240 item pools the number of items with exposure rates above 0.30 was relatively low with an increase seen in the smaller 60 item pool. The implementation of the Sympson-Hetter procedure had little effect on the number of items with an exposure rate above 0.30. Increasing the item pool size and implementation of the Sympson-Hetter procedure helped to decrease the amount of overall average overlap and overlap for examines whose known theta values differed by 2 logits or less. The use of a variable length test also helped reduce these overlap rates while the implementation of content balancing had no effect on overlap rates.

This study supports the conclusion that the use of the Sympson-Hetter exposure control procedure will not significantly impact measurement precision. This study also found, however, that this procedure did not significantly decrease average item exposure rate. The authors note that the amount of gain seen when implementing the Sympson-Hetter procedure was also produced by increasing item pool size. This fact in conjunction with the complexities of implementing the Sympson-Hetter procedure gave the authors reason to conclude that the gains seen did not lend support for the justification of utilizing this procedure.

Davis and Dodd (2003) investigated the performance of the modified within .10 logits procedure with an item group size of six and the computerized adaptive sequential testing (CAST) system in comparison to maximum information item selection and selecting item for administration randomly with the partial credit model. CAST systems (Luecht & Nungester, 1998) control exposure rates by the preconstruction of adaptive test forms. Items are grouped into modules which are then arranged in multistage panels. Modules within each stage are divided by item difficulty. Ability estimation is conducted once all items in the module have been administered. (For more information the reader is referred to Luecht & Nungester, 1998.)

The modified within .10 logits procedure and the CAST system produced a distribution of estimated theta that was quite similar to, yet slightly above, the known theta distribution. The no exposure control condition produced 27 nonconvergent cases while the random condition produced 114 cases, the modified within .10 logits procedure produced 44 cases, and the CAST condition produced 0 cases. The modified within .10 logits and CAST conditions resulted in identical standard error values.

Random item selection resulted in the lowest correlation between known and estimated theta (0.93) and the largest values for root mean squared error (0.41) and average absolute difference (0.30). The no exposure control condition, modified within .10 logits procedure, and CAST conditions produced correlations between known and estimated theta ranging from 0.95 to 0.96. The no exposure control condition produced a bias of 0.00 with the CAST condition producing a bias of -0.02 and both the random and modified within .10 logits conditions producing bias of -0.04. The no exposure control condition also yielded the lowest values of root mean squared error and average absolute difference (0.31 & 0.24). The next lowest values of these statistics were produced by the modified within .10 logits condition (0.33 & 0.25) followed by the CAST condition (0.35

& 0.27). On these indices of measurement precision, the modified within .10 logits procedure produced slightly better results than the CAST condition.

For the indices of test security the authors provided two sets of results. This was done because of the low frequency of 8 and 10 item passages when content and passage type balancing was used. When all item were included in the analysis the no exposure control condition resulted in an a standard deviation of exposure rates of 0.101, maximum exposure rate of 0.513, and 62% of it's item pool not being administered. The modified within .10 logits procedure resulted in a standard deviation of exposure rates of 0.063, maximum exposure rate of 0.444, and failed to administer 18% of its item pool. The CAST condition resulted in the lowest standard deviation of exposure rates (0.044) and maximum exposure rate (0.165). The random and CAST conditions both utilized all of the items in their item banks.

When the 8 and 10 item passages were removed from the analysis the no exposure control condition resulted in a standard deviation of exposure rates of 0.088, maximum exposure rate of 0.474, and failed to administer 66% of its item pool. Again, the random and CAST conditions both utilized all of the items in their item banks. In this analysis the random item selection condition presented the best exposure rate indices. This condition resulted in a standard deviation of exposure rates of 0.019 and a maximum exposure rate of 0.104. The CAST condition produced the next most desirable results with a standard deviation of exposure rates of 0.036 and a maximum exposure rate of 0.165. The modified within .10 logits procedure resulted in a standard deviation of exposure rates of 0.042, a maximum exposure rate of 0.191, and failed to administer 21% of its item pool.

In regards to exposure rate indices the CAST system outperformed the modified within .10 logits condition.

When all items were included in the analysis the CAST condition produced the lowest overall average overlap (9%), average overlap between examinees with similar abilities (10%), and average overlap for examinees with different abilities (6%). The no exposure control condition produced overlap rates of 26%, 30%, and 7%. The random item selection condition resulted in an average overall overlap, overlap rate for examinees with similar abilities, and overlap rate for examinees with different abilities of 10%. The modified within .10 logits condition also showed significant improvement over the no exposure control condition in average overall overlap (13%) and overlap rate for examinees with similar abilities (14%) but a higher overlap rate for examinees with different abilities (9%).

When the 8 and 10 item passages were removed from the analysis the no exposure control condition produced the highest overlap rates for average overall overlap (17%), overlap rate for examinees with similar abilities (20%), and overlap rate for examinees with different abilities (4%). The random item selection condition resulted in an average overall overlap, overlap rate for examinees with similar abilities, and overlap rate for examinees with different abilities of 3%. The modified within .10 logits condition produced an average overlap rate of 6%, average overlap between examinees with similar abilities of 7%, and average overlap for examinees with different abilities of 2%. The CAST condition produced an average overlap rate of 5%, average overlap between examinees with similar abilities of 6%, and average overlap for examinees with different abilities of 3%. While the modified within .10 logits condition demonstrated an improvement over the no exposure control condition, the CAST condition resulted in lower overlap rates regardless of the inclusion or exclusion of the 8 and 10 item passages.

While random item selection would logically aid in lowering exposure rates it was somewhat surprising to find that the measurement precision was not more effected. The authors suggest that this may have been due to the greater amount of psychometric information provided by polytomous item pools or by the ability of models based on the Rash model not be as effected by suboptimal items. The authors conclude that while both the modified within .10 procedure and CAST system provided desirable results the CAST system appeared to display a slight edge over the modified within .10 logits procedure. Not withstanding this, the modified within .10 logits procedure was able to produce good test security with only a minimal loss in measurement precision when compared to the maximum information, no exposure control, condition.

Boyd (2003) examined the efficiency of the randomesque, progressive restricted, modified within .10 logits, and Sympson-Hetter procedures with the partial credit model. The randomesque and modified within .10 logits procedures were both implemented with an item group size of six. In both the progressive restricted and Sympson-Hetter procedures the maximum exposure rate parameters used were .20 and .30. Boyd encountered difficulties in the simulations with the progressive restricted procedure with a maximum exposure rate parameter of .20.

In eight of the ten replications performed with the progressive restricted procedure with a maximum exposure rate of .20 the CATs were not successfully completed. Boyd notes that the nature of the item pool resulted in a situation where the progressive restricted algorithm had no item groups to administer. That is to say, there occurred in these eight replications a situation in which no item group had an exposure rate below the prespecified .20 maximum exposure rate. As such, it was impossible to continue the CAT since no item group satisfied the progressive restricted algorithm. Therefore, results involving the progressive restricted procedure with a maximum exposure rate of .20 are averages of the two successful replications while the other conditions represent the averages of the ten successful replications of each condition.

In reference to the distribution of the estimated theta values across conditions it can be seen that all conditions performed roughly equivalently. Each condition yielded a grand mean of estimated thetas approximately equal to that of the distribution of known thetas. Each condition yielded a mean of the standard deviation of estimated thetas slightly below the mean of the standard deviation of known theta values. Boyd notes, however, that this is not surprising as it is an artifact of the estimation procedure used. Given these findings it can be concluded that the estimated theta values were approximately normally distributed which accurately portrays the normal distribution of the known theta values.

Moving on to indices of measurement precision, we first examine the standard error of the estimated theta values. While minor variances were seen between conditions, these differences were small enough to conclude that all conditions performed roughly equivalently. In terms of the recovery of known theta values all conditions once again performed equivalently. In terms of bias, standardized difference between means, average absolute difference, root mean squared error, and standardized root mean squared difference all conditions again preformed equivalently. From these indices of measurement precision it is justifiable to conclude that no condition significantly decreased measurement precision.

The next indices concerning efficiency of the item selection strategies centers upon exposure control. In terms of maximum exposure rates each condition investigated performed very well. Both the randomesque and modified within .10 logits procedures resulted in maximum exposure rates around 0.19. The Sympson-Hetter and progressive restricted procedures yielded maximum exposure rates that were consonant with the maximum exposure rate parameter utilized for the condition. The progressive restricted procedure with a maximum exposure rate of 0.20 produced the most even item usage with the randomesque, modified within .10 logits, and progressive restricted procedures with a maximum exposure rate of 0.30 following closely behind it. The Sympson-Hetter procedures outperformed the maximum information condition in this regard but both were outperformed by the previously mentioned conditions.

Examining the frequency distributions of exposure rates it was found that the exposure control conditions, i.e. excluding the maximum information condition, yielded exposure rates below 0.20. Most impressively, both progressive restricted procedures resulted in every item in the item pool being administered. The randomesque and modified within .10 logits procedures both resulted in a modest 28% of the item pool not being administered. The Sympson-Hetter procedures with a maximum exposure rate of 0.20 and 0.30 resulted in an unsettlingly large percentage of the item pool having never been administered, 52% and 57% respectively.

The last set of indices of efficacy for the exposure control procedures investigated is item overlap. In the case of item overlap, statistics are reported for the overall number, and percentage, of items that examinees have in common as well as the number of items, and percentage, that examinees who share similar and different abilities have in common. Examinees were first defined to have similar abilities if they differed by two logits or less and different if they differed by more than two logits. This definition was then tightened to one logit in either case to focus on examinees with similar abilities.

Using the 2 logit definition, the maximum information condition yielded overall overlap rates of three item groups, approximately two item groups being shared among examinees with similar abilities, and less than one item group being shared among examinees with different abilities. Average overall overlap rates for the investigated conditions revealed a difference of, at most, 1 item group. When the broader 2 logit definition was used the randomesque, modified within .10 logits, and both progressive restricted procedure resulted in approximately 1 item being shared by examinees with similar abilities, while the Sympson-Hetter procedures resulted in one to two items being shared. Examinees with different abilities, as defined by a difference of two logits or more, were likely to have less than one item group in common.

When the stricter definition of similar and different abilities was used, the randomesque, modified within .10 logits, and both progressive restricted procedures maintained overlap rates seen with the broader two logit definition. The maximum information condition now yielded an overlap of three item groups for examinees with less than one logit difference in known abilities. The Sympson-Hetter procedures however increased their overlap rates for examinees with similar abilities. These procedures now resulted in approximately two item groups being shared among examinees with similar abilities.

In this research study it seems to be the case that the imposition of an exposure rate constraint of 0.20 in the progressive restricted procedure was overly restrictive. When an exposure rate of constraint all ten replications were successfully completed while only two replications were successful with the more restricted 0.20 constraint. This study found that the exposure control procedures investigated did not significantly impact measurement precision. In terms of test security a clear pattern emerged. With the exception of pool use the randomesque procedure produced the best protection of test security. This was followed by the modified within .10 logits procedure, the progressive restricted procedure with an exposure constraint of 0.20, the progressive restricted procedure with an exposure constraint of 0.30, the Sympson-Hetter with a target exposure rate 0.20, and the Sympson-Hetter with a target exposure rate 0.30.

Davis (2002) evaluated the Sympson-Hetter, conditional Sympson-Hetter, randomesque, and modified within .10 logits procedures when implemented with the partial credit model. The randomesque and modified within .10 logits procedures were both implemented with an item group size of three and six. In all conditions employed in the simulation study the distribution of estimated theta values were normally distributed. The Sympson-Hetter procedure yielded a standard error value equally low as the maximum information condition (0.27). The three item variants of the randomesque and modified within .10 logits procedure as well as the conditional Sympson-Hetter procedure yielded slightly higher standard error values (0.28), and the six item variants of the randomesque and modified within .10 logits procedure yielded the highest standard error values, 0.29 and 0.30 respectively. These two conditions also produced the highest number of nonconvergent cases, 19 and 20, when compared to all other conditions, 7 - 9.

In terms of correlation of known theta to estimated theta, bias, standardized difference between means, root mean squared error, standardized root mean squared error, and average absolute difference all conditions performed equally well. Given these results it can be concluded that all conditions investigated preformed equally well in terms of measurement precision.

In terms of exposure control the randomization procedures with an item group size of six and the conditional Sympson-Hetter clearly performed the best. The maximum information condition produced the highest maximum exposure rate (0.655) followed by the within .10 logits procedure with an item group size of three (0.535), the randomesque procedure with an item group size of three (0.503), the Sympson-Hetter procedure ((0.434), the within .10 logits procedure with an item group size of six (0.398), the randomesque procedure with an item group size of six (0.396), and finally the conditional Sympson-Hetter procedure (0.395). The randomization procedures with an item group size of six and the conditional Sympson-Hetter procedure yielded the most even item usage as indicated by the standard deviation of exposure rates. The values for these conditions ranged from 0.8 to 0.11 while the other conditions investigated ranged from 0.123 to 0.147, excluding the maximum information condition (0.167).

The within .10 logits procedure with an item group size of six, the randomesque procedure with an item group size of six, and the conditional Sympson-Hetter procedure also produced the lowest percentage of pool never administered, 8%, 8%, and 15% respectively. This was significantly lower than the other investigated procedures which ranged form 20% to 31%, excluding the maximum information condition (37%). It is worth noting that in the case of percent of pool not administered and standard deviation of exposure rates, the randomization procedures with an item group size of three, while outperformed by their six item variants and the conditional Sympson-Hetter condition, performed better than the Sympson-Hetter procedure.

In terms of item overlap the randomization procedures, when implemented with an item group size of six, outperformed their three item variants as well as both conditional procedures investigated. As a baseline comparison, the maximum information condition yielded an overall average overlap rate of 34% and an average overlap rate for simulees with similar abilities of 39%. The randomization procedures with an item group size of six produced an overall average overlap rate of 20% and an average overlap rate for simulees with similar abilities of 22%. The conditional Sympson-Hetter procedure produced the next lowest levels of overlap (22% and 25%) followed by the randomesque procedure with an item group size of three (24% and 28%), the within .10 logits procedure with an item group size of three (25% and 28%), and finally the Sympson-Hetter procedure (29% and 34%).

It is clear that the three best performing procedures in this study were the randomization procedures with an item group size of six and the conditional Sympson-Hetter. On several of the measures just described the randomization procedures with an item group size of six outperformed the conditional Sympson-Hetter. Given this fact and the fact that the randomization procedures are far less complex to implement and maintain once in use these procedures emerge as the best performing exposure control procedures investigated by Davis (2002).

Davis and Dodd (2005) studied the randomesque, modified within .10 logits, Sympson-Hetter, and conditional Sympson-Hetter exposure control procedures with the partial credit model. The randomization strategies were explored using item group sizes of three and six. The conditional procedures investigated were implemented with target exposure rates of 0.39. In this study the randomesque procedure with an item group size of three and the conditional Sympson-Hetter procedure produced the same number of nonconvergent cases as the no exposure control condition. The three item variant of the randomesque procedure resulted in 8 nonconvergent cases. The six item variants of the randomesque and modified within .10 logits procedures and the Sympson-Hetter procedure resulted in a slightly higher number of nonconvergent cases, ranging from 12 to 13.

The correlation between known and estimated theta was equivalent in all conditions examined. Root mean squared error was also very similar, ranging from 0.29 to 0.30. The six item variant of the modified within .10 logits procedure resulted in the smallest level of bias (-0.03) followed closely by the six item variant of the randomesque procedure (-0.04) and the conditional Sympson-Hetter procedure (-0.04). The three item variants of the randomesque and modified within .10 logits procedures as well as the Sympson-Hetter procedure resulted in bias values equal to the no exposure control condition (-0.05).

On the measures of test security included, the randomesque and modified within .10 logits procedures with an item group size of six as well as the conditional Sympson-Hetter procedures provided superior performance. The conditional Sympson-Hetter produced the lowest maximum exposure rate (0.348) and lowest standard deviation of exposure rates (0.093). This was followed by the randomesque procedure with an item groups size of six (0.374 & 0.095) and then closely by the modified within .10 logits procedure with an item group size of six (0.391 & 0.096). The conditional Sympson-Hetter procedure resulted in 10% of the item bank not being administered while the randomesque and modified within .10 logits procedures with an item group size of six resulted in only 9% of the item bank not being administered.

In terms of item overlap these same three procedures provided the lowest overall average item overlap and average overlap for simulees whose known theta differed by 2 logits or less. The conditional Sympson-Hetter and the six item variants of the randomesque and modified within .10 logits procedure resulted in an overall average item overlap of 20%. For simulees whose known theta values differed 2 logits or less these three conditions resulted in an average overlap of 22%.

Davis and Dodd (2005) also found that when item group size increased from three to six, in the randomization procedures, test security increased while measurement precision was only minimally impacted. This study provided further support for the use of randomization procedures over the more complicated, yet widely accepted, Sympson-Hetter procedure. Performance by these procedures was, in fact comparable to the conditional Sympson-Hetter. It was also noted that the maximum exposure rate of 0.39 for the conditional procedure was chosen by necessity because of convergence failures with lower exposure rate targets of 0.19 and 0.29. This highlights one of the advantages of the randomization procedures over conditional procedures. As is often noted with the conditional procedures, implementation of these procedures is complicated and often mitigates it desirability in operational use.

Davis (2004) examined the performance of 8 item selection strategies using the generalized partial credit model. The item selection procedures examined were the maximum information, randomesque, modified within .10 logits, Sympson-Hetter, *a*-stratified, and enhanced *a*-stratified procedures. In the randomesque and modified within .10 logits procedure two variants were used. Each was tested using both three and six items as their item group size. Davis' research demonstrated no appreciable difference in the theta estimates yielded. In this study the three item variants of the randomesque and modified within .10 logits procedures yielded the lowest average standard error values, being outperformed in this regard by only the maximum information condition. The six item variants of these same procedures yielded average standard error values that were comparable to the more complex conditional and stratified procedures investigated. In this regard the less complex randomization procedures performed as well or better than all of the more complex procedures.

In addition, the randomization procedures resulted in roughly equivalent correlation and bias values as the more complex conditional and stratified procedures. The root mean squared error values for all conditions except the enhanced a-stratified condition were also roughly equivalent, with values ranging from 0.30 to 0.33. The six item variants of the randomization procedures, along with the conditional Sympson-Hetter condition, provided the lowest exposure rate standard deviation, which demonstrated the most even item usage, and the lowest percent of pool not administered.

In this study the conditional procedures yielded the lowest maximum exposure rate. This finding needs to be qualified by the trend the randomization procedures demonstrated when larger item size variants were used. As the item group size was increased from three to six the maximum exposure rate decreased. This may suggest that the use of an item group size larger than six may further reduce the maximum exposure rate. In regard to both overall average overlap and overlap rates for examinees with similar abilities, the six item variants of the randomization procedures as well as the conditional Sympson-Hetter demonstrated the lowest overlap rates. This research study found a trend for overlap rates, exposure rate standard deviation, and maximum exposure rate to decrease as item group size was increased.

The Sympson-Hetter and conditional Sympson-Hetter procedures were quite complex to implement. Additionally when the targeted exposure control rate was not achieved, the simulation had to be rerun. Davis (2004) also noted that the programming requirements for these procedures were considerable. These two procedures resulted in observed exposure rates slightly above the targeted rate. While a reduction in overlap rates and percent of the pool not used when compared to the no exposure control condition was seen, their performance was less than desirable when compared to other options.

Pastor, Dodd, and Chang (2002) examined the efficacy of the *a*-stratified design, Sympson-Hetter, enhanced *a*-stratified design, conditional Sympson-Hetter, and the conditional enhanced *a*-stratified design with item banks of sixty items and one hundred items using the generalized partial credit model. Looking first at the one hundred item pool we can examine the performance of the stratified design in comparison to the more complex Sympson-Hetter, enhanced stratified design, conditional Sympson-Hetter, and the conditional enhanced stratified design. In regards to pool utilization the stratified design reduced the percentage of pool not utilized from 44% in the maximum information condition to the same level as the Sympson-Hetter (28%). While both were outperformed in this regard by the enhanced stratified design (13%), conditional

Sympson-Hetter (3%), and conditional enhanced stratified design (3%) the fact that it performed as well as the widely accepted Sympson-Hetter should not be overlooked.

In regards to the standard deviation of exposure rates the stratified design (0.19) resulted in slightly more even item usage than the maximum information item selection (0.22) but was outperformed by the Sympson-Hetter (0.14), enhanced stratified design (0.13), conditional Sympson-Hetter (0.09), and conditional enhanced stratified design (0.08). When examining the overall average overlap and the overlap for examinees with similar abilities the authors found that the stratified design reduced the percentage of overlap in both cases but was outperformed by the more complex conditional procedures.

In this larger item pool the correlation of estimated theta to known theta values for the stratified design, Sympson-Hetter, and enhanced stratified design were very similar and had standard errors of .30, .31, and .31 respectively. These were slightly higher than the average standard error of the maximum information condition (.28). The more complex procedures yielded similar theta estimates with higher average standard error values of .34 for both the conditional Sympson-Hetter and conditional enhanced stratified design. The RMSE statistics increased from the least restrictive no exposure condition (.32) to the more restrictive conditional enhanced stratified design (.38) demonstrating the relative loss of measurement precision accompanying more restrictive procedures.

In the 60 item pool the stratified design again outperformed the maximum information procedure. It, in fact, reduced the percent of pool not administered from 30% to 13%, a reduction of over half. It was, however, greatly outperformed by the Sympson-Hetter (2%), enhanced stratified design (2%), conditional Sympson-Hetter (0%), and the conditional enhanced stratified design (0%). The stratified design reduced the standard deviation of exposure rates from .28 for the maximum information condition to .24. The Sympson-Hetter, enhanced stratified design, conditional Sympson-Hetter, and conditional enhanced stratified design yielded significantly lower standard deviations of .10, .08, .08, and .07 respectively. In this case, with the smaller item bank, the more complex designs clearly outperformed the stratified design.

The average standard error of the stratified design (.31) was very similar to that of the no exposure control condition (.30). The more complex conditions yielded average standard errors ranging from .34 to .38. The stratified design did reduce the average overlap rate and overlap rate for simulees of similar ability in comparison to the no exposure control maximum information condition, but it was greatly outperformed by the more complex conditional procedures investigated. As in the larger item pool this smaller item pool yielded correlations that were very similar across conditions. Again as seen with the larger item pool the RMSE statistics increased from less complex conditions to more complex conditions.

Burt, Kim, Davis, and Dodd (2003) investigated the functioning of the maximum information, randomesque, modified within .10 logits, and Sympson-Hetter procedures with the generalized partial credit model. In their research both the randomesque and modified within .10 logits procedure were examined with an item group size of three and six. The Sympson-Hetter procedure utilized a target exposure rate of .29. They used an item pool of 210 items from the NAEP 1996 Science Assessment. They found that the highest mean standard error was found for the randomesque procedure with an item group size of six (0.282) and modified within .10 logits procedure with an item group size of six (0.281) procedures followed by the Sympson-Hetter procedure (.274). When they examined the recovery of known theta values and bias the values of all procedures were roughly equivalent.

When they examined their results of exposure rates several findings were of note. When looking at the difference between the three item conditions and the six item conditions maximum exposure rate dropped and better pool utilization was seen. The Sympson-Hetter procedure produced the lowest maximum exposure rate of .33 followed by the modified within .10 logits procedure with an item group size of six (.50) and the randomesque procedure with an item group size of six (.52) conditions. These findings are not consistent with the results reported by Davis (2004) where the randomization procedure outperformed the Sympson-Hetter procedure. The modified within .10 logits and the randomesque procedures with an item group size of six produced the lowest percentage of pool not administered at 23.3% and 24.8%, respectively, followed by the modified within .10 logits and randomesque procedures with an item group size of three procedures with 34.3% and 33.8% of pool not administered. The Sympson-Hetter procedure resulted in 39.5% of the pool not being administered.

When looking at the results for the mean item overlap rates for the three exposure control procedures a similar trend appeared between the three and six item conditions of the randomesque and modified within .10 logits procedures. Overall average overlap was reduced from 30.6% to 22.8% and from 30.5% to 22.8% from the randomesque procedure with an item group size of three to the randomesque procedure with an item group size of six conditions and from the modified within .10 logits procedure with an item group size of three to the modified within .10 logits procedure with an item group size of six. Both the modified within .10 logits procedure with an item group size of six and randomesque procedure with an item group size of six conditions outperformed the Sympson-Hetter procedure which had an overall average overlap of 23.7%.

When they examined the average overlap for simulees with similar abilities the same trend appeared. Average overlap for simulees with similar abilities was reduced from 35.8% to 26.4% and from 35.7% to 26.4% from the randomesque three to randomesque six conditions and from the modified within .10 logits procedure with an item group size of three to the modified within .10 logits procedure with an item group size of six conditions, respectively. Again, both the modified within .10 logits procedure with an item group size of six and randomesque procedure with an item group size of six conditions outperformed the Sympson-Hetter procedure which had an overall average overlap of 27.9%.

Johnson (2006) expanded upon the previous research involving the *a*-stratified and *a*-stratified with *b*-blocking procedures by attempting to determine if an optimum number of strata could be determined for practical use with polytomous item pools. Johnson's simulation research examined the functioning of these procedures with item pools of 175 items and 85 items. In the 85 item pool both stratification procedures were implemented with two and three strata, while in the 157 item pool both stratification procedures were implemented with two, three, four, and five strata. All conditions were conducted with ten replications so the results reported are the average results in each condition.

As indicators of measurement precision, Johnson examined several indices that suggested that there was no appreciable difference in measurement precision across the item pools or conditions examined. For example, all conditions resulted in a distribution of theta estimates and their respective standard deviations indicating a normal distribution of theta estimates. Mean standard error statistics and correlations between known and estimated theta values indicated that all conditions achieved roughly the same level of precision of measurement and recovery of known theta values. Bias and root mean squared error statistics were also roughly equivalent across conditions also indicating no discernable difference in measurement precision.

In terms of the maximum exposure rate produced the stratification procedures did reduce this statistic in comparison to the no exposure control maximum information item selection condition. However, none of the stratification procedures yielded maximum exposure rates as low as the randomesque procedure with an item group size of six. Johnson noted that it was surprising that the *a*-stratified with *b*-blocking condition produced maximum exposure rates similar to that of the *a*-stratified design. This result was surprising because the *a*-stratified with *b*-blocking procedure was designed to compensate for any correlation that may exist between discrimination and difficulty parameters and thus was proposed as a method to control exposure rates more strictly than the *a*-stratified procedure. In terms of the standard deviation of exposure rates, a measure of even item usage, the stratification procedures and the maximum information item selection procedure functioned approximately equivalently in the larger 157 item pool, but all performed worse than the randomesque procedure with an item group size of six which yielded the most even item use of all conditions investigated. In the smaller 85 item pool the stratification procedures did outperform the maximum information condition; however the *a*-stratified with *b*-blocking did not outperform the *a*-stratified condition. Additionally in this smaller item pool the randomesque procedure with an item group size of six again yielded the most even item usage.

In terms of item pool usage the randomesque condition investigated clearly outperformed the maximum information and stratification procedures investigated. In fact, in both item pool sizes examined, the stratification procedures produced roughly equivalent levels of lack of pool use while the randomesque procedure yielded significantly lower levels of lack of pool use. In the larger item pool size all stratification procedures resulted in almost 50% of the item pool never being administered while the randomization procedure resulted in only 33% of the pool never being administered. In the smaller item pool, the stratification procedure resulted in approximately 20% of the item pool never being administered while the randomization procedure investigated resulted in a significantly lower percentage of the pool never being administered, 3%.

Turning next to the performance of the stratification procedures in terms of item overlap a similar trend was seen. Johnson (2006) examined the index of item overlap at two levels. In this research study, simulees were considered to be similar if their known theta values were equal to or less than 2 logits or equal to or less than 1 logit and they were considered different if their known theta values were different by more than 2 logits or more than 1 logit.

When a criterion of 2 logits was used the results are as follows: In the larger item pool, the randomesque condition yielded the lowest level of overall item overlap and item overlap amongst simulees with similar abilities. When simulees of different abilities were examined the level of item overlap was functionally equivalent; the conditions differed by at most one item. In the smaller item pool, the randomesque procedure again yielded the lowest overall item overlap and lowest overlap for simulees with similar abilities. The stratification procedures produced overlap rates similar to that of the maximum information condition. In terms of item overlap between similar examines using a criterion of 1 logit the conditions performed the same as when the larger two logit criterion was used.

Statement of Problem

Now that a foundational understanding of the theory and implementation of CAT systems has been addressed an argument can be made for the present research study. If we look at the example given previously of using an average maximum exposure rate of 0.951 as well as the results reported by Parshall et al. (1998) and McLeod (1998) we can gain a sense of the importance of employing some method of directly controlling item exposure rates. This facet of test security and validity must be closely monitored to endeavor to maximize the advantages offered by CATs while minimizing potential pitfalls.

In regards to the present research Pastor et al.'s (2002) study suggests that the performance of the less complex stratified design was, in a couple cases, functionally equivalent to more complex designs, and was always better than the maximum information condition. In the 100 item pool the *a*-stratified design reduced the percentage of pool not utilized to the same level as the Sympson-Hetter. In the 60 item pool the stratified design again outperformed the maximum information procedure. Also, the stratified design reduced the percentage of overlap in both the case of overall average overlap and the overlap for examinees with similar abilities. In both item pools the correlation of estimated theta to known theta values and standard errors for the stratified design, Sympson-Hetter, and enhanced stratified design were very similar.

Davis (2004) also produced several key findings that support the argument for a closer examination of the randomization procedures. This study not only examined the randomesque, modified within .10 logits, and Sympson-Hetter procedures but also included the more restrictive enhanced stratified design and conditional Sympson-Hetter. This study also included an item group size of three and six for both the randomesque and modified within .10 logits procedures.

In both the three and six item variants of the randomesque and modified within .10 logits procedures the average standard error was lower than the more complex conditional Sympson-Hetter and enhanced *a*-stratified design and only slightly outperformed by the Sympson-Hetter condition. The correlations and bias among all conditions examined were roughly equivalent. Of note is that with the exception of the enhanced *a*-stratified condition, the less complex procedures yielded roughly equivalent RMSE statistics as the more complex Sympson-Hetter and conditional Sympson-Hetter with the enhanced a-stratified condition yielding a slightly higher value.

This study revealed that the six item conditions of the randomesque and modified within .10 logits procedures performed well in regards to both exposure control, percent of pool not administered, and overlap rates. The six item variants of the two randomization procedures were only slightly outperformed by the conditional Sympson-Hetter which yielded the most even item usage. While outperforming the *a*-stratified design these two randomization conditions did not yield maximum exposure rates as low as the more complex conditional procedures; the drop from the three to six item variants suggest that it may be possible that a larger item group size would further drop the maximum exposure rate. Additionally the six item variants of the randomesque and modified within .10 logits procedures, along with the conditional Sympson-Hetter condition, yielded the best pool utilization of all procedures examined. These same three conditions yielded the lowest average overlap rates and overlap rates for simulees with similar abilities.

In this study, a pattern of improvement was also clearly seen when the item group size was increased from three to six which again raises the issue of how a larger item group size would perform. Davis (2004) also noted that the programming requirements and time needed to conduct the simulations necessary to set the exposure control parameters for the Sympson-Hetter and enhanced stratified design were quite intensive. This in conjunction with the relative loss of measurement precision seen with the more complex procedures argued strongly for the use of the randomization procedures examined.

The research conducted by Burt, Kim, Davis, and Dodd (2003) investigated several of the exposure control procedures of interest in the present dissertation. Of note is the finding that all correlations of known to estimated theta values were roughly equivalent. The six item variants of the randomesque and modified within .10 logits conditions produced average standard errors only slightly above the more complex Sympson-Hetter procedure and their three item variants both outperformed the Sympson-Hetter condition.

Burt et al. (2003) also found that in their study the Sympson-Hetter was outperformed by both the three and six item variants of the randomesque conditions as well as by the three item variant of the modified within .10 logits condition in terms of RMSE while all conditions produced negligibly different bias and correlations between known and estimated theta values. Another interesting finding was that when the item group sizes were increased from three to six items the maximum exposure rate dropped and pool utilization increased for both the randomesque and modified within .10 logits procedures while the RMSE increased only slightly. In terms of percent of pool not administered both the three and six item variants outperformed the more complex Sympson-Hetter.

The authors found that the Sympson-Hetter procedure did reduce the maximum exposure rate more than the less complex randomization procedures but one should weigh this finding with the alternative finding by Davis (2004) where they outperformed the Sympson-Hetter procedure. Additionally, they found that the six item variants yielded a standard deviation of exposure rates lower than that of the Sympson-Hetter. In regards to both average overlap and overlap for simulees with similar abilities the six item variants of both the randomesque and modified within .10 logits conditions again outperformed the Sympson-Hetter procedure.

The last notable finding of this research study was the pattern of improvement that was seen when the randomesque and modified within .10 logits procedures were modified in regards to item group size. With the exception of the average standard error and RMSE when more items were available for selection in the randomization procedures the larger six item condition outperformed its three item variant.

Johnson's (2006) simulation study provides a strong argument for the use of the less complex randomesque procedure over the more complex *a*-stratified and *a*-stratified with *b*-blocking procedures. It is interesting to note that in no case did the stratification procedures outperform the randomesque procedure with an item group size of six. Johnson concluded that no optimal number of strata could be determined. Also of note was the fact that in every case the *a*-stratified procedure performed the same as the *a*-stratified with *b*-blocking procedure. As Johnson noted this is rather surprising since the *a*-stratified with *b*-blocking procedure was designed as a refinement to the *a*-stratified procedure.

This research failed to find support for the use of either stratification method over the more simplistic randomesque procedure. It must be noted however that the *a*-stratified with *b*-blocking procedure was designed to be implemented with dichotomous item pools and the method of extension to the polytomous case may have attenuated the potential benefits of the procedure. Not withstanding this cautionary note the research provided clear support for the less complex randomization strategy over the more complex stratification procedures investigated.

In the Pastor et al. (2002) research while it was shown that the stratified design was the least favorable condition in regards to item overlap and exposure control when compared to more complex procedures it was able to reduce exposure and overlap rates and increase item use while only demonstrating a minor loss in measurement precision. This study could not be used to argue for the use of the stratified design over more complex procedures but it can be seen as the beginning of the development of an argument for the exploration of less complex procedures as an alternative for more complex procedures. Pastor et al. taken in conjunction with Johnson (2006) provides the basis for an argument for the investigation of less complex randomization procedures over more complicated conditional or stratification procedures.

Both the Burt et al. (2003) study and the Davis (2004) study provide good support for the investigation of more simplistic randomization procedures as a viable alternative for more complex conditional procedures. Davis' (2004) research provides the strongest support for such an argument. On many criteria for efficacious performance both studies found that the less complex randomization procedures outperformed or were at least comparable to the more complex conditional procedures investigated. In the case of the modified within .10 logits procedure and randomesque procedures the restricted forms of these procedures investigated in the present study was expected counter the issue of randomization strategies yielding larger maximum exposure rates found by both Davis and Burt et al. Using these four studies as its basis, the present research focused on the less complex randomization procedures using the generalized partial credit model. Specifically this research endeavored to answer the following questions:

1. What effect will the implementation of a maximum exposure rate constraint have on the performance of the modified within .10 logits procedure in regards to measurement precision and test security?
2. What effect will the implementation of a maximum exposure rate constraint have on the performance of the randomesque procedure in regards to measurement precision and test security?
3. What impact will item group size have on the randomesque, modified within .10 logits, restricted randomesque, and restricted modified within

.10 logits procedures in regards to measurement precision and test security?

4. How will measurement precision be affected in the exposure control procedures investigated?
5. Which exposure control procedure will protect test security in regards to exposure control, pool utilization, and item overlap the best?

Chapter III: Methodology

In this dissertation four exposure control strategies were evaluated for efficacy under the generalized partial credit model. As a base-line condition, the maximum information item selection technique was used where no exposure control was to be examined. The randomization procedures used were the modified within .10 logits, restricted modified within .10 logits, randomesque, and progressive restricted procedures. The modified within .10 logits, restricted modified within .10 logits, randomesque, and restricted randomesque methods were each examined with three item group sizes. In each case the number of items available for selection was three, six, and nine. To examine the effect of item pool size on the exposure control procedures an item pool of 100 items and 200 items was used. This design resulted in 13 exposure control conditions and a no exposure control condition across two item pools yielding 28 conditions.

Item Pools

The data used in this study was based on a national admissions test. The data set, as provided, had 157 items where 63% of the items had three response categories, 18.5% had four response categories, and the final 18.5% of items had five response categories. The items bank was balanced across three content areas. These contents are denoted as content 1, content 2 and content 3. Content 1 represented 39% of items, content 2 represented 37.5% of items, and content 3 represented 23.5% of items. Two item pools were constructed from this data. The desired item pool sizes were 100 and 200 items. To attain the proper number of items for the 100 item pool 36 of the items with three response options, eleven of the items with four response options, and twelve of the items with five response options were removed. For the two hundred item pool the 100 item parameters were duplicated to produce item parameters for 200 items. The joint distribution of percentages of the number of items in the content by number of response options in the item banks are presented in Table 1.

Data Generation

Eleven separate datasets were generated for use in the current study via the IRTGEN SAS data generation macro (Whittaker et al., 2003). IRTGEN begins the process of generating item responses by first assigning a random number for the normal distribution to serve as a known theta value. Using this known theta value and the item parameters from the national admissions test the macro calculated the probability of responding in each of the response categories. The cumulative subtotals for the response categories are calculated by summing these values. Following this a random number is generated from the uniform distribution and compared to the cumulative subtotal for each of the response categories. The simulee is then assigned the response category that is at or below the random number. This process is then repeated on every item for every simulee.

A calibration dataset with a sample size of 10,000 was generated through the IRTGEN SAS data generation macro utilizing the original parameters from the national admissions test. For the 200 item pool 10 datasets were generated for use as replications and one dataset was generated for use in item calibration. Each of the ten datasets used as replications were generated with 1,000 simulees. The item parameters calibrated through PARSCALE, see parameter estimation section for details, for the 200 item pool were used to generate responses for all data sets by utilizing the IRTGEN SAS data generation

Table 1: Distribution of Items in the 100 and 200 Item Bank by Content Area and Number of Response Options

Content Area	Number of Response Options			Cumulative %
	3	4	5	
1	26.75%	6.37%	5.73%	38.85%
2	26.75%	3.82%	7.00%	37.57%
3	9.55%	8.28%	5.73%	23.56%
Cumulative %	63.05%	18.47%	18.46%	99.98%[a]

[a] Percentages do not sum to 100 due to rounding error.

macro developed by Whittaker, Fitzpatrick, Williams, and Dodd (2003). This produced ten datasets with responses for 1,000 simulees on the 200 items. This dataset generated responses according to the generalized partial credit model. To attain the dataset for the 100 item pool the first 63 items with three response options, 18 items with 4 response options, and 19 items with 5 response options were used.

Parameter Estimation

The 10,000 responses from the calibration sample to the 200 item pool were entered into PARSCALE (Muraki & Bock, 1993) for calibration according to the generalized partial credit model. This program uses marginal maximum likelihood EM algorithm. This algorithm uses a two step process to estimate parameters. In the first step the provisional expected frequency and sample size are calculated. Following this step the marginal maximum likelihood is estimated. This iterative process is continued until stable estimates are achieved.

CAT Simulations

Using a modified version of a program (Davis, 2002) developed originally by Chen (1996) the simulations were run on each CAT condition. The initial theta for all simulated test takers was 0.00 and a variable step size was used initially until responses were made in two different categories at which point maximum likelihood estimation was used. All conditions were simulated using a fixed length test of 20 items. The Kingsbury and Zara (1989) procedure was used to balance both the content and the number of response options, categories, in the item pool. In this way items were categorized by both content and number of categories. This produced nine distinct item category by content target groups. The percentages for each of these nine target groups were reported in Table 1. In this method, after an item was administered, the proportion of items given in each item category by content group was computed and compared to the target desired proportion which was attained from the observed proportions of the item pool. From these the item category by content group with the largest discrepancy was used as a constraint for selection of the next item.

The Maximum Information Procedure

In this condition items were selected in order to maximize the information given the trait estimate at the time.

The Modified within .10 logits Procedure

Three variants of this procedure were utilized. This procedure selects items that provide the most psychometric information at the estimated theta, estimated theta minus .10 logits, and estimated theta plus .10 logits levels. The item to be administered is then randomly selected from this subset of items. The number of items selected at the three levels was varied in the current study for evaluative purposes. In the first variant the most informative item at each level was selected. This produced a group of three items from which to choose. In the second variant the two most informative items at each level were selected. This produced a group of six items from which to choose. In the third variant the three most informative items at each level were selected. This produced a group of nine items from which to choose.

The Restricted Modified within .10 logits Procedure

Prior to items being considered for selection and administration, an exposure rate maximum constraint was utilized. All items were initially assigned an exposure rate of 0.0. As items were administered in the simulated CATs the exposure rate for every item was recalculated. This procedure selected items for administration in the same manner as the modified within .10 logits procedure with one modification. Before creation of the item group, items whose exposure rate exceeded 0.30 were blocked from inclusion in the item group. Only items whose exposure rate was less than or equal to 0.30 were considered for selection in the item group. Once this subpool had been defined, item selection proceeded according to the modified within .10 logits procedure.

The Randomesque Procedure

The randomesque procedure (Kingsbury and Zara, 1989) for controlling item exposure randomly selects an item for administration from a group of the most informative items. For the purposes of this study three group sizes were used for comparison purposes. The group sizes were three, six, and nine. For illustration proposes, in the condition with a group size of nine, throughout the examination when items were selected there was a group of the nine most informative items from which the item to be administered was randomly chosen. This procedure differs from the modified within .10 logits procedure in that while that procedure creates an item group from three points on the theta scale this procedure selected all items from the theta estimate.

The Restricted Randomesque Procedure

The restricted randomesque procedure for controlling item exposure randomly selects an item for administration from a group of the most informative items given that those items exposure rate do not exceed a maximum exposure rate of 0.30. For the purposes of this study three group sizes were used for comparison purposes. The group sizes used were three, six, and nine. Therefore after every administration of an item its exposure rate is calculated and only items that have not been administered in the current test and whose exposure rate did not exceed 0.30 were considered for administration via the randomesque procedure.

The Progressive Restricted Procedure

In this procedure items are chosen for administration by a weight value which is calculated in such a way as to reduce the impact of item information in item selection in the early stages of the test. This procedure also explicitly restricts the maximum exposure rate of any item. In the present research the maximum exposure rate constraint for any item was 0.30. All items were initially assigned an exposure rate of 0.00. As items were administered in the simulated CATs the exposure rate was recalculated. If an items exposure rate ever exceeded 0.30 it was not considered for administration. Therefore in this procedure the item whose exposure rate did not exceed 0.30 had a weight computed, and the item with the largest weight value was administered.

Data Analysis

To evaluate the recovery of known theta in the 24 conditions, descriptive statistics as well as the correlation between the known and estimated theta values were examined. The grand mean, standard error of the mean, minimum mean, and maximum mean for the estimated theta values in each condition, across all ten replications, are reported. For all descriptive statistics where a mean, minimum, and maximum are provided the mean value represents a more stable indicator of the statistic while the minimum and maximum are meant to provide an index of the variability among the replications. The standard deviation and standard error of the estimated theta values in each condition include the mean, minimum, and maximum, across all ten replications. Similarly the report of the correlation of the estimated theta values to known theta values for each condition include the mean, minimum, and maximum correlations across all ten replications. Additionally bias, root mean squared error (RMSE), standardized root mean square difference (SRMSD), and average absolute difference statistics were calculated. The formulas are as follows:

$$Bias = \frac{\sum_{k=1}^{n}(\hat{\theta}_k - \theta_k)}{n},$$

$$RMSE = \left[\frac{\sum_{k=1}^{n}(\hat{\theta}_k - \theta_k)^2}{n}\right]^{.5}, \text{ and}$$

$$SRMSD = \left[\frac{\frac{1}{n}\sum_{k=1}^{n}(\hat{\theta}_k - \theta_k)^2}{\frac{s_{\hat{\theta}}^2 + s_{\theta}^2}{2}} \right]^{.5},$$

where $\hat{\theta}_k$ is the trait estimate for simulee k, θ_k is the known trait level for simulee *k*, $s_{\hat{\theta}}^2$ is the variance of estimated trait levels, and s_{θ}^2 is the variance for known trait levels. As before these statistics include the mean, minimum, and maximum values across all ten replications.

Item exposure rates were attained by dividing the number of times an item was administered by the number of simulated test takers. Reported results for the descriptive statistics for the exposure rates and standard deviation of exposure rates for each condition include the mean, minimum, and maximum value across all ten replications. Across conditions, the average frequency distributions of exposure rates, average exposure rate, and average maximum exposure rate were attained. Pool utilization was measured by the average percentage of the items in the pool that were never used across replications for each condition.

The audit trails of the examinees were compared for the purpose of measuring test overlap between simulated test takers with similar and different θ estimates. Simulees were referred to as having similar ability levels if their known theta levels differed by one logit or less and were referred to as different if their known theta levels differed by more than one logit (Boyd, 2003). Reported results of item overlap for each condition include the mean, minimum, and maximum values across all ten replications.

For each of the indices of measurement precision and test security listed above each condition was compared to the baseline no exposure control maximum information condition as well as to every other exposure control condition. The comparison to the no exposure control condition serves the purpose of providing a measure of improvement provided by each condition. Comparing the exposure control conditions to each other provides a measure of the relative performance of each condition. It needs to be noted that there is no research which makes use of the restricted randomesque or the restricted within .10 logits procedures investigated here. As such the expectations of these conditions were based, primarily, on the performance in previous research of the unrestricted procedures from which they are derived.

The grand mean and mean of standard deviation values across replications of the same condition gave a summary of the distribution of the estimated theta values when each condition has been implemented. It was expected that the distribution of estimated abilities would match those of the known abilities. The recovery of known theta values was evaluated by the correlation of estimated theta values to known theta values. It was expected that each condition would yield a strong positive correlation indicating that the estimated theta values closely match those of the known theta values. The average values across replications of the bias, RMSE, SRMSE, and AAD serve as indicators of measurement precision. It was expected that all conditions would perform equivalently.

The average maximum exposure rate served as the primary indicator of control of exposure rates. It was anticipated that the progressive restricted, restricted modified within .10 logits, and restricted randomesque procedure would produce maximum exposure rates slightly above their prespecified maximum exposure rates. It was further anticipated that as item group size increased from three to six to nine in the restricted modified within .10 logits, restricted randomesque, modified within .10 logits, and randomesque procedures maximum exposure rates would decrease. Even item usage, as indicated by the standard deviation of exposure rates, was anticipated to be well produced by all conditions.

Pool utilization rates were described by the percentage of the pool not administered. The expansion of item group size in the restricted modified within .10 logits, restricted randomesque, modified within .10 logits, and randomesque procedures was expected to produce a decrease in the percent of the item pool not administered. All procedures investigated were expected to outperform the no exposure control condition in this regard. The last indicator of test security examined was the item overlap rate. Here again, all exposure control procedures examined ere expected to outperform the no exposure control condition. Each exposure control condition was compared to determine which yields the lowest levels of overall item overlap, overlap between examinees with similar abilities, and examinees with different abilities.

Chapter IV: Results

The following reporting of results for each of the 28 conditions investigated in the current study is reported separately by item pool size. For each condition, results shall be reported based upon measurement precision indices, exposure rates, and overlap rates. All results reported were based upon ten replications of each condition. For each of the results reported the 100 item pool shall be discussed first, followed by the results for the 200 item pool.

Measurement Precision Indices

The items in both the 100 and 200 item pool consisted of item which had three, four, or five response options. The mean, standard deviation, minimum and maximum for the item parameters in the 100 item pool are listed in Table 2. The mean, standard deviation, minimum and maximum for the item parameters in the 200 item pool are listed in Table 3. In both tables the descriptive statistics are given for the total item pool as well as for each individual content by number of categories item type. As a result of using items with varying number of response options, the number of step difficulties also varied.

The information function for the 100 item pool is provided in Figure 1. Figure 2 provides the information function for each of the nine content area by number of category item types. The information function for the 200 item pool is provided in Figure 3. Figure 4 provides the information function for each of the nine content area by number of category item types. The test information functions for both item pools peaked at a theta

Table 2: Descriptive Statistics for the Parameters for the 100 Item Pool and by Content Area and Number of Categories

Content Area and Number of Categories	Total Pool	Content 1			Content 2			Content 3		
		3	4	5	3	4	5	3	4	5
Item Discrimination										
Mean	0.933	1.013	0.783	0.729	1.010	0.801	0.815	1.074	0.781	0.790
Standard Deviation	0.202	0.187	0.144	0.106	0.149	0.145	0.121	0.276	0.128	0.118
Minimum	0.599	0.759	0.635	0.613	0.774	0.662	0.669	0.758	0.599	0.674
Maximum	1.568	1.452	0.995	0.886	1.376	1.005	0.991	1.568	0.940	1.001
n	100	27	6	6	26	4	7	10	8	6
Step Difficulty 1										
Mean	-1.078	-1.162	0.242	-1.182	-1.092	-0.488	-1.821	-1.454	-0.265	-1.845
Standard Deviation	0.860	0.633	0.652	0.837	0.778	0.388	0.662	0.798	0.758	0.784
Minimum	-3.030	-2.202	-0.655	-2.342	-2.161	-0.765	-2.627	-2.681	-0.931	-3.030
Maximum	1.532	0.200	1.152	-0.101	0.707	0.066	-0.936	-0.281	1.532	-0.830
n	100	27	6	6	26	4	7	10	8	6
Step Difficulty 2										
Mean	0.047	0.617	-0.660	-0.827	0.516	-0.740	-0.857	0.317	-0.683	-0.869
Standard Deviation	0.987	0.906	0.504	0.419	0.980	0.580	0.417	0.751	0.254	0.458
Minimum	-1.692	-1.158	-1.203	-1.433	-1.348	-1.268	-1.692	-0.511	-1.019	-1.438
Maximum	3.105	2.338	0.031	-0.303	3.105	0.061	-0.425	1.829	-0.297	-0.328
n	100	27	6	6	26	4	7	10	8	6
Step Difficulty 3										
Mean	-0.576	NA	-0.568	-0.411	NA	-0.598	-0.741	NA	-0.445	-0.718
Standard Deviation	0.432	NA	0.608	0.438	NA	0.414	0.194	NA	0.522	0.355
Minimum	-1.486	NA	-1.306	-0.819	NA	-0.875	-1.162	NA	-1.486	-1.222
Maximum	0.342	NA	0.060	0.342	NA	0.015	-0.576	NA	0.339	-0.269
n	37	NA	6	6	NA	4	7	NA	8	6
Step Difficulty 4										
Mean	-0.494	NA	NA	-0.470	NA	NA	-0.519	NA	NA	-0.489
Standard Deviation	0.578	NA	NA	0.927	NA	NA	0.408	NA	NA	0.376
Minimum	-2.231	NA	NA	-2.231	NA	NA	-1.211	NA	NA	-1.026
Maximum	0.318	NA	NA	0.318	NA	NA	0.160	NA	NA	-0.063
n	19	NA	NA	6	NA	NA	7	NA	NA	6

Note: NA = Not Applicable

Table 3: Descriptive Statistics for the Parameters for the 200 Item Pool and by Content Area and Number of Categories

Content Area and Number of Categories	Total Pool	Content 1			Content 2			Content 3		
		3	4	5	3	4	5	3	4	5
Item Discrimination										
Mean	0.932	1.012	0.779	0.729	1.012	0.810	0.811	1.066	0.780	0.785
Standard Deviation	0.200	0.185	0.133	0.105	0.148	0.141	0.115	0.258	0.126	0.115
Minimum	0.585	0.755	0.627	0.591	0.764	0.662	0.648	0.729	0.585	0.632
Maximum	1.568	1.522	0.995	0.911	1.387	1.040	0.991	1.568	0.968	1.001
n	200	54	12	12	52	8	14	20	16	12
Step Difficulty 1										
Mean	-1.084	-1.162	0.245	-1.204	-1.085	-0.533	-1.838	-1.456	-0.275	-1.892
Standard Deviation	0.862	0.632	0.621	0.794	0.771	0.373	0.607	0.777	0.762	0.751
Minimum	-3.246	-2.306	-0.719	-2.439	-2.210	-0.829	-2.627	-2.681	-1.100	-3.246
Maximum	1.641	0.200	1.152	-0.101	0.763	0.066	-0.936	-0.281	1.641	-0.830
n	200	54	12	12	52	8	14	20	16	12
Step Difficulty 2										
Mean	0.046	0.616	-0.666	-0.813	0.506	-0.714	-0.848	0.320	-0.679	-0.877
Standard Deviation	0.977	0.888	0.464	0.398	0.968	0.506	0.415	0.726	0.247	0.420
Minimum	-1.754	-1.158	-1.203	-1.433	-1.380	-1.268	-1.754	-0.515	-1.152	-1.458
Maximum	3.105	2.338	0.031	-0.211	3.105	0.061	-0.425	1.829	-0.297	-0.328
n	200	54	12	12	52	8	14	20	16	12
Step Difficulty 3										
Mean	-0.568	NA	-0.566	-0.401	NA	-0.569	-0.742	NA	-0.435	-0.708
Standard Deviation	0.432	NA	0.587	0.440	NA	0.397	0.182	NA	0.492	0.337
Minimum	-1.486	NA	-1.329	-0.819	NA	-0.875	-1.162	NA	-1.486	-1.278
Maximum	0.477	NA	0.070	0.477	NA	0.118	-0.576	NA	0.339	-0.269
n	74	NA	12	12	NA	8	14	NA	16	12
Step Difficulty 4										
Mean	-0.4992	NA	NA	-0.4778	NA	NA	-0.5221	NA	NA	-0.4936
Standard Deviation	0.5878	NA	NA	0.9191	NA	NA	0.4015	NA	NA	0.3543
Minimum	-2.3955	NA	NA	-2.3955	NA	NA	-1.2108	NA	NA	-1.1024
Maximum	0.3425	NA	NA	0.3425	NA	NA	0.2388	NA	NA	-0.0627
n	38	NA	NA	12	NA	NA	14	NA	NA	12

Note: NA = Not Applicable

Figure 1: Test Information Function for the 100 Item Pool

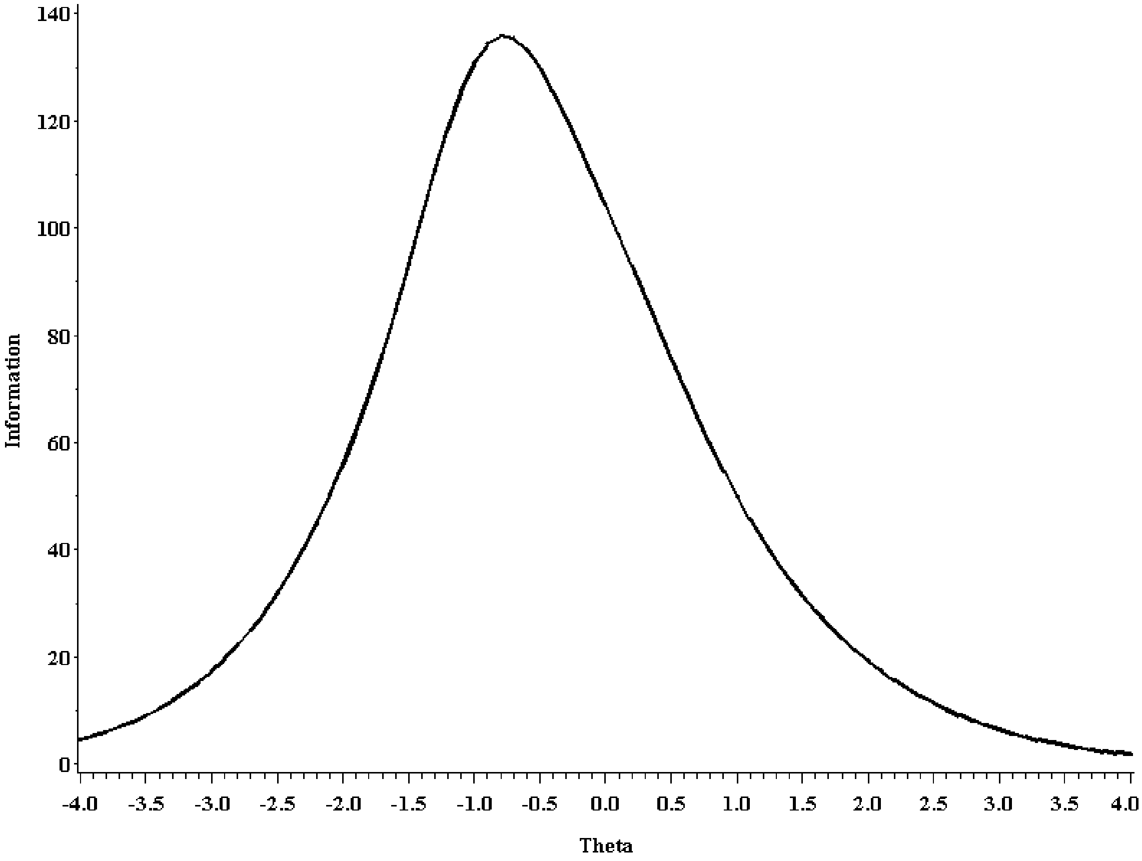


Figure 2: Information Functions for the Nine Content by Number of Category Item Types for the 100 Item Pool

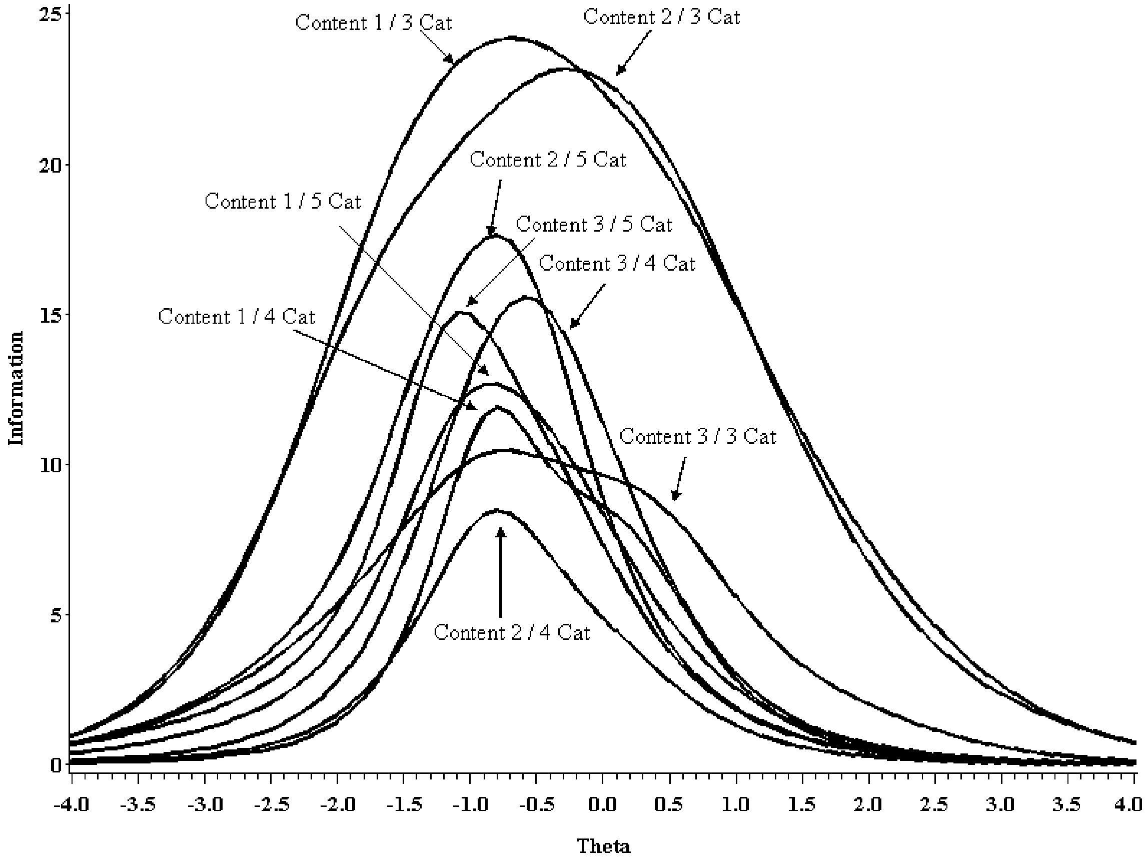


Figure 3: Test Information Function for the 200 Item Pool

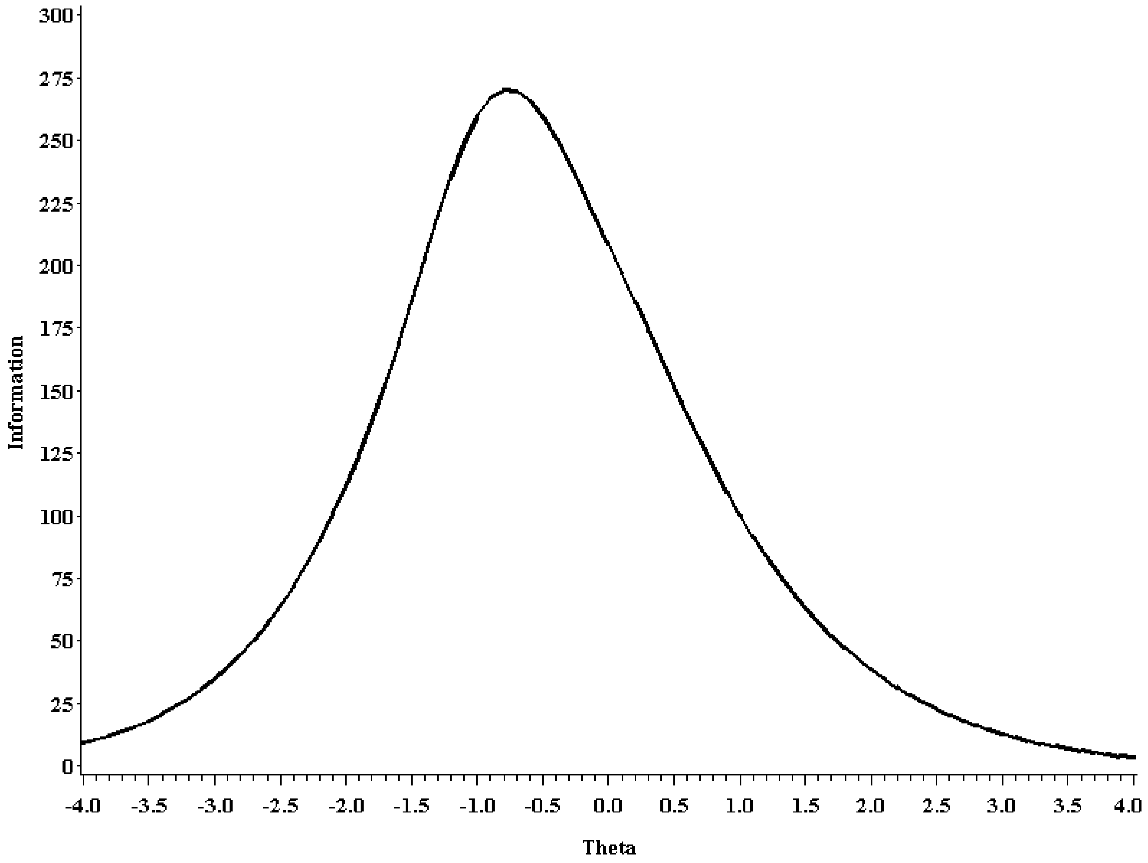
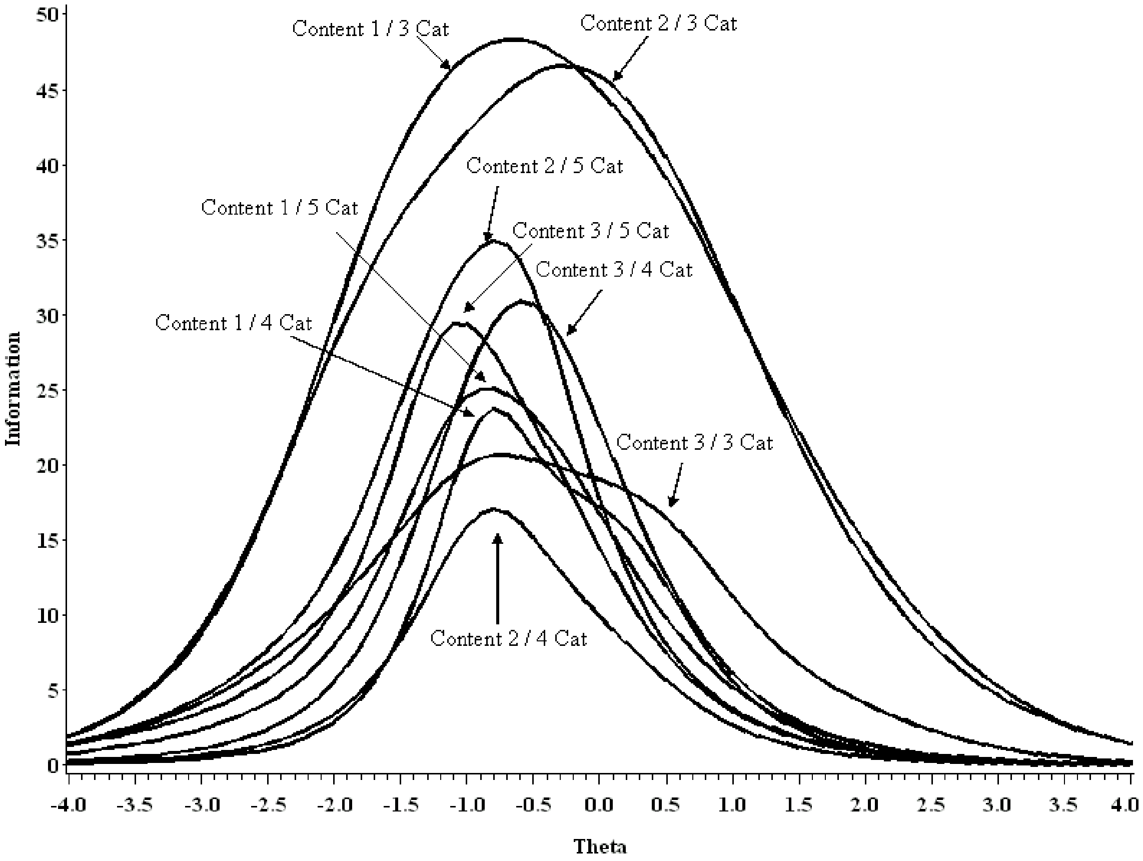


Figure 4: Information Functions for the Nine Content by Number of Category Item Types for the 200 Item Pool



value of -0.8. The information functions for each of the nine content by number of category item types peaked, in both item pools, between theta values of -0.3 and -1.1.

Table 4 and Table 5 list the grand mean, standard error of the mean, minimum mean, and maximum mean for the estimated theta values across all conditions for the 100 and 200 item pools respectively. Table 6 provides the number of nonconvergent cases produced by each condition as well as the average, minimum, and maximum number of nonconvergent cases in both the 100 and 200 item pools. Nonconvergent cases were defined as simulees whose theta estimate was greater than or equal to 4.0 or less then or equal to -4.0, or if a maximum likelihood estimate was never attained in the CAT. The nonconvergent cases in each condition were excluded from subsequent analysis for that condition.

In the 100 item pool, the number of nonconvergent cases was relatively high (33 to 69 cases) in comparison to the no exposure control condition (20 cases). In the 200 item pool the three item variants of the randomesque, restricted randomesque, modified within .10 logits, and restricted modified within .10 logits procedures yielded approximately equal number of nonconvergent cases (26, 24, 25, and 23 cases respectively) as the no exposure control condition (24 cases). The remaining procedures yielded a somewhat higher number of nonconvergent cases (34 to 61 cases).

From Tables 4 and 5 it can be observed that the grand mean of estimated theta in each condition, for both item pools, is very close to zero. In Tables 7 and 8 the descriptive statistics for the standard deviations of the estimated thetas in the 100 and 200 item pool are provided. For each condition the grand mean of the standard deviation of estimated theta values was approximately 1. Based on the information provided in Tables

Table 4: Descriptive Statistics for the Estimated Theta Values of the 100 Item Pool Averaged Across Ten Replications

Exposure Control Procedure	Grand Mean	Standard Error of the Mean	Minimum Mean	Maximum Mean
No Exposure Control	0.010	0.028	-0.020	0.064
Progressive Restricted	0.011	0.028	-0.038	0.053
Randomesque (3)	0.007	0.030	-0.036	0.052
Randomesque (6)	0.009	0.028	-0.034	0.060
Randomesque (9)	0.006	0.022	-0.021	0.042
Restricted Randomesque (3)	0.011	0.030	-0.031	0.061
Restricted Randomesque (6)	0.013	0.027	-0.029	0.059
Restricted Randomesque (9)	0.007	0.022	-0.027	0.044
Modified Within .10 Logits (3)	0.009	0.028	-0.029	0.050
Modified Within .10 Logits (6)	0.011	0.026	-0.039	0.061
Modified Within .10 Logits (9)	0.008	0.030	-0.034	0.057
Rest. Modified Within .10 Logits (3)	0.012	0.030	-0.032	0.062
Rest. Modified Within .10 Logits (6)	0.014	0.034	-0.032	0.075
Rest. Modified Within .10 Logits (9)	0.009	0.030	-0.046	0.042

Table 5: Descriptive Statistics for the Estimated Theta Values of the 200 Item Pool Averaged Across Ten Replications

Exposure Control Procedure	Grand Mean	Standard Error of the Mean	Minimum Mean	Maximum Mean
No Exposure Control	0.017	0.023	-0.012	0.056
Progressive Restricted	0.010	0.035	-0.042	0.063
Randomesque (3)	0.014	0.023	-0.017	0.053
Randomesque (6)	0.009	0.030	-0.032	0.056
Randomesque (9)	0.000	0.029	-0.036	0.055
Restricted Randomesque (3)	0.014	0.026	-0.026	0.048
Restricted Randomesque (6)	0.010	0.034	-0.033	0.071
Restricted Randomesque (9)	0.004	0.028	-0.046	0.053
Modified Within .10 Logits (3)	0.018	0.025	-0.025	0.057
Modified Within .10 Logits (6)	0.012	0.028	-0.028	0.053
Modified Within .10 Logits (9)	0.006	0.030	-0.038	0.057
Rest. Modified Within .10 Logits (3)	0.016	0.027	-0.040	0.053
Rest. Modified Within .10 Logits (6)	0.015	0.031	-0.026	0.063
Rest. Modified Within .10 Logits (9)	0.008	0.031	-0.042	0.052

Table 6: Number of Nonconvergent Cases for the 100 and 200 Item pool in Each Condition

	Replication	Nonconvergent Cases													
		MI	PR	RD-3	RD-6	RD-9	RRD-3	RRD-6	RRD-9	MW-3	MW-6	MW-9	RMW-3	RMW-6	RMW-9
100 Item Pool	1	26	74	35	49	54	35	52	56	35	49	64	26	50	60
	2	20	66	38	37	56	35	46	55	37	39	48	34	50	45
	3	20	73	47	55	54	32	50	56	48	52	61	32	43	49
	4	21	65	38	29	58	38	30	55	41	32	57	33	37	45
	5	15	69	42	34	53	37	40	41	42	46	40	32	44	42
	6	21	59	33	49	51	31	37	45	34	50	59	32	35	59
	7	21	69	30	38	54	27	40	58	24	37	56	24	32	51
	8	19	75	27	32	56	41	44	61	24	36	51	30	38	52
	9	16	64	51	44	41	45	44	49	54	55	54	44	53	54
	10	23	80	45	46	50	44	41	54	38	48	52	43	47	59
	Average (Min, Max)	20 (15, 26)	69 (59, 80)	39 (27, 51)	41 (29, 55)	53 (41, 58)	37 (27, 45)	42 (30, 52)	53 (41, 61)	38 (24, 54)	44 (32, 55)	54 (40, 64)	33 (24, 44)	43 (32, 53)	52 (42, 60)
200 Item Pool	1	34	63	27	43	58	32	39	57	25	42	47	26	37	50
	2	19	64	22	45	40	19	38	45	25	36	30	18	33	34
	3	31	58	26	41	55	27	43	58	21	33	39	22	34	40
	4	27	60	24	31	53	22	38	57	19	28	32	21	34	37
	5	20	72	30	43	44	27	39	43	33	35	38	33	35	41
	6	21	52	27	46	52	19	38	53	25	47	41	17	40	46
	7	23	54	22	39	45	24	34	44	23	30	42	18	32	42
	8	24	59	20	50	42	21	36	45	15	39	31	18	29	32
	9	17	58	32	33	44	23	30	45	36	29	35	28	31	33
	10	20	68	31	34	57	25	32	59	31	35	41	25	34	45
	Average (Min, Max)	24 (17, 34)	61 (52, 72)	26 (20, 32)	41 (31, 50)	49 (40, 58)	24 (19, 32)	37 (30, 43)	51 (43, 59)	25 (15, 36)	35 (28, 47)	38 (30, 47)	23 (17, 33)	34 (29, 40)	40 (32, 50)

Table 7: Descriptive Statistics for the Standard Deviation of the Estimated Theta Values for the 100 Item Pool Averaged Across Ten Replications

Exposure Control Procedure	Standard Deviation		
	Grand Mean	Minimum Mean	Maximum Mean
No Exposure Control	1.068	1.016	1.109
Progressive Restricted	1.099	1.054	1.138
Randomesque (3)	1.074	1.025	1.111
Randomesque (6)	1.088	1.054	1.125
Randomesque (9)	1.088	1.054	1.143
Restricted Randomesque (3)	1.078	1.027	1.113
Restricted Randomesque (6)	1.087	1.046	1.128
Restricted Randomesque (9)	1.086	1.053	1.133
Modified Within .10 Logits (3)	1.075	1.031	1.114
Modified Within .10 Logits (6)	1.093	1.052	1.137
Modified Within .10 Logits (9)	1.093	1.063	1.127
Rest. Modified Within .10 Logits (3)	1.074	1.034	1.116
Rest. Modified Within .10 Logits (6)	1.088	1.032	1.145
Rest. Modified Within .10 Logits (9)	1.095	1.050	1.146

Table 8: Descriptive Statistics for the Standard Deviation of the Estimated Theta Values for the 200 Item Pool Averaged Across Ten Replications

Exposure Control Procedure	Standard Deviation		
	Grand Mean	Minimum Mean	Maximum Mean
No Exposure Control	1.068	1.030	1.098
Progressive Restricted	1.082	1.054	1.125
Randomesque (3)	1.066	1.034	1.095
Randomesque (6)	1.076	1.028	1.112
Randomesque (9)	1.075	1.039	1.103
Restricted Randomesque (3)	1.068	1.026	1.110
Restricted Randomesque (6)	1.073	1.027	1.106
Restricted Randomesque (9)	1.077	1.035	1.107
Modified Within .10 Logits (3)	1.066	1.031	1.097
Modified Within .10 Logits (6)	1.071	1.029	1.103
Modified Within .10 Logits (9)	1.074	1.029	1.112
Rest. Modified Within .10 Logits (3)	1.067	1.022	1.110
Rest. Modified Within .10 Logits (6)	1.067	1.032	1.102
Rest. Modified Within .10 Logits (9)	1.075	1.032	1.107

4, 5, 7, and 8 we can conclude that each condition produced a distribution of estimated thetas which was approximately normal with a mean of zero and standard deviation of one.

The first statistical index of measurement precision was provided by the standard error of the theta estimate provided in Tables 9 and 10 for the 100 and 200 item pools respectively. In the 100 item pool the no exposure control condition yielded the lowest grand mean of the standard error of 0.288. The three item variants of the randomesque (0.303) and modified within .10 logits (0.303) procedures yielded the next lowest grand mean standard error followed closely by the three item variants of the restricted randomesque (0.317) and restricted modified within .10 logits (0.317) procedures. The progressive restricted (0.320) procedure, along with the six item variants of the randomesque (0.323) and modified within .10 logits (0.324) procedures, yielded the next highest grand mean standard error. The six item variants of the restricted randomesque (0.330) and restricted modified within .10 logits (0.330) procedures along with the nine item variants of the randomesque (0.334), restricted randomesque (0.336), modified within .10 logits (0.336), and restricted modified within .10 logits (0.339) procedures yielded the highest levels of grand mean standard error.

In the 200 item pool, the no exposure control procedure again, as expected, yielded the lowest grand mean standard error (0.279). The three item variants of the randomesque (0.286), modified within .10 logits (0.287), restricted randomesque (0.293), restricted modified within .10 logits (0.293) produced the next lowest grand mean standard error. These were followed by the six item variants of the modified within .10

Table 9: Descriptive Statistics for the Standard Error for Theta Estimates for the 100 Item Pool Averaged Across Ten Replications

Exposure Control Procedure	Standard Error		
	Grand Mean	Minimum Mean	Maximum Mean
No Exposure Control	0.288	0.284	0.292
Progressive Restricted	0.320	0.316	0.323
Randomesque (3)	0.303	0.299	0.307
Randomesque (6)	0.323	0.319	0.328
Randomesque (9)	0.334	0.331	0.337
Restricted Randomesque (3)	0.317	0.313	0.321
Restricted Randomesque (6)	0.330	0.325	0.334
Restricted Randomesque (9)	0.336	0.334	0.339
Modified Within .10 Logits (3)	0.303	0.299	0.307
Modified Within .10 Logits (6)	0.324	0.319	0.330
Modified Within .10 Logits (9)	0.336	0.333	0.340
Rest. Modified Within .10 Logits (3)	0.317	0.313	0.321
Rest. Modified Within .10 Logits (6)	0.330	0.324	0.336
Rest. Modified Within .10 Logits (9)	0.339	0.335	0.343

Table 10: Descriptive Statistics for the Standard Error for Theta Estimates for the 200 Item Pool Averaged Across Ten Replications

Exposure Control Procedure	Standard Error		
	Grand Mean	Minimum Mean	Maximum Mean
No Exposure Control	0.279	0.276	0.282
Progressive Restricted	0.302	0.299	0.306
Randomesque (3)	0.286	0.283	0.289
Randomesque (6)	0.299	0.295	0.303
Randomesque (9)	0.309	0.306	0.313
Restricted Randomesque (3)	0.293	0.289	0.295
Restricted Randomesque (6)	0.301	0.297	0.306
Restricted Randomesque (9)	0.310	0.306	0.313
Modified Within .10 Logits (3)	0.287	0.284	0.290
Modified Within .10 Logits (6)	0.299	0.295	0.303
Modified Within .10 Logits (9)	0.310	0.306	0.313
Rest. Modified Within .10 Logits (3)	0.293	0.289	0.296
Rest. Modified Within .10 Logits (6)	0.301	0.298	0.305
Rest. Modified Within .10 Logits (9)	0.311	0.306	0.314

logits (0.299), randomesque (0.299), restricted randomesque (0.301), restricted modified within .10 logits (0.301) procedures and the progressive restricted procedure (0.302). The nine item variants of the randomesque (0.309), modified within .10 logits (0.310), restricted randomesque (0.310), and restricted modified within .10 logits (0.311) procedures produced the highest grand mean standard error.

Recovery of known theta values as defined as the correlation between known and estimated theta values is reported in Table 11 and 12 for the 100 and 200 item pools respectively. For both item pools and in all conditions, the correlation between known and estimated theta values was high, ranging from .95 to .96. From this, it is fair to conclude that the implementation of the exposure control procedures did not significantly alter the estimation of theta in relation to the no exposure control condition and provided good recovery of known theta values.

The measurement precision indices of average absolute difference (AAD), bias, root mean squared error (RMSE), and standardized root mean squared error (SRMSE) are provided in Table 13 and 14 for the 100 and 200 item pool, respectively. For the 100 item pool AAD ranged from 0.233 to 0.273, bias ranged from -0.020 to -0.010, RMSE ranged from 0.302 to 0.354, and SRMSE ranged from 0.531 to 0.565. The no exposure control condition produced the lowest levels of each of the statistics indicating the best measurement precision of the conditions investigated. The nine item variant of the restricted within .10 logits procedure produced the highest values for all the measurement precision indices. Across conditions, however, the values produced in each of the indices were functionally equivalent indicating that for the 100 item pool the utilization of the

Table 11: Descriptive Statistics for the Correlation between Known and Estimated Theta Values for the 100 Item Pool Averaged Across Ten Replications

Exposure Control Procedure	Correlation Coefficient		
	Mean	Minimum	Maximum
No Exposure Control	0.959	0.954	0.962
Progressive Restricted	0.952	0.947	0.955
Randomesque (3)	0.955	0.951	0.958
Randomesque (6)	0.950	0.944	0.954
Randomesque (9)	0.948	0.944	0.953
Restricted Randomesque (3)	0.951	0.945	0.955
Restricted Randomesque (6)	0.948	0.942	0.954
Restricted Randomesque (9)	0.947	0.939	0.953
Modified Within .10 Logits (3)	0.956	0.952	0.959
Modified Within .10 Logits (6)	0.951	0.946	0.956
Modified Within .10 Logits (9)	0.949	0.942	0.956
Rest. Modified Within .10 Logits (3)	0.950	0.943	0.955
Rest. Modified Within .10 Logits (6)	0.949	0.940	0.954
Rest. Modified Within .10 Logits (9)	0.947	0.940	0.954

Table 12: Descriptive Statistics for the Correlation between Known and Estimated Theta Values for the 200 Item Pool Averaged Across Ten Replications

Exposure Control Procedure	Correlation Coefficient		
	Mean	Minimum	Maximum
No Exposure Control	0.962	0.958	0.965
Progressive Restricted	0.957	0.953	0.962
Randomesque (3)	0.960	0.954	0.963
Randomesque (6)	0.957	0.949	0.961
Randomesque (9)	0.954	0.951	0.957
Restricted Randomesque (3)	0.959	0.956	0.962
Restricted Randomesque (6)	0.956	0.950	0.961
Restricted Randomesque (9)	0.954	0.950	0.958
Modified Within .10 Logits (3)	0.960	0.953	0.963
Modified Within .10 Logits (6)	0.957	0.951	0.962
Modified Within .10 Logits (9)	0.954	0.950	0.960
Rest. Modified Within .10 Logits (3)	0.959	0.952	0.963
Rest. Modified Within .10 Logits (6)	0.955	0.944	0.960
Rest. Modified Within .10 Logits (9)	0.954	0.950	0.958

Table 13: Descriptive Statistics for Average Absolute Difference (AAD), Bias, Root Mean Squared Error (RMSE), and Standardized Root Mean Squared Error (SRMSE) for the 100 Item Pool Averaged Across Ten Replications

Exposure Control Procedure	AAD Mean (Min, Max)	Bias Mean (Min, Max)	RMSE Mean (Min, Max)	SRMSE Mean (Min, Max)
No Exposure Control	0.233 (0.228, 0.238)	-0.010 (-0.036, 0.001)	0.302 (0.290, 0.319)	0.531 (0.516, 0.552)
Progressive Restricted	0.260 (0.253, 0.266)	-0.015 (-0.030, 0.003)	0.337 (0.327, 0.343)	0.548 (0.533, 0.571)
Randomesque (3)	0.246 (0.238, 0.256)	-0.015 (-0.029, 0.002)	0.319 (0.309, 0.336)	0.543 (0.526, 0.561)
Randomesque (6)	0.262 (0.253, 0.274)	-0.015 (-0.026, -0.008)	0.339 (0.329, 0.352)	0.556 (0.538, 0.579)
Randomesque (9)	0.268 (0.256, 0.279)	-0.018 (-0.030, 0.002)	0.347 (0.329, 0.363)	0.563 (0.544, 0.583)
Restricted Randomesque (3)	0.259 (0.251, 0.275)	-0.016 (-0.034, 0.001)	0.335 (0.325, 0.354)	0.555 (0.538, 0.577)
Restricted Randomesque (6)	0.268 (0.261, 0.280)	-0.016 (-0.035, -0.005)	0.347 (0.334, 0.360)	0.563 (0.543, 0.584)
Restricted Randomesque (9)	0.268 (0.261, 0.275)	-0.019 (-0.034, -0.006)	0.349 (0.341, 0.373)	0.564 (0.540, 0.581)
Modified Within .10 Logits (3)	0.246 (0.234, 0.256)	-0.016 (-0.027, 0.001)	0.318 (0.304, 0.338)	0.541 (0.525, 0.557)
Modified Within .10 Logits (6)	0.261 (0.256, 0.268)	-0.016 (-0.026, -0.001)	0.339 (0.328, 0.347)	0.554 (0.535, 0.571)
Modified Within .10 Logits (9)	0.268 (0.253, 0.284)	-0.018 (-0.039, -0.002)	0.347 (0.334, 0.358)	0.561 (0.534, 0.581)
Rest. Modified Within .10 Logits (3)	0.259 (0.254, 0.265)	-0.015 (-0.031, 0.003)	0.335 (0.329, 0.345)	0.557 (0.536, 0.582)
Rest. Modified Within .10 Logits (6)	0.265 (0.252, 0.277)	-0.017 (-0.034, -0.003)	0.344 (0.321, 0.358)	0.559 (0.537, 0.586)
Rest. Modified Within .10 Logits (9)	0.273 (0.264, 0.290)	-0.020 (-0.048, 0.002)	0.354 (0.343, 0.366)	0.565 (0.537, 0.584)

Table 14: Descriptive Statistics for Average Absolute Difference (AAD), Bias, Root Mean Squared Error (RMSE), and Standardized Root Mean Squared Error (SRMSE) for the 200 Item Pool Averaged Across Ten Replications

Exposure Control Procedure	AAD Mean (Min, Max)	Bias Mean (Min, Max)	RMSE Mean (Min, Max)	SRMSE Mean (Min, Max)
No Exposure Control	0.226 (0.220, 0.232)	-0.011 (-0.020, -0.001)	0.292 (0.286, 0.304)	0.521 (0.501, 0.543)
Progressive Restricted	0.243 (0.230, 0.254)	-0.013 (-0.024, 0.000)	0.316 (0.306, 0.333)	0.536 (0.515, 0.553)
Randomesque (3)	0.232 (0.226, 0.241)	-0.012 (-0.026, 0.001)	0.300 (0.292, 0.312)	0.529 (0.512, 0.553)
Randomesque (6)	0.243 (0.232, 0.255)	-0.015 (-0.029, -0.001)	0.315 (0.296, 0.334)	0.537 (0.515, 0.565)
Randomesque (9)	0.248 (0.240, 0.258)	-0.010 (-0.025, 0.004)	0.322 (0.313, 0.339)	0.545 (0.527, 0.560)
Restricted Randomesque (3)	0.236 (0.228, 0.247)	-0.012 (-0.037, 0.005)	0.304 (0.295, 0.316)	0.532 (0.513, 0.547)
Restricted Randomesque (6)	0.246 (0.236, 0.253)	-0.013 (-0.034, 0.005)	0.316 (0.308, 0.330)	0.540 (0.517, 0.563)
Restricted Randomesque (9)	0.249 (0.242, 0.258)	-0.013 (-0.024, 0.002)	0.323 (0.316, 0.336)	0.545 (0.528, 0.562)
Modified Within .10 Logits (3)	0.233 (0.220, 0.242)	-0.012 (-0.022, -0.003)	0.300 (0.286, 0.316)	0.529 (0.511, 0.556)
Modified Within .10 Logits (6)	0.240 (0.231, 0.251)	-0.015 (-0.026, -0.004)	0.310 (0.300, 0.328)	0.535 (0.510, 0.559)
Modified Within .10 Logits (9)	0.250 (0.236, 0.265)	-0.011 (-0.022, 0.000)	0.324 (0.305, 0.347)	0.547 (0.519, 0.567)
Rest. Modified Within .10 Logits (3)	0.237 (0.226, 0.247)	-0.012 (-0.021, 0.011)	0.304 (0.290, 0.315)	0.532 (0.511, 0.559)
Rest. Modified Within .10 Logits (6)	0.243 (0.23, 0.256)	-0.015 (-0.033, 0.000)	0.315 (0.299, 0.351)	0.541 (0.521, 0.578)
Rest. Modified Within .10 Logits (9)	0.249 (0.239, 0.257)	-0.013 (-0.030, 0.003)	0.323 (0.310, 0.340)	0.546 (0.528, 0.566)

exposure control procedures did not significantly impact measurement precision in comparison to the no exposure control condition.

For the 200 item pool AAD ranged from 0.230 to 0.250, Bias ranged from -0.015 to -0.010, RMSE ranged from 0.292 to 0.324, and SRMSE ranged from 0.521 to 0.547. Here again the no exposure control condition shared the lowest values of each of the indices of measurement precision. The nine item variant of the modified within .10 logits procedure produced the highest levels of AAD, RMSE, and SRMSE. The nine item variant of the randomesque procedure produced the most bias. As in the 100 item pool the introduction of the exposure control procedures did not significantly impact measurement precision in comparison to the no exposure control condition.

Appendix A contains conditional bias plots for both the 100 and 200 item pool intended to convey the level of bias at 17 discrete points along the known theta scale. Recall that bias is the average difference between estimated and known theta. Therefore in these plots the values at the 17 discrete points represent the average difference of estimated and known theta across replications. From each of these plots it is clear that regardless of item pool the lowest levels of bias, and therefore lowest average difference between estimated and known theta, was seen at and around the average ability level of 0.0. At the extreme values of the theta scale we can see that across replications there was a greater variability in the bias values. This represents the expected phenomenon of decreased measurement precision for extreme scores. We can also see that this variability was not systematically higher or lower indicating that there was no systematic bias at the extremes. From a visual inspection it appears that the variability of bias in the upper extreme of theta was greater than in the lower extreme of theta.

Appendix B contains conditional standard error plots intended to convey the level of standard error at 17 discrete points on the theta scale. Recall that standard error is the reciprocal of the square root of the test information and is a measure of precision at various levels of known theta. Across conditions the plots reveal that regardless of item pool, the lowest levels of standard error were seen in the middle of the distribution around the theta range of -1.0 to -0.5. Across conditions, the extremes of the theta scale yielded higher standard error values. Across conditions it can also been seen that, in general, the standard error values for higher ability were higher than the standard error values for low ability levels indicating worse measurement precision in the upper ability range.

Exposure Rate and Pool Utilization

Table 15 and 16 provide the grand mean, minimum mean, and maximum mean exposure rate for each of the conditions investigated in the 100 and 200 item pool, respectively. The exposure rate of any item is the number of times an item has been administered divided by the number of tests that have been administered. Within each item pool the grand mean of exposure rates was equal. This is a function of the exposure rate being the ratio of test length to item pool size. In a fixed length CAT, as each condition was, the grand mean would be equal. For the 100 item pool the minimum mean exposure rate for the six and nine item variants of the restricted randomesque and restricted within .10 logits procedure along with the progressive restricted procedure were slightly above zero. This is a result of all items in the item pool having been administered at least once.

Table 15: Descriptive Statistics for the Exposure Rates of the 100 Item Pool Averaged Across Ten Replications

Exposure Control Procedure	Item Exposure Rate		
	Grand Mean	Minimum Mean	Maximum Mean
No Exposure Control	0.200	0.000	0.935
Progressive Restricted	0.200	0.050	0.322
Randomesque (3)	0.200	0.000	0.739
Randomesque (6)	0.200	0.000	0.527
Randomesque (9)	0.200	0.000	0.436
Restricted Randomesque (3)	0.200	0.000	0.312
Restricted Randomesque (6)	0.200	0.000[a]	0.314
Restricted Randomesque (9)	0.200	0.004	0.317
Modified Within .10 Logits (3)	0.200	0.000	0.760
Modified Within .10 Logits (6)	0.200	0.000	0.530
Modified Within .10 Logits (9)	0.200	0.000	0.424
Rest. Modified Within .10 Logits (3)	0.200	0.000	0.311
Rest. Modified Within .10 Logits (6)	0.200	0.000[b]	0.314
Rest. Modified Within .10 Logits (9)	0.200	0.004	0.316

[a] Minmum Mean = 0.0002

[b] Minmum Mean = 0.0003

Table 16: Descriptive Statistics for the Exposure Rates of the 200 Item Pool Averaged Across Ten Replications

Exposure Control Procedure	Item Exposure Rate		
	Grand Mean	Minimum Mean	Maximum Mean
No Exposure Control	0.100	0.000	0.927
Progressive Restricted	0.100	0.009	0.320
Randomesque (3)	0.100	0.000	0.718
Randomesque (6)	0.100	0.000	0.486
Randomesque (9)	0.100	0.000	0.367
Restricted Randomesque (3)	0.100	0.000	0.308
Restricted Randomesque (6)	0.100	0.000	0.312
Restricted Randomesque (9)	0.100	0.000	0.315
Modified Within .10 Logits (3)	0.100	0.000	0.732
Modified Within .10 Logits (6)	0.100	0.000	0.504
Modified Within .10 Logits (9)	0.100	0.000	0.383
Rest. Modified Within .10 Logits (3)	0.100	0.000	0.308
Rest. Modified Within .10 Logits (6)	0.100	0.000	0.311
Rest. Modified Within .10 Logits (9)	0.100	0.000	0.312

The no exposure control condition yielded a maximum mean exposure rate of 0.935 indicating that on average 93.5% of the simulees saw some of the same items. The three, six, and nine item variants of the restricted randomesque and restricted modified within .10 logits procedures along with the progressive restricted procedure yielded acceptable maximum exposure rates ranging from 0.311 to 0.322. The three, six, and nine item variants of the randomesque and modified within .10 logits procedures produced moderate to high maximum exposure rates ranging from 0.424 to 0.760. For the 200 item pool maximum exposure rate displayed the same trend.

The no exposure control condition yielded a maximum mean exposure rate of 0.927. The three, six, and nine item variants of the restricted randomesque and restricted modified within .10 logits procedures along with the progressive restricted procedure yielded acceptable maximum exposure rates ranging from 0.308 to 0.320. The three, six, and nine item variants of the randomesque and modified within .10 logits procedures produced moderate to high maximum exposure rates ranging from 0.367 to 0.732.

The grand mean, minimum mean, and maximum mean of exposure rates provide an overview of the exposure rate in the conditions across replications, but they do not convey the level of exposure rate control. For this reason the standard deviation of exposure rates is also provided in Tables 17 and 18 for the two item pools. The standard deviation of exposure rates provides an indication of the even use of items. When the standard deviation of exposure rates is small it indicates that the exposure rates were similar and therefore they were equally administered. Conversely, when the standard deviation of exposure rates is high this suggests that the exposure rates were spread across a wider range and therefore the items were less evenly administered.

Table 17: Descriptive Statistics for the Standard Deviation of Exposure Rates for the 100 Item Pool Averaged Across Ten Replications

Exposure Control Procedure	Standard Deviation of Exposure Rates		
	Mean	Minimum	Maximum
No Exposure Control	0.252	0.249	0.257
Progressive Restricted	0.090	0.088	0.091
Randomesque (3)	0.189	0.186	0.193
Randomesque (6)	0.139	0.136	0.140
Randomesque (9)	0.109	0.107	0.111
Restricted Randomesque (3)	0.110	0.109	0.111
Restricted Randomesque (6)	0.099	0.098	0.100
Restricted Randomesque (9)	0.090	0.089	0.091
Modified Within .10 Logits (3)	0.189	0.186	0.192
Modified Within .10 Logits (6)	0.138	0.136	0.141
Modified Within .10 Logits (9)	0.110	0.109	0.112
Rest. Modified Within .10 Logits (3)	0.110	0.109	0.111
Rest. Modified Within .10 Logits (6)	0.099	0.098	0.100
Rest. Modified Within .10 Logits (9)	0.091	0.090	0.091

Table 18: Descriptive Statistics for the Standard Deviation of Exposure Rates for the 200 Item Pool Averaged Across Ten Replications

Exposure Control Procedure	Standard Deviation of Exposure Rates		
	Mean	Minimum	Maximum
No Exposure Control	0.186	0.183	0.188
Progressive Restricted	0.086	0.086	0.087
Randomesque (3)	0.146	0.145	0.148
Randomesque (6)	0.111	0.109	0.112
Randomesque (9)	0.091	0.089	0.093
Restricted Randomesque (3)	0.112	0.111	0.112
Restricted Randomesque (6)	0.098	0.097	0.099
Restricted Randomesque (9)	0.088	0.086	0.089
Modified Within .10 Logits (3)	0.146	0.144	0.148
Modified Within .10 Logits (6)	0.111	0.110	0.112
Modified Within .10 Logits (9)	0.091	0.089	0.093
Rest. Modified Within .10 Logits (3)	0.112	0.111	0.113
Rest. Modified Within .10 Logits (6)	0.098	0.098	0.099
Rest. Modified Within .10 Logits (9)	0.087	0.086	0.088

For the 100 item pool the no exposure control condition, predictably, resulted in the highest standard deviation of exposure rates (0.252). The progressive restricted procedure as well as the six and nine item variants of the restricted randomesque and restricted modified within .10 logits procedures yielded the most even item usage with standard deviation of exposure rates ranging from 0.090 to 0.099. In the 200 item pool the no exposure control condition once again yielded the most uneven item usage with a standard deviation of exposure rates equal to 0.186. In this larger item pool the six item variants of the restricted randomesque and restricted modified within .10 logits procedures, the nine item variants of the randomesque, modified within .10 logits, restricted randomesque, and restricted modified within .10 logits procedures in addition to the progressive restricted procedure resulted in the most even item usage with standard deviations of exposure rates ranging from 0.086 to 0.098.

In Tables 19 and 20 the average frequency distribution of exposure rates across all ten replications for each condition is provided for the 100 and 200 item pools respectively. For the 100 item pool the maximum information (MI) procedure resulted in the highest percentage of the item pool not being administered (28%). The progressive restricted (PR) procedure and the nine item variants of the restricted randomesque (RRD9) and restricted modified within .10 logits (RMW9) procedure utilized the entire item pool. The three item variants of the restricted randomesque (RRD3) and restricted modified within .10 logits (RMW3) procedures, the six item variants of the randomesque (RD6), restricted randomesque (RRD6), modified within .10 logits (MW6), and restricted modified within .10 logits (RMW6), and the nine item variants of the randomesque (RD9) and modified within .10 logits (MW9) all produced very low percentages of the

Table 19: Frequency of Exposure Rates for the 100 Item Pool Averaged Across Ten Replications

	Exposure Control Condition													
Exposure Rate	MI	PR	RD3	RD6	RD9	RRD3	RRD6	RRD9	MW3	MW6	MW9	RMW3	RMW6	RMW9
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.91-.99	1	0	0	0	0	0	0	0	0	0	0	0	0	0
.81-.90	1	0	0	0	0	0	0	0	0	0	0	0	0	0
.71-.80	4	0	1	0	0	0	0	0	1	0	0	0	0	0
.61-.70	5	0	3	0	0	0	0	0	3	0	0	0	0	0
.51-.60	6	0	5	3	0	0	0	0	5	3	0	0	0	0
.41-.50	4	0	9	8	4	0	0	0	8	7	4	0	0	0
.36-.40	2	0	1	3	7	0	0	0	3	4	8	0	0	0
.31-.35	7	24	4	9	6	29	18	15	3	8	7	28	17	17
.26-.30	4	12	7	11	15	18	25	22	8	13	12	20	25	19
.21-.25	3	13	11	9	17	10	8	17	10	9	16	11	9	15
.16-.20	5	15	10	25	26	11	24	24	10	25	27	9	24	26
.11-.15	4	20	13	6	6	8	5	5	13	5	6	9	5	5
.06-.10	7	16	10	7	8	9	7	9	9	7	7	9	7	9
.01-.05	20	1	13	14	10	13	12	9	13	14	10	13	12	8
0	28	0	14	5	3	2	1	0	14	5	2	3	1	0
(% Not Admin)	(28%)	(0%)	(14%)	(5%)	(3%)	(2%)	(1%)	(0%)	(14%)	(5%)	(2%)	(3%)	(1%)	(0%)

Note: Items in each condition may not sum to 100 due to rounding error.

Table 20: Frequency of Exposure Rates for the 200 Item Pool Averaged Across Ten Replications

	Exposure Control Condition													
Exposure Rate	MI	PR	RD3	RD6	RD9	RRD3	RRD6	RRD9	MW3	MW6	MW9	RMW3	RMW6	RMW9
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.91-.99	1	0	0	0	0	0	0	0	0	0	0	0	0	0
.81-.90	1	0	0	0	0	0	0	0	0	0	0	0	0	0
.71-.80	2	0	2	0	0	0	0	0	2	0	0	0	0	0
.61-.70	3	0	0	0	0	0	0	0	1	0	0	0	0	0
.51-.60	5	0	3	0	0	0	0	0	3	1	0	0	0	0
.41-.50	8	0	7	5	0	0	0	0	7	5	0	0	0	0
.36-.40	7	0	4	6	2	0	0	0	5	5	2	0	0	0
.31-.35	4	14	5	2	9	26	16	10	5	3	8	27	16	9
.26-.30	3	6	6	10	5	7	9	7	5	11	6	5	10	8
.21-.25	5	9	11	14	10	14	15	9	10	14	12	12	16	11
.16-.20	6	13	10	9	23	12	10	24	12	9	25	15	9	25
.11-.15	10	28	20	30	37	20	27	37	17	27	31	19	27	34
.06-.10	9	59	22	36	43	24	40	44	22	37	47	25	38	47
.01-.05	32	72	44	50	41	48	49	40	46	52	38	48	52	38
0	106	0	66	38	31	50	34	29	65	37	31	49	32	29
(% Not Admin)	(53%)	(0%)	(33%)	(19%)	(16%)	(25%)	(17%)	(15%)	(33%)	(19%)	(16%)	(25%)	(16%)	(15%)

Note: Items in each condition may not sum to 200 due to rounding error.

pool that were not used, ranging from 1% to 5%. The three item variants of the randomesque (RD3) and modified within .10 logits (MW3) procedures both resulted in 14% of the item pool not being administered. With the exceptions of the three item variants of the randomesque and modified within .10 logits procedures, who reduced the percent of pool unadministered from 28% in the no exposure control condition to 14%, all exposure control conditions investigated significantly reduced the percentage of the item pool that were not administered.

Looking at the frequency distribution of exposure rates it is possible to determine the percentage of items in the item pool whose exposure rates are moderate to high. Exposure rates were defined as moderate to high if their exposure rate was above 0.35. The exposure control procedures that condition item administration on exposure rate had no items with an exposure rate above an exposure rate of 0.35. These procedures were the progressive restricted procedure and the three, six, and nine item variants of the restricted randomesque and restricted modified within .10 logits procedures.

The no exposure control condition resulted in 23% of the item pool having items whose exposure rates were moderate to high. The three, six, and nine item variants of the randomesque procedure resulted in 19%, 14%, and 11% of the item pool having items whose exposure rates exceeded 0.35. The three, six, and nine item variants of the modified within .10 logits procedure resulted in 20%, 14%, and 12% of the item pool having items whose exposure rates exceeded 0.35. It is interesting to note that in the randomization procedures the increase in item group size had a more significant impact on decreasing the percentage of the pool that was never administered than it did on reducing the percentage of items whose exposure rates were moderate to high.

In the 200 item pool the no exposure control, maximum information, condition again resulted in the highest percentage of the item pool not being administered (53%). This means that despite the availability of 200 items, in this condition trait estimation was occurring with a functional item pool of less than 100 items. In this larger item pool only the progressive restricted procedure was able to maintain complete utilization of the item pool. The three, six, and nine item variants of the randomesque and modified within .10 logits procedures resulted in 33%, 19%, and 16% of the item pool not being administered. The three and nine item variants of the restricted randomesque and restricted modified within .10 logits procedure resulted in 25% and 15% of the item pool not having been administered. The six item variants of the restricted randomesque and restricted modified within .10 logits procedure resulted in 16% and 17% of the item pool not being administered.

In this item pool the no exposure control, maximum information, condition resulted in 14% of the item bank having exposure rates above 0.35. In comparison to the 100 item pool, the progressive restricted procedure as well as the combinatorial procedures, that condition item administration on exposure rate, maintained the percentage of items whose exposure rate exceeded 0.35 at 0%. The three item variants of the randomesque and modified within .10 logits procedures resulted in 8% and 9% of the item pool having exposure rates that were moderate or high. The six and nine item variants of the randomesque and modified within .10 logits procedures resulted in 19% and 16% of the item pool having moderate or high exposure rates.

Item Overlap

To determine the amount of items shared by simulees the audit trails of every simulee was compared to the audit trail of every other simulee. Three indices of item overlap were examined. First, the overall item overlap grand mean provides an index of overlap regardless of the abilities of the simulees. Secondly, if simulee abilities are within 1 logit they were considered to have similar abilities and their grand mean overlap rates were reported. Finally if simulees have differences in ability greater than 1 logit they were considered to have difference abilities and their grand mean overlap rates were reported. These results are reported in Tables 21 and 22 for the 100 and 200 item pools.

In the 100 item pool the highest grand mean overall overlap occurred for the no exposure control condition where on average simulee tests shared 51% of items overall. The three item variants of the randomesque and modified within .10 logits procedure resulted in an overall overlap grand mean of 38%. The next largest grand mean overlap was demonstrated by the six item variants of the randomesque and modified within .10 logits procedure, 29%. The progressive restricted procedure, nine item variants of the randomesque and modified within .10 logits, and the three, six, and nine item variants of the combinatorial procedures each yielded similar grand mean overall overlap ranging from 24% to 26%.

For simulees with similar abilities the no exposure control condition produced grand mean overlap of 32% of items. The three item variants of the randomesque and modified within .10 logits procedure both yielded a grand mean item overlap for simulees with similar abilities of 28%. The six and nine item variants of the randomesque and modified within .10 logits procedure both yielded a grand mean item overlap for simulees

Table 21: Descriptive Statistics for Item Overlap of the 100 Item Pool Averaged Across Ten Replications

Exposure Control Condition	Item Overlap		
	Overall Overlap Grand Mean (Min, Max)	Similar Abilities Grand Mean (Min, Max)	Different Abilities Grand Mean (Min, Max)
No Exposure Control	10.288 (51%) (10.103, 10.553)	6.383 (32%) (6.168, 6.577)	13.915 (70%) (13.838, 14.068)
Progressive Restricted	4.777 (24%) (4.756, 4.801)	3.857 (19%) (3.825, 3.930)	5.662 (28%) (5.592, 5.736)
Randomesque (3)	7.526 (38%) (7.401, 7.662)	5.543 (28%) (5.440, 5.670)	9.381 (47%) (9.305, 9.475)
Randomesque (6)	5.892 (29%) (5.828, 5.932)	5.000 (25%) (4.904, 5.052)	6.730 (34%) (6.695, 6.754)
Randomesque (9)	5.150 (26%) (5.128, 5.205)	4.631 (23%) (4.606, 4.701)	5.637 (28%) (5.617, 5.665)
Restricted Randomesque (3)	5.186 (26%) (5.164, 5.213)	4.043 (20%) (3.990, 4.108)	6.257 (31%) (6.150, 6.336)
Restricted Randomesque (6)	4.953 (25%) (4.942, 4.965)	4.305 (22%) (4.254, 4.341)	5.562 (28%) (5.483, 5.629)
Restricted Randomesque (9)	4.785 (24%) (4.769, 4.805)	4.346 (22%) (4.324, 4.380)	5.197 (26%) (5.173, 5.219)
Modified Within .10 Logits (3)	7.524 (38%) (7.418, 7.636)	5.531 (28%) (5.445, 5.627)	9.388 (47%) (9.319, 9.446)
Modified Within .10 Logits (6)	5.875 (29%) (5.811, 5.944)	4.974 (25%) (4.893, 5.035)	6.724 (34%) (6.690, 6.746)
Modified Within .10 Logits (9)	5.191 (26%) (5.155, 5.235)	4.627 (23%) (4.583, 4.694)	5.722 (29%) (5.698, 5.743)
Rest. Modified Within .10 Logits (3)	5.186 (26%) (5.165, 5.214)	4.050 (20%) (3.959, 4.093)	6.247 (31%) (6.171, 6.335)
Rest. Modified Within .10 Logits (6)	4.949 (25%) (4.930, 4.966)	4.294 (21%) (4.234, 4.335)	5.567 (28%) (5.522, 5.616)
Rest. Modified Within .10 Logits (9)	4.795 (24%) (4.785, 4.809)	4.327 (22%) (4.296, 4.352)	5.237 (26%) (5.188, 5.292)

Table 22: Descriptive Statistics for Item Overlap of the 200 Item Pool Averaged Across Ten Replications

	Item Overlap		
Exposure Control Condition	Overall Overlap Grand Mean (Min, Max)	Similar Abilities Grand Mean (Min, Max)	Different Abilities Grand Mean (Min, Max)
No Exposure Control	8.841 (44%) (8.661, 9.016)	4.936 (25%) (4.771, 5.075)	12.493 (62%) (12.428, 12.572)
Progressive Restricted	3.467 (17%) (3.45, 3.498)	2.283 (11%) (2.261, 2.312)	4.598 (23%) (4.537, 4.672)
Randomesque (3)	6.224 (31%) (6.15, 6.338)	3.921 (20%) (3.854, 4.01)	8.38 (42%) (8.272, 8.432)
Randomesque (6)	4.448 (22%) (4.368, 4.5)	3.11 (16%) (3.054, 3.165)	5.706 (29%) (5.657, 5.767)
Randomesque (9)	3.628 (18%) (3.569, 3.686)	2.782 (14%) (2.701, 2.856)	4.423 (22%) (4.389, 4.456)
Restricted Randomesque (3)	4.462 (22%) (4.423, 4.497)	2.773 (14%) (2.683, 2.842)	6.045 (30%) (5.951, 6.104)
Restricted Randomesque (6)	3.911 (20%) (3.865, 3.941)	2.759 (14%) (2.705, 2.802)	4.995 (25%) (4.97, 5.029)
Restricted Randomesque (9)	3.51 (18%) (3.467, 3.553)	2.691 (13%) (2.621, 2.753)	4.282 (21%) (4.252, 4.304)
Modified Within .10 Logits (3)	6.239 (31%) (6.103, 6.334)	3.955 (20%) (3.81, 4.053)	8.371 (42%) (8.322, 8.451)
Modified Within .10 Logits (6)	4.44 (22%) (4.372, 4.502)	3.098 (15%) (3.043, 3.15)	5.701 (29%) (5.668, 5.728)
Modified Within .10 Logits (9)	3.629 (18%) (3.576, 3.692)	2.782 (14%) (2.726, 2.863)	4.42 (22%) (4.393, 4.439)
Rest. Modified Within .10 Logits (3)	4.464 (22%) (4.43, 4.504)	2.778 (14%) (2.732, 2.844)	6.041 (30%) (5.94, 6.116)
Rest. Modified Within .10 Logits (6)	3.913 (20%) (3.877, 3.946)	2.754 (14%) (2.717, 2.789)	5 (25%) (4.949, 5.034)
Rest. Modified Within .10 Logits (9)	3.5 (18%) (3.462, 3.54)	2.681 (13%) (2.623, 2.748)	4.268 (21%) (4.239, 4.318)

with similar abilities ranging from 23% to 25%. The progressive restricted procedure and the three, six, and nine item variants of the combinatorial procedures each yielded a grand mean item overlap for simulees with similar abilities ranging from 19% to 22%.

When focusing on the simulees who had different abilities the no exposure control condition resulted in a grand mean of 70% of items being shared. The three item variants of the randomesque and modified within .10 logits procedures yielded the next largest amount of item overlap for simulees with different abilities, 47% of items. The six item variants of the randomesque and modified within .10 logits procedures (34%) and the three item variants of the combinatorial procedures (31%) yielded lower grand mean overlap rates. The nine item variants of the randomesque and modified within .10 logits procedures, the six item variants of the combinatorial procedures, and the progressive restricted procedure resulted in similar grand mean overlap rates (28% to 29%). The nine item variants of the combinatorial procedures produced the lowest grand mean overlap rates for simulees with different abilities (26%).

In the 200 item pool the no exposure control conditions resulted in an overall item overlap grand mean of 44% of items. The three item variants of the randomesque and modified within .10 logits procedures yielded the next largest overall overlap grand mean with simulees sharing an average of 31% of items. The three item variants of the combinatorial procedures as well as the six item variants of the randomesque and modified within .10 logits procedures reduced the overlap rate grand mean to 22%. The six item variants of the combinatorial procedures further reduced the overall overlap grand mean to 20%. The nine item variants of the combinatorial procedures produced a low overlap rate grand mean of 18%. Finally, the progressive restricted procedure yielded the lowest overall item overlap grand mean with simulees sharing 17% of items.

For simulees with similar abilities the no exposure control condition produced an overlap rate grand mean of 25% of items. The three item variants of the randomesque and modified within .10 logits procedures reduced this overlap rate grand mean slightly (20%). The six item variants of the randomesque and modified within .10 logits procedures further reduced this overlap rate grand mean to 15% to 16%. The three and six item variants of the combinatorial procedures along with the nine item variants of the randomesque and modified within .10 logits procedures each resulted in an overlap rate grand mean for simulees with similar abilities of 14%. The nine item variants of the combinatorial procedures yielded the second lowest level of overlap producing an overlap rate grand mean of 13%. The progressive restricted procedure yielded the lowest item overlap grand mean for simulees with similar abilities where simulees shared 11% of items.

For simulees with different abilities the no exposure control condition again produced the highest overlap grand mean, 62%. The three item variants of the randomesque and modified within .10 logits procedures yielded the next largest grand mean overlap rates for simulees with different abilities, sharing 42% of items. The three item variants of the combinatorial procedures (30%) along with the six item variants of the randomesque and modified within .10 logits procedures (29%) resulted in similar grand mean overlap rates among simulees with different abilities. The six item variants of the combinatorial procedures (25%) and the progressive restricted (23%) yielded similar grand mean overlap rates for simulees with different abilities. For the simulees with different abilities the nine item variants of the randomesque and modified within .10 logits (22%) along with the nine item variants of the combinatorial procedures (21%) produced the lowest grand mean overlap among simulees with different abilities.

Chapter V: Discussion

The present chapter shall take as its primary goal the exploration of the results detailed in the previous chapter in an effort to draw conclusions to the five research questions previously posed. Relevant data is discussed as a method to support the conclusions being drawn. Subsequent to a discussion of each of the research questions a brief discussion of the conclusions that can be drawn from the research questions will be offered. Finally, limitations of the present research along with suggestions for future research are offered.

Research Questions

What effect will the implementation of a maximum exposure rate constraint have on the performance of the modified within .10 logits procedure in regards to measurement precision and test security?

The results reported in the previous section provide evidence to the effect that the restricted modified within .10 logits procedure is an advantageous evolution of the modified within .10 logits procedure. The distribution of estimated thetas and standard errors between the base model and its restricted form were functionally equivalent in both item pools. The indices of recovery of known theta, average absolute difference, bias, root mean squared error, and standardized root mean squared error also showed no appreciable difference when the base model was compared to its restricted form. From this it is justifiable to conclude that the implementation of a maximum exposure rate constraint on the modified within .10 logits procedure did not negatively impact measurement precision in either item pool.

In the 100 item pool, the three item variant of the modified within .10 logits procedure produced an average maximum exposure rate of 0.760 while its restricted from produced an average maximum exposure rate of 0.311. The six item variant of the base procedure produced an average maximum exposure rate of 0.530 while its restricted form reduced this number to 0.314. The nine item variant of the base model resulted in an average maximum exposure rate of 0.424 and its restricted analog reduced this to 0.316. In every item variant the restricted procedure was successful in reducing the average maximum exposure rate produced by its unrestricted counterpart.

In regards to the standard deviation of exposure rates the restricted modified within .10 logits procedure yielded more even item usage than the modified within .10 logits procedure. The three item variant of the base procedure yielded an average standard deviation of exposure rates of 0.189 while the restricted form yielded an average standard deviation of exposure rates of 0.110. The six item variants resulted in values of 0.138 and 0.099 for the base and restricted procedures. The nine item variants resulted in values of 0.110 and 0.091 for the base and restricted procedures.

For pool utilization and percent of items with moderate to high exposure rates the same pattern of improvement is seen. The three item variant of the base procedure resulted in 14% of the item pool not being administered and 20% of the item pool having moderate to high exposure rates while the restricted procedure reduced these to 3% and 0%. For the six item variants these values were 5% and 14% for the base procedure and 1% and 0% for the restricted form. The nine item variant produced values of 2% and 12% while the restricted form surprisingly used 100% of the item pool while maintaining all items below an exposure rate of 0.36.

This pattern of improvement is also seen in average overlap rate, average overlap rate for simulees with similar abilities, and average overlap rate for simulees with different abilities. In the three item variant, the base procedure resulted in overlap rates of 38%, 28%, and 47% for average overlap, average overlap for simulees with similar abilities, and average overlap for simulees with different abilities while the restricted form reduced these values to 26%, 20%, and 31%. The six item variant produced overlap rates of 29%, 25%, and 34% with the restricted form producing overlap rates of 25%, 21%, and 28%. The nine item variant yielded overlap rates of 26%, 23%, and 29% with the restricted form reducing these rates to 24%, 22%, and 26%.

In the 200 item pool the three item variant of the modified within .10 logits procedure model resulted in an average maximum exposure rate of 0.732 while the restricted modified within .10 logits procedure reduced this number to 0.308. The six item variant of the base model resulted in an average maximum exposure rate 0.504 while its restricted counterpart reduced this to 0.311. In the nine item variant the values for the two procedures were 0.383 and 0.312. In the average standard deviation of the exposure rates the trend of improvement continues. The three item variant of the restricted form of the base procedure was able to reduce the average standard deviation of the exposure rates from 0.146 to 0.112, the six item variant of the restricted procedure reduced this value from 0.111 to 0.098, and the nine item variant reduced this value from 0.091 to 0.087.

In terms of pool utilization and percent of items with moderate to high exposure rates the three item variant of the base procedure resulted in 33% of the item pool not being administered and 9% of the item pool having moderate to high exposure rates while the restricted procedure reduced these to 25% and 0%. For the six item variants these values were 19% and 6% for the base procedure and 16% and 0% for the restricted form. The nine item variant produced values of 16% and 1% while the restricted form resulted in only 15% of the pool not being administered and no items having moderate or high exposure rates.

This pattern of improvement was also present for average overlap rate, average overlap rate for simulees with similar abilities, and average overlap rate for simulees with different abilities. In the three item variant the base procedure resulted in overlap rates of 31%, 20%, and 42% for average overlap, average overlap for simulees with similar abilities, and average overlap for simulees with different abilities while the restricted form reduced these values to 22%, 14%, and 30%. The six item variant produced overlap rates of 22%, 15%, and 29% with the restricted form producing overlap rates of 20%, 14%, and 25%. The nine item variant yielded overlap rates of 18%, 14%, and 22% with the restricted form reducing these rates to 18%, 13%, and 21%.

Given these results we can see that the implementation of an exposure rate constraint in the restricted modified within .10 logits procedure did not negatively impact measurement precision in comparison to its base procedure. In terms of test security modest to significant improvements were observed. More significant gains were seen in the 100 item pool than in the larger 200 item pool. Given these improvements and the ease of implementation of the restricted form of the modified within .10 logits procedure the current research study finds clear support for the use of the restricted modified within .10 logits procedure over the modified within .10 logits procedure.

What effect will the implementation of a maximum exposure rate constraint have on the performance of the randomesque procedure in regards to measurement precision and test security?

As with the comparison in the previous research question this question concerns itself only with differences between the randomesque and restricted randomesque procedures. The results of the present research study support the conclusion that the restricted form of the base procedure increases test security while not significantly negatively impacting measurement precision. The distribution of estimated theta values in conjunction with the measurement precision indices of recovery of known theta, average absolute difference, bias, root mean squared error, and standardized root mean squared error suggest that the restricted randomesque procedure provided equivalent levels measurement precision as the randomesque procedure.

In the 100 item pool the restricted randomesque procedure was able to reduce the average maximum exposure rate from 0.739 to 0.312 in the three item variant, from 0.527 to 0.314 in the six item variant, and from 0.436 to 0.317 in the nine item variant. In this item pool the six and nine item variants of the restricted procedure raised the average minimum exposure rate above 0.0 indicating that in some replications all items in the item pool were administered. The restricted randomesque procedure also reduced the standard deviation of exposure rates from 0.189 to 0.110 in the three item variant, from 0.139 to 0.099 in the six item variant, and from 0.109 to 0.090 in the nine item variant.

The restrict procedure displays significant improvements in pool utilization and in the percentage of the pool whose items resulted in a moderate to high exposure rate. The percent of the item pool whose items were never administered were reduced from 14% to

2% in the three item variant, from 5% to 1% in the six item variant, from 3% to 0% in the nine item variant. In the randomesque procedures the percentage of items with a moderate to high exposure rate ranged from 11% to 19% while all of the item variants of the restricted randomesque procedure were successful in preventing items in the item pool from exceeding an exposure rate of 0.35.

The restricted randomesque procedure was also able to produce improved average overall overlap rates, average overlap rates for simulees with similar abilities, and average overlap rates for simulees with different abilities in comparison to the randomesque procedure. The restricted randomesque procedure reduced the average overall overlap rate from 38% to 26% in the three item variant, from 29% to 25% in the six item variant, and from 26% to 24% in the nine item variant. The restricted randomesque procedure reduced the average overlap rate for simulees with similar abilities from 28% to 20% in the three item variant, from 25% to 22% in the six item variant, and from 23% to 22% in the nine item variant. The restricted randomesque procedure reduced the average overlap rate for simulees with different abilities from 47% to 31% in the three item variant, from 34% to 28% in the six item variant, and from 28% to 26% in the nine item variant.

In the 200 item pool, the restricted randomesque procedure was also able to reduce the high average maximum exposure rates produced by the randomesque procedure while demonstrating more even use of the item pool. The restricted randomesque procedure reduced the average maximum exposure rate from 0.718 to 0.308 in the three item variants, from 0.486 to 0.312 in the six item variant, and from 0.367 to 0.315 in the nine item variant. The restricted randomesque procedure reduced the standard deviation of exposure rates from 0.146 to 0.112 in the three item variant, from 0.111 to 0.098 in the six item variant, and from 0.091 to 0.088 in the nine item variant.

Improvements were also demonstrated in the percentage of the item pool that was never administered as well as in the percentage of items in the item pool with moderate to high exposure rates. The restricted randomesque procedure reduced the percent of the item pool that was unadministered from 33% to 25% in the three item variant, from 19% to 17% in the six item variant, and from 16% to 15% in the nine item variant. The randomesque procedure resulted in the percent of items above an exposure rate of 0.35 of 8% in the three item variant, 6% in the six item variant, and 1% in the nine item variant while all item variants of the restricted randomesque procedure reduced these percentages to 0%.

In comparison to the randomesque procedure modest improvements were seen by the restricted randomesque procedure in the average overall overlap rate, average overlap rate for simulees with similar abilities, and average overlap rate for simulees with different abilities. In the average overall overlap rate the restricted randomesque procedure was able to reduce this rate from 31% to 22% in the three item variant, from 22% to 20% in the six item variant. Both nine item variants produced an average overall overlap rate of 18%. For simulees with similar abilities the average overlap rate was reduced from 20% to 14% in the three item variant, from 16% to 14% in the six item variant, and from 14% to 13% in the nine item variant. The average overlap rate for simulees with different abilities was reduced from 42% to 30% in the three item variant, from 29% to 25% in the six item variant, and from 22% to 21% in the nine item variant.

With the exception of the average overall overlap rate of the nine item variants of the restricted randomesque procedure and the randomesque procedure in the 200 item pool, every measure of test security demonstrated improvement in the restricted from of the randomesque procedure. These improvements were produced with no significant degradation of measurement precision. This, in conjunction with the ease of implementation of the exposure rate constraint, argues strongly for the use of the restricted randomesque procedure over the randomesque procedure.

What impact will item group size have on the randomesque, restricted randomesque, modified within .10 logits, and restricted modified within .10 logits procedures in regards to measurement precision and test security?

Item group size appeared to produce insignificantly small differences in the indices of measurement precision. In the indices of average absolute difference, bias, root mean squared error, and standardized root mean squared error as item group size increased minor increases were generally seen however these variances were so slight as to not be of major concern. For example, in the 100 item pool the largest difference observed within item group variants in average absolute difference, bias, root mean squared error, or standardized root mean squared error was 0.022, 0.005, 0.029, and 0.02. In the 200 item pool the largest difference observed within item group variants in average absolute difference, bias, root mean squared error, or standardized root mean squared error was 0.017, 0.005, 0.024, and 0.018. It is judged, therefore, that the use of an item group size of three, six, or nine did not significantly impact measurement precision in any of the four procedures utilizing item group variants.

In both item pools as item group size increased the average maximum exposure rate decreased in the randomesque procedure. For the 100 item pool as item group size increased from three to six and from three to nine the average maximum exposure rate decreased 29% and 41%. In this procedure when item group size was increased from six to nine the average maximum exposure rate decreased 17%. As item group size increased from three to six and from three to nine, in the randomesque procedure with an item pool of 200, the average maximum exposure rate decreased 32% and 49%. In this procedure when item group size was increased from six to nine the average maximum exposure rate decreased 24%.

For the 100 item pool with the modified within .10 logits procedure as item group size increased from three to six and from three to nine the average maximum exposure rate decreased 30% and 44%. In this procedure when item group size was increased from six to nine the average maximum exposure rate decreased 20%. In the case of the 200 item pool as item group size increased from three to six and from three to nine the average maximum exposure rate decreased 31% and 48%. In this procedure when item group size was increased from six to nine the average maximum exposure rate decreased 24%.

In the restricted randomesque and restricted modified within .10 logits procedures the average maximum exposure rate demonstrated slight increases as item group size increased. In the restricted randomesque procedure as item group size increased from three to six and from three to nine the average maximum exposure rate increased 1%. In this procedure when item group size was increased from six to nine the average maximum exposure rate increased 1%. In the restricted modified within .10 logits procedure as item group size increased from three to six and from three to nine the average maximum exposure rate increased 1% and 2%. In this procedure when item group size was increased from six to nine the average maximum exposure rate increased 1%.

In both item pools and in all procedures that utilized item group size variants, as item group size increased the standard deviation of exposure rates decreased. In the 100 item pool as item group size increased from three to six, from three to nine, and from six to nine the standard deviation of exposure rates decreased 27%, 43%, and 22% for the randomesque procedure. For the restricted randomesque procedure these statistics fell 10%, 18%, and 9%. In the modified within .10 logits procedure these statics fell 27%, 42%, and 20%. With the restricted modified within .10 logits procedure these statics dropped 10%, 18%, and 8%. In the randomesque and modified within .10 logits procedure with an item pool of 200 these statistics fell 24%, 38%, and 18%. With the restricted randomesque procedure these same statistics decreased 12%, 21%, and 11%. Increasing the item group size in the modified within .10 logits procedures caused the standard deviation of exposure rates to decrease 24%, 28%, and 18%. In the restricted modified within 10 logits procedure they were reduced 12%, 22%, and 11%.

On the test security index of pool utilization the incrementation of item group size again demonstrated improvements. The percentage of the item pool that was never administered was significantly decreased as item group size increased from three to six in all conditions. In the 100 item pool, as item group size was increased from three to six and finally to nine the randomesque procedure produced lack of pool utilization rates of 14%, 5%, and 3%. For the modified within .10 logits procedure with an item pool of 100 these statistics were 14%, 5%, and 2%. In the 200 item pool the randomesque and modified within .10 logits procedure these statistics were 33%, 19%, and 16%. In the restricted forms of these two procedures with an item pool of 100 all three item variants produced very low percentages of items never administered. For the restricted randomesque procedure as item group size increased these percentages were 2%, 1%, and 0%. For the restricted modified within .10 logits procedure as item group size increased these percentages were 3%, 1%, and 0%. In both of these restricted procedures the nine item variant utilized all of the item pool.

In the 200 item pool, for the restricted randomesque procedure the percent of item unadministered were 25% for the three item variant, 17% for the six item variant, and 15% for the nine item variant. The randomesque procedure produced lack of pool rates of 33%, 19%, and 16%. In this larger item pool the restricted modified with .10 logits procedure resulted in percentages of 25%, 16%, and 15% while the modified within .10 logits procedure resulting percentages of 33%, 19%, and 16%.

Improvements were also seen in all conditions and in both item pools for overlap rates. In the 100 item pool as item group size increased from three to six to nine, the average overall overlap rate for the randomesque procedure decreased from 38% to 29% and finally to 26%. In the 200 item pool average overall overlap rate decreased from 31% to 22% and finally to 18% in the nine item variant. For the restricted randomesque procedure these rates were 26%, 25%, and 24% for the 100 item pool and 22%, 20%, and 18% for the 200 item pool. In the modified within .10 logits procedure these rates were 38%, 29%, and 26% with an item pool of 100 and 31%, 22%, and 18% with an item pool of 200. In the restricted modified within .10 logits procedure these rates were 26%, 25%, and 24% with an item pool of 100 and 22%, 20%, and 18% with an item pool of 200.

The same trend is seen with the average overlap rate for simulees with similar abilities for all four procedures in the 200 item pool and with the randomesque and modified within .10 logits procedure in the 100 item pool. For the randomesque procedure these overlap rates were 28%, 25%, and 23% for the 100 item pool and 20%, 16%, and 14% for the 200 item pool. With the modified within .10 logits procedure these overlap rates were 28%, 25%, and 23% in the 100 item pool and 20%, 15%, and 14% in the 200 item pool. In the 200 item both the restricted randomesque and restricted modified within .10 logits procedure displayed the same trend, though to a lesser extent (14%, 14%, &13%). The restricted randomesque and restricted modified within .10 logits procedure displayed the opposite trend in the smaller 100 item pool. In this item pool the average overlap rate for simulees with similar abilities increased slightly from 20% in the three item variant to 22% in the six and nine item variants. The restricted modified within .10 logits procedure also demonstrated slight increases from 20% in the three item variant to 21% in the six item variant and finally to 22% in the nine item variant. This may suggest the smaller item pool was showing signs of being stressed.

The average overlap rate for simulees with different abilities decreased as item group size increased from three to six to nine for all procedures incorporating item group variants in both item pools. In the randomesque procedure this overlap rate fell from 47% to 34% and finally to 28% in the nine item variant with and item pool of 100 and from 42% to 29% and finally to 22% in the nine item variant with an item pool of 200. In the restricted randomesque procedure this overlap rate decreased from 31% in the three item variant to 28% in the six item variant and down to 26% in the nine item variant in the 100 item pool and from 30% to 25% and finally to 21% in the 200 item pool. With the modified within .10 logits procedure this overlap rate was reduced from 47% to 34% and down to 29% in the 100 item pool and from 42% to 29% and was finally reduced to 22% in the nine item variant with an item pool of 200. For the restricted modified within .10 logits procedure this overlap rate was reduced from 31% to 28% and finally to 26% in the 100 item pool and from 30% to 25% and finally down to 21% in the 200 item pool.

It is clear from all of this that the increase in item group size demonstrated unequivocal improvement in test security while demonstrating only a minor loss of measurement precision. The present research suggests that, in almost every case, for the randomesque, restricted, randomesque, modified within .10 logits procedure, and the restricted modified within .10 logits procedures an item group size of nine will outperform an item group size of six and both will outperform an item group size of three in terms of test security indices. Although there was some degradation in the indices of average maximum exposure rate and average overlap rate for simulees with similar abilities for the restricted randomesque and restricted modified within .10 logits procedures these increases were very minor. Also, while the increases in item group size did demonstrate an expected loss of measurement precision the loss is minor and judged acceptable when weighed against the improvements in test security.

How will measurement precision be affected in the exposure control procedures investigated?

Generally speaking, the current research study found that little to no differences were observed in the exposure control procedures when compared to the no exposure control condition. A brief summary of the slight differences in the observed measures of measurement precision are here given. In terms of the distribution of estimated theta and the recovery of known theta all exposure control conditions performed equivalently to the no exposure control condition.

While all exposure control conditions demonstrated slightly higher levels of average standard error of estimated theta in comparison to the no exposure control condition, within exposure control conditions the variance was slight. In the 100 item pool average standard error ranged from 0.303 to 0.339 with the three item variants of the randomesque and modified within .10 logits procedure producing the lowest statistics and the nine item variant of the restricted modified within .10 logits procedure producing the highest values of this statistic. In the 200 item pool average standard error ranged from 0.286 to 0.311 with the three item variant of the randomesque procedure producing the lowest value of this statistic and the nine item variant of the restricted modified within .10 logits producing the highest value of this statistic.

Average absolute difference (AAD) ranged from 0.246 to 0.273 in the 100 item pool with the lowest value of AAD being produced by the three item variant of the randomesque procedure and the highest value of AAD being produced by the nine item variant of the restricted modified within .10 logits procedure. In the 200 item pool AAD ranged from 0.232 to 0.250 with the three item variant of the randomesque procedure and the nine item variant of the modified within .10 logits procedure producing the lowest and highest value of this statistic. In the 100 item pool bias ranged from -0.020 to -0.015 with the six item variant of the randomesque method being closest to zero and the nine item variant of the restricted modified within .10 logits procedure being furthest from zero. In the 200 item pool bias ranged from -0.015 to -0.010 with the nine item variant of the randomesque procedure being closest to zero and the six item variants of the randomesque, modified within .10 logits, and restricted modified within .10 logits procedure being furthest from zero.

In the 100 item pool root mean squared error ranged from 0.318 to 0.354 and standardized root mean squared error ranged from 0.541 to 0.565 with the three item variant of the modified within .10 logits procedure and the nine item variant of the restricted modified within .10 logits procedure producing the lowest and highest values of these statistics. In the 200 item pool root mean squared error ranged from 0.300 to 0.324 and standardized root mean squared error ranged from 0.529 to 0.547 with the three item variants of the randomesque and modified within .10 logits procedures producing the lowest value and the nine item variant of the modified within .10 logits procedure producing the highest values of these statistics. Across item pools the range of these statistics was very small. In the 100 item pool the difference between the highest and lowest value for AAD, bias, RMSE, and SRMSE was 0.027, 0.005, 0.036, and 0.024. In the 200 item pool the difference between the highest and lowest value for AAD, bias, RMSE, and SRMSE was 0.018, 0.005, 0.024, and 0.018.

While the variance between conditions is relatively minor it is worthy of note that a pattern seems to have emerged between the randomesque procedure and the modified within .10 logits procedure as well as between restricted randomesque procedure and the restricted modified within .10 logits procedure. In the current research when one compares these two pairs of procedures their respective measurement precision indices are almost exactly equal. For example, in the 100 item pool with an item group size of six in standard error, average absolute difference, bias, root mean squared error, and standardized root mean squared error the six item variants of the restricted randomesque and restricted modified within .10 logits procedures produced differences in these statistics of 0.000, 0.003, 0.001, 0.003, and 0.004. With an item group size of nine the two procedures produced differences in statistics of 0.003, 0.005, 0.001, 0.005, and 0.001. These statistics are even closer in the 200 item pool. As another example in the 200 item pool the differences between the randomesque and modified within .10 logits procedures with an item group size of nine for these statistics were 0.000, 0.002, 0.001, 0.002, and 0.002. This suggests that, at least for the present research, the randomesque and modified within .10 logits procedures and the restricted randomesque and restricted modified within .10 logits procedures performed equivalently in terms of measurement precision.

Which exposure control procedure will protect test security in regards to exposure control, pool utilization, and item overlap the best?

To determine which exposure control procedure best controlled exposure rates the indices of average maximum exposure rate and average standard deviation of exposure rates will be examined. To categorize levels of maximum exposure rate let us first define acceptable maximum exposure rates to be an exposure rate of 0.35 or lower, moderate exposure rates to be exposure rates in the range of 0.36 to 0.49, and unacceptably high exposure rates to be any exposure rate equal to or above 0.50. For the current study in the 100 item pool all item variants of the restricted randomesque (0.312, 0.314, & 0.317) and restricted modified within .10 logits (0.311, 0.314, & 0.316) procedures and the progressive restricted procedure (0.322) were successful at controlling maximum exposure rate to an acceptable level. In the 200 item pool, the three, six, and nine item variants of the restricted randomesque (0.308, 0.312, & 0.315) and restricted modified within .10 logits (0.308, 0.311, & 0.312) procedures, and the progressive restricted procedure (0.320) produced acceptable maximum exposure rates. We can also note that in both item pools the six and nine item variants of the restricted modified within .10 logits procedure tended to be slightly lower than the six and nine item variants of the restricted randomesque procedure.

On the index of the average standard deviation of exposure rates three exposure control procedures stand out as providing the best increase in even item usage when compared to the no exposure control condition. Specifically, in the 100 item pool the progressive restricted procedure along with the nine item variants of the restricted randomesque and restricted modified within .10 logits procedures produced the most even item usage. These three procedures reduced the average standard deviation of exposure rates produced by the no exposure control procedure by 64%. In the 200 item pool the progressive restricted procedure and nine item variants of the restricted randomesque and restricted modified within .10 logits procedures demonstrated the most significant improvement in even item usage ranging in a reduction of the average standard deviation of exposure rate from 53% in the progressive restricted procedure and 54% for both restricted procedures.

In terms of the utilization of the item pool the same three procedures emerged as the top performers. In the 100 item pool the progressive restricted procedure and the nine item variants of the randomesque and modified within .10 logits procedure were successful at utilizing the entire item pool. In the 200 item pool only the progressive restricted procedure was successful at utilizing the complete item pool. The nine item variants of the restricted randomesque and restricted modified within .10 logits procedures produced the next lowest lack of pool use with 15% of the item pool not being administered. While the three procedures performed equivalently in this regard in the 100 item pool the progressive restricted was clearly superior in the larger 200 item pool.

The superior performance of the progressive restricted procedure is explained by the weighting algorithm employed. In the initial stages of the CAT the weighting algorithm deemphasizes the impact of item information in item selection. This results in greater pool utilization at this stage of the test in comparison to the restricted modified within .10 logits and restricted randomesque procedures whose item selection algorithms rely entirely upon item information at this stage of the test. In later stages the three procedures perform similarly since information has a greater emphasis in later stages of the CAT when using the progressive restricted procedure. For this reason the progressive restricted procedure should always provide better pool utilization in comparison to the restricted modified within .10 logits and restricted randomesque procedures.

In terms of overall average item overlap for the 100 item pool the progressive restricted and the nine item variants of the restricted randomesque and restricted modified within .10 logits procedures produced the lowest overlap rate percentage with 24%. In the 200 item pool the progressive restricted procedure produced the lowest average overall item overlap with 17% of items being shared. This was followed closely by the nine item variants of the randomization procedures and the nine item variants of the restricted form of the randomization procedures (18%). In regards to average item overlap for simulees with similar abilities in both the 100 (19%) and 200 (11%) item pool the progressive restricted procedure produced the lowest amount of item overlap. The nine item variants of the restricted randomesque and restricted modified within .10 logits procedures produced the lowest percent of average item overlap for simulee with different abilities with 26% in the 100 item pool and 21% in the 200 item pool. In terms of the three indicators of item overlap discussed the progressive restricted procedure emerged as the best performing procedure across item pools followed closely by the nine item variants of the restricted randomesque and restricted modified within .10 logits procedures.

In terms of test security the current research finds that the progressive restricted procedure and the nine item variants of the restricted randomesque and restricted modified within .10 logits procedure protected test security the best. No single procedure was however found to be superior on all indices of test security. When looking at the maximum exposure rate the restricted modified within .10 logits procedure with an item group size of nine performed the best. In terms of even item usage the three procedures performed very similarly. In terms of pool utilization the progressive restricted emerged clearly superior having utilized 100% of both item pools. In terms of average overall overlap the progressive restricted procedure had a slight edge in the larger item pool but performed equivalently as the nine item variants of the restricted randomesque and restricted modified within .10 logits procedures in the 100 item pool. The progressive restricted procedure also demonstrated the lowest average overlap rates for simulees with similar abilities in both item pools.

Conclusions

The present research explored the efficacy of five exposure control procedures using the generalized partial credit model (Muraki, 1992). The exposure control procedures examined were the randomesque, modified within .10 logits, restricted randomesque, restricted modified within .10 logits, and progressive restricted procedures. The randomesque, modified within .10 logits procedures, restricted randomesque and restricted modified within .10 logits procedures were each implemented with an item group size of three, six, and nine. The maximum information item selection procedure was included as a no exposure control baseline comparison condition. The restricted randomesque and restricted modified within .10 logits procedures are introduced as procedures that incorporate the demonstrated advantages of randomization procedures in polytomous item pools while curbing the negative effects these procedures tend to have on maximum exposure rate. Each condition was evaluated from its performance on a number of indices of measurement precision and test security.

As expected the maximum information condition resulted in the best measurement precision indices and the poorest test security. The restricted randomesque and restricted modified within .10 logits procedures demonstrated significant improvements over the randomesque and modified within .10 logits procedures in terms of test security. Given that they come at the cost of only minor loss in measurement precision the present research strongly supports these procedures over their base procedures.

That said, the present research found that the progressive restricted procedure and the nine item variant of these procedures produced the most desirable levels of test security on the indices examined. The progressive restricted procedure tended to produce slightly better overlap rates particularly in the 200 item pool. It also used the item pool slightly more evenly in the larger 200 item pool. Most significantly the progressive restricted procedure resulted in complete use of the item pool. Both of the nine item variants of the restricted randomesque and restricted modified within .10 logits procedures tended to produce lower maximum exposure rates than the progressive restricted procedure. While the nine item variants of the restricted randomesque and restricted modified within .10 logits procedures tended to produce equal pool use and overlap rates the nine item variant of the restricted modified within .10 logits procedure tended to produce slightly lower maximum exposure rates. All three procedures were successful at preventing any item in the item pool from exceeding an exposure rate of 0.35 which has implications for the longevity of use of items in the item pool.

As a cautionary note, the differences seen between these three procedures were small and may have been idiosyncratic. While the use of ten replications of each condition lends support to the reality of these differences practically speaking these three procedures performed equivalently. The present research study adds to the body of research studies which supports the use of randomization procedures over conditional procedures for polytomous item pools. In addition to supporting the use of randomization procedures the present research finds strong support for the use of the combinatorial procedures investigated over randomization procedures. The present research finds strong support for the use of these combinatorial procedures with larger item pools.

Limitations and Suggestions for Future Research

The present dissertation deliberately focused on the randomization and combinatorial procedures because of the success that has been demonstrated in previous research of randomization procedures over more complicated conditional procedures. This however necessarily limits the generalizations that are possible as the conditional procedures were not examined head to head with the randomization procedures and combinatorial procedures. This holds particularly true for the nine item variant of the randomesque and modified within .10 logits procedures which was investigated here as well as for the restricted randomesque and restricted modified within .10 logits procedure extensions which were introduced in the present dissertation. Future research should endeavor not only to replicate the successes of the current study but to conduct head to head comparisons with the more complex conditional procedures. Of particular interest may be the ability of the nine item variants of the randomization procedures and the combinatorial procedures to produce maximum exposure rates comparable to those traditionally seen by conditional procedures with polytomous item pools.

The present research investigated the functioning of the exposure control procedure with a small item pool of 100 and a moderate 200 item pool. As larger item pools provided better measurement precision and, in general, better test security future research should investigate the combinatorial procedures with even larger item pools. This should be done with special attention given to the maximum exposure rate produced by the restricted randomesque and restricted modified within .10 logits procedures. Given that the present research found that increasing the item pool resulted in these procedures generating slightly higher maximum exposure rates future research should investigate if this pattern holds true. If this increase is seen with larger item pools it needs to be determined if the increases are substantial or marginal. The increases in test security observed in the larger item pool seemed to have begun showing signs that further increasing the item pool may not result in significantly more improvement of test security. Future research should endeavor to examine the gains seen when the item pool is increase beyond 200 items.

As in previous research, the impact of item group size proved to be substantial. In every case, the nine item variants of the randomesque, modified within .10 logits procedures, restricted randomesque, and restricted modified within .10 logits procedures produced more advantageous results than their three or six item variant counterparts. It may prove productive to investigate if increasing the item group size further will continue this trend or if, perhaps, an item group size of nine is optimal for these procedures as the expansion may introduce too many suboptimal items into the estimation procedure.

This dissertation is also necessarily limited by its scope of IRT model. Future research should endeavor to explore if the patterns and trends found in the current study are found with other models. As the trend regarding the effects of an increase in item group size has been demonstrated with other models the incorporation of a group size of nine is justified. As well the success of the restricted randomesque and restricted modified within .10 logits procedures over their unrestricted forms needs to be verified for other models.

In addition to the replication and verification of the superiority of the restricted procedures introduced in this dissertation the performance of these procedures in relation to the progressive restricted procedure need to be investigated further. In the present research these three procedures were found to produce the most desirable results in regards to test security while maintaining good measurement precision. Future research should attempt to tease out the differences in the performance of these procedures.

Appendix A: Conditional Bias Plots

Figure A1: Conditional Bias Plot for Maximum Information (No Exposure Control) with the 100 Item Pool

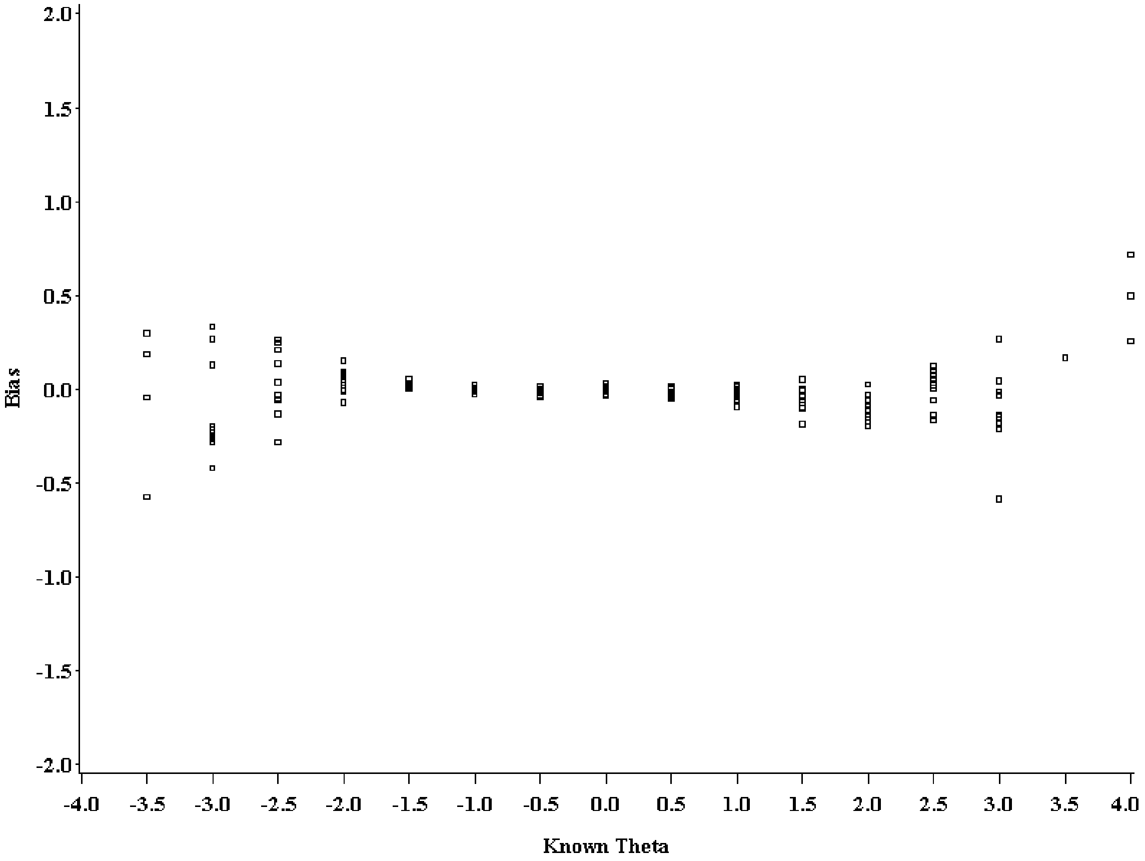


Figure A2: Conditional Bias Plot for Progressive Restricted with the 100 Item Pool

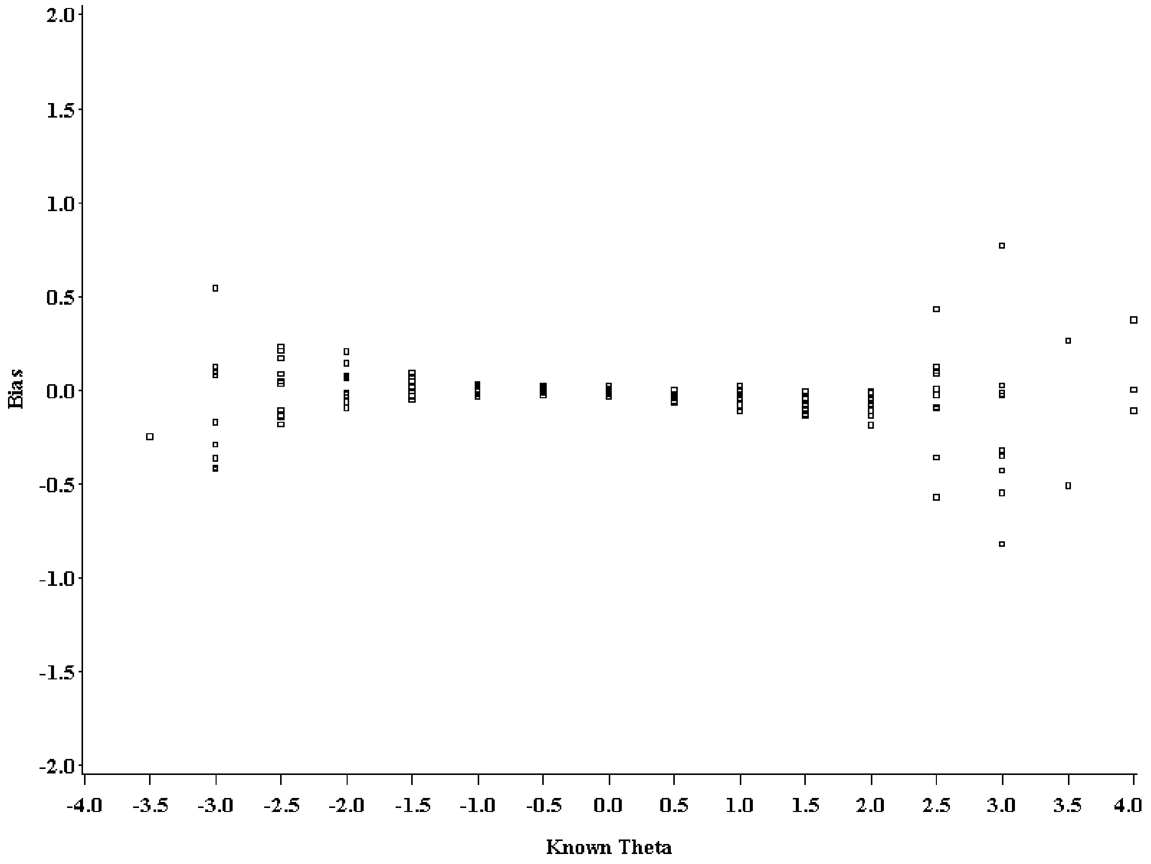


Figure A3: Conditional Bias Plot for Randomesque (Item Group Size = 3) with the 100 Item Pool

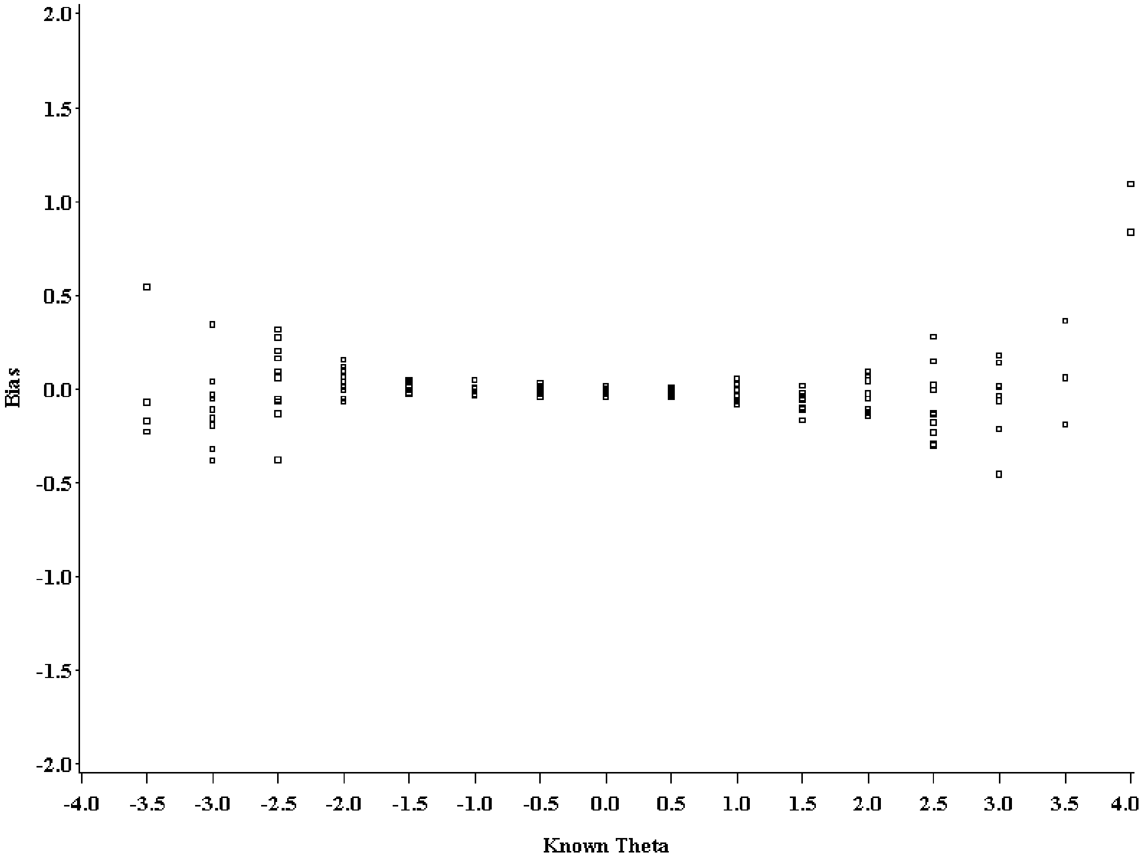


Figure A4: Conditional Bias Plot for Randomesque (Item Group Size = 6) with the 100 Item Pool

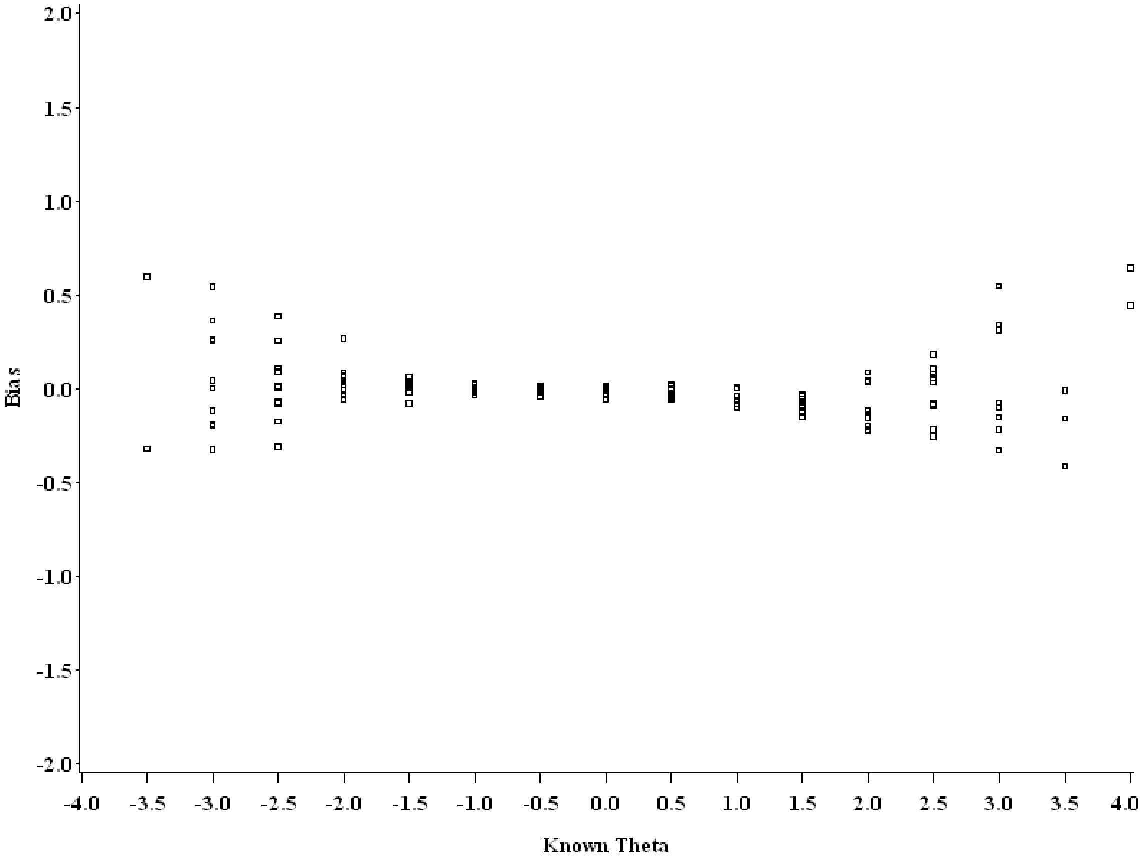


Figure A5: Conditional Bias Plot for Randomesque (Item Group Size = 9) with the 100 Item Pool

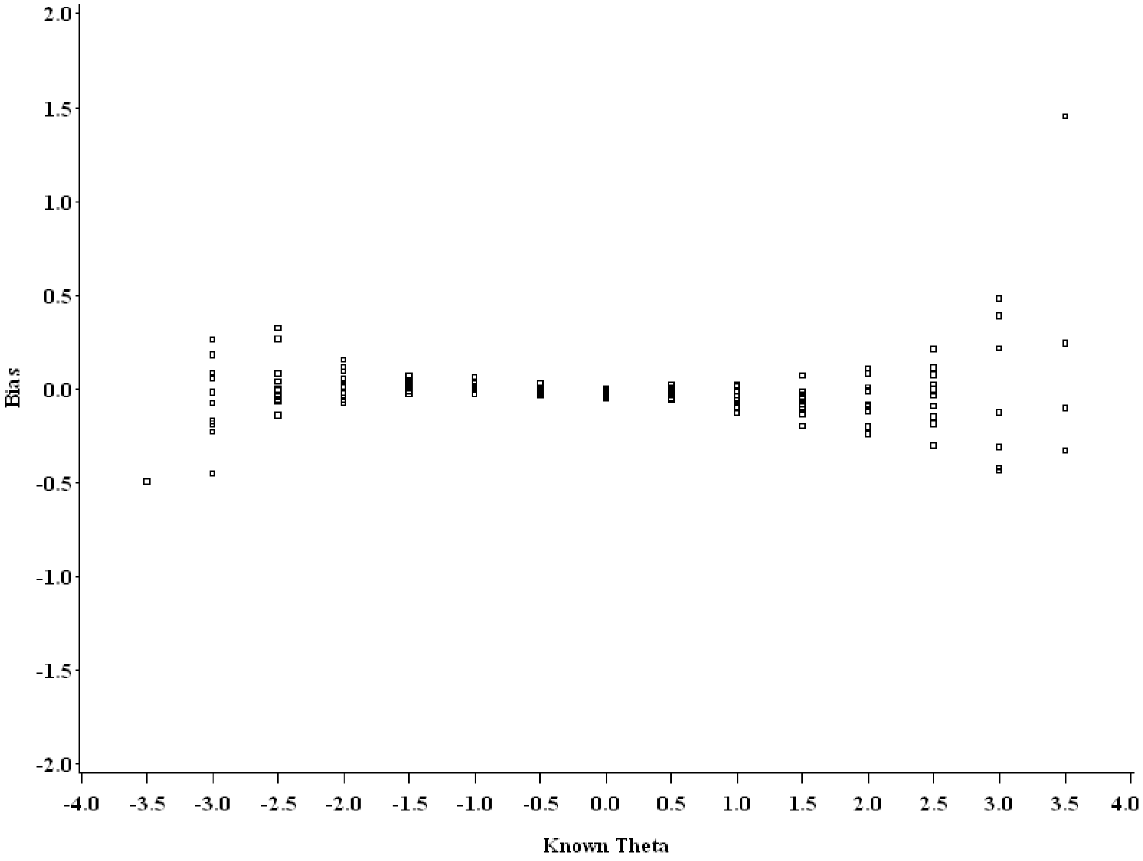


Figure A6: Conditional Bias Plot for Restricted Randomesque (Item Group Size = 3) with the 100 Item Pool

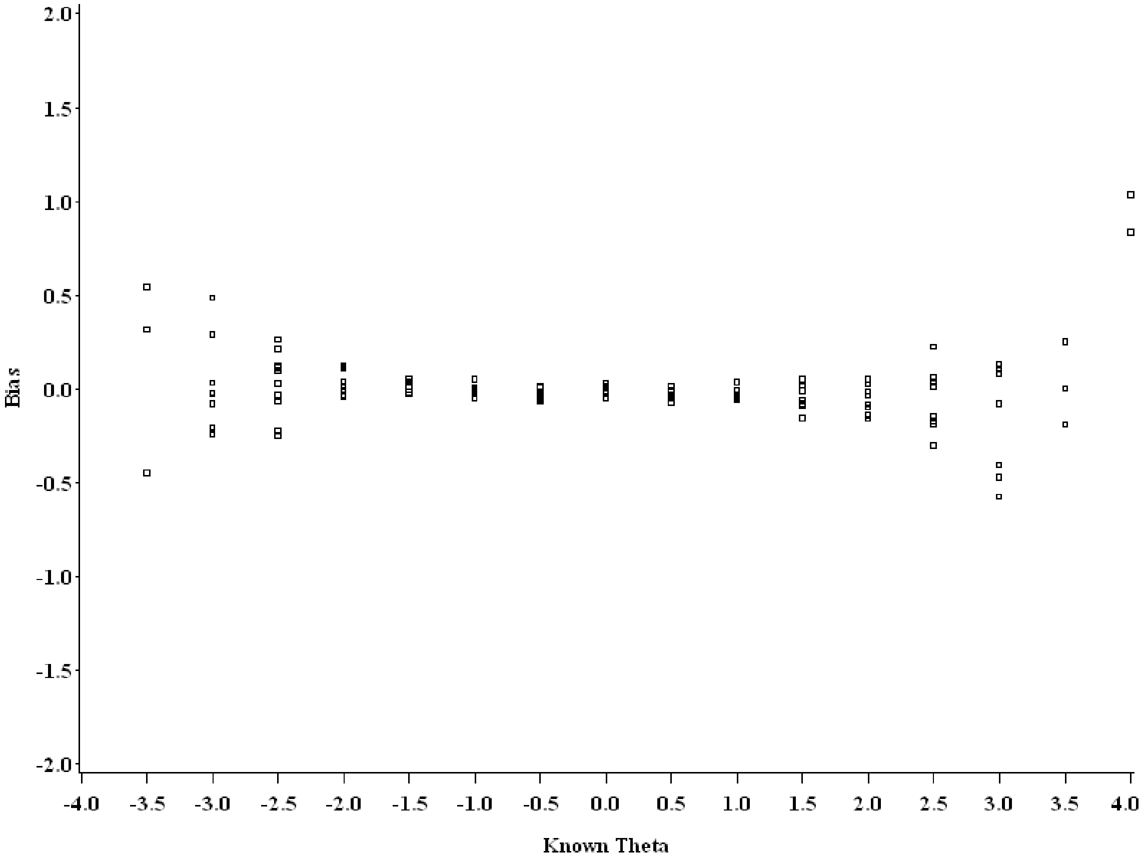


Figure A7: Conditional Bias Plot for Restricted Randomesque (Item Group Size = 6) with the 100 Item Pool

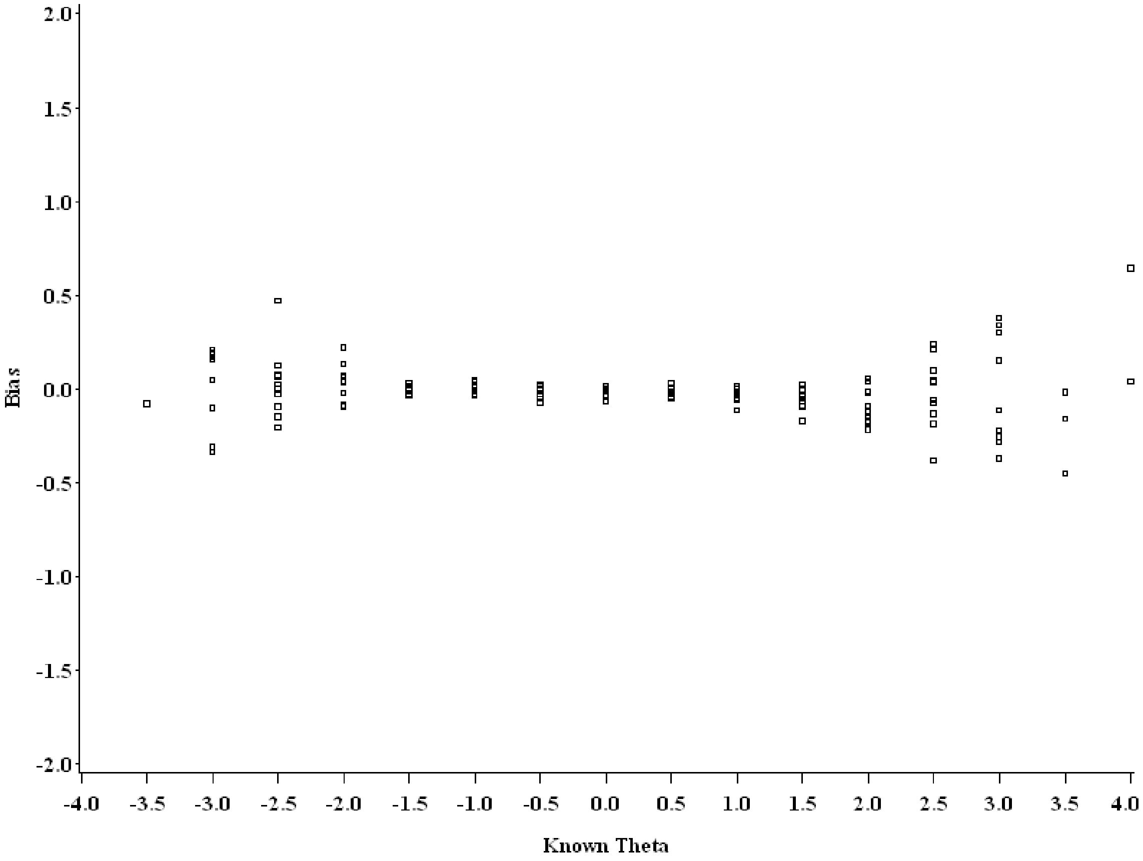


Figure A8: Conditional Bias Plot for Restricted Randomesque (Item Group Size = 9) with the 100 Item Pool

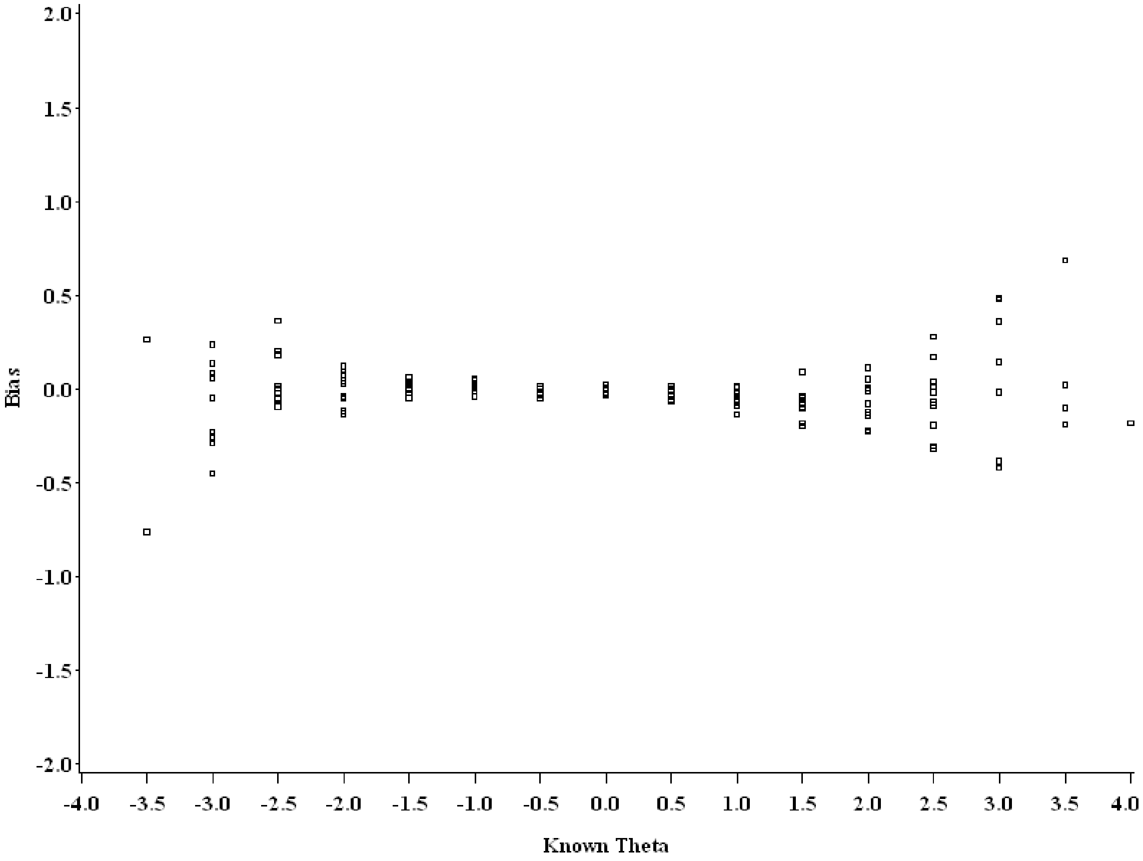


Figure A9: Conditional Bias Plot for Modified Within 0.10 logits (Item Group Size = 3)

with the 100 Item Pool

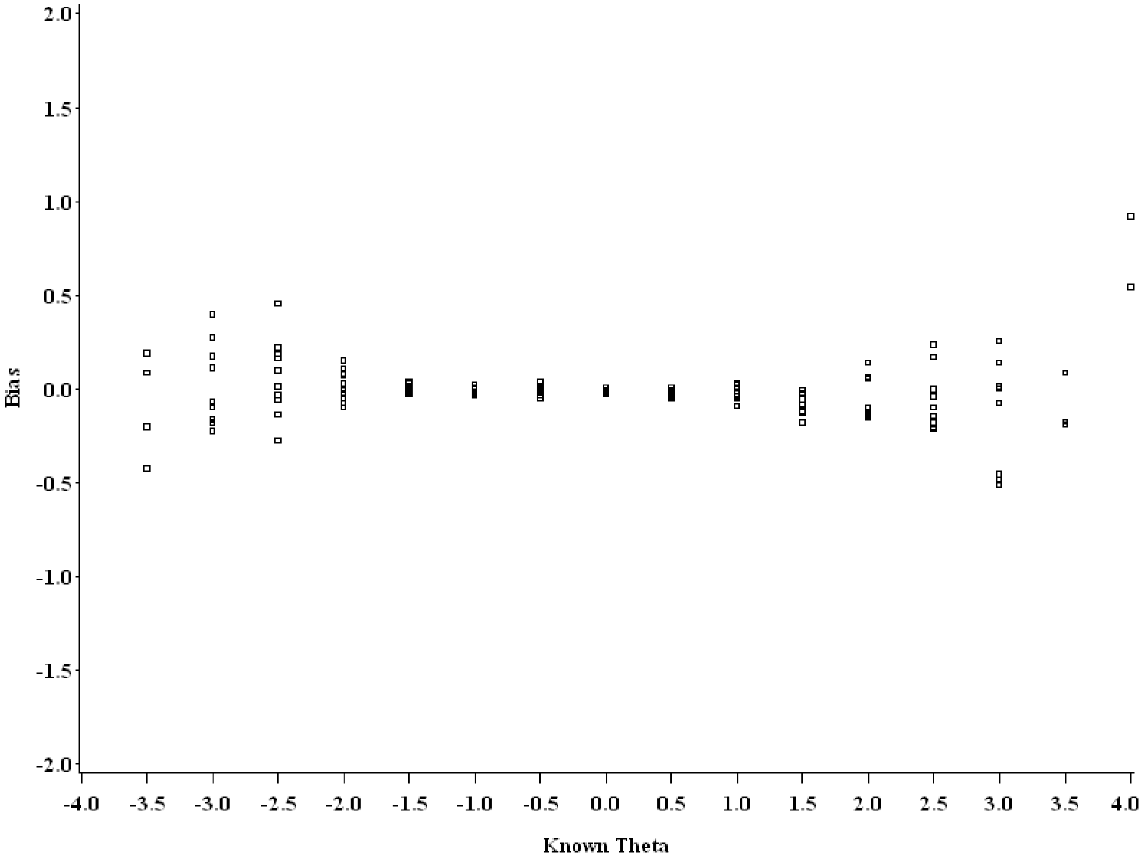


Figure A10: Conditional Bias Plot for Modified Within 0.10 logits (Item Group Size = 6) with the 100 Item Pool

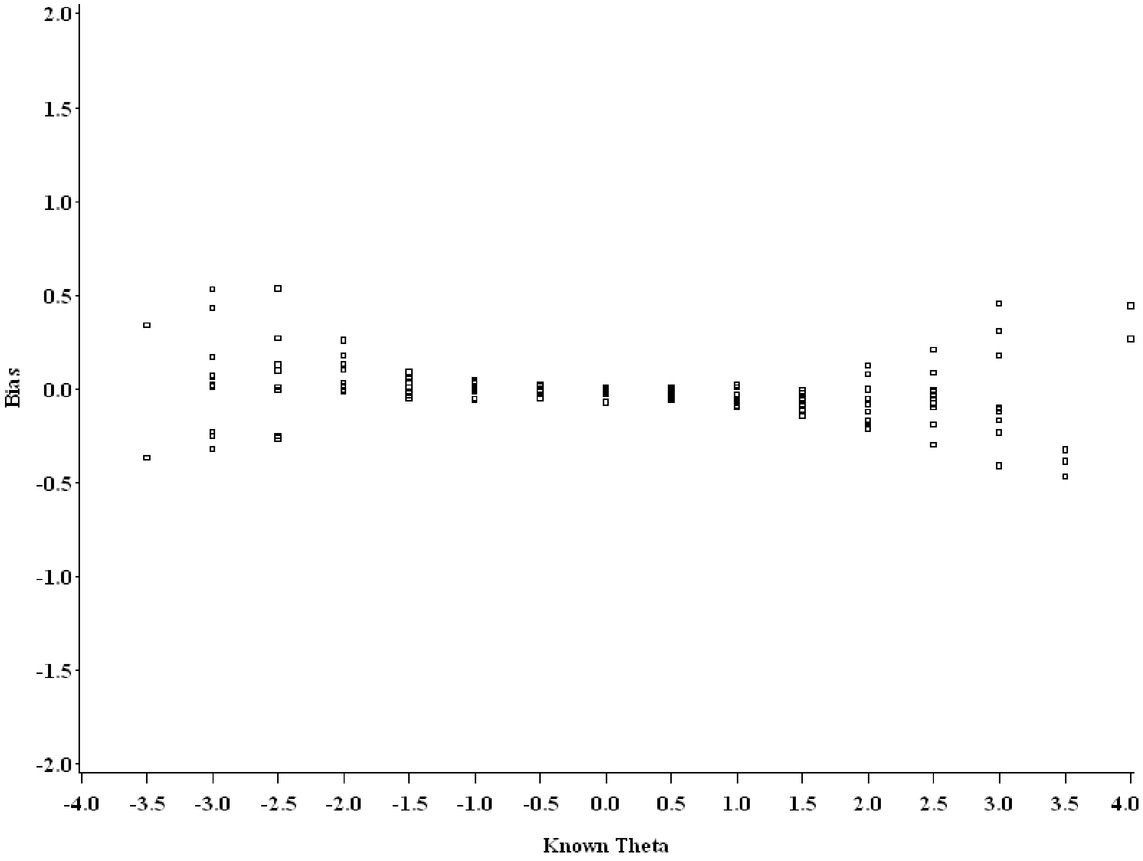


Figure A11: Conditional Bias Plot for Modified Within 0.10 logits (Item Group Size = 9) with the 100 Item Pool

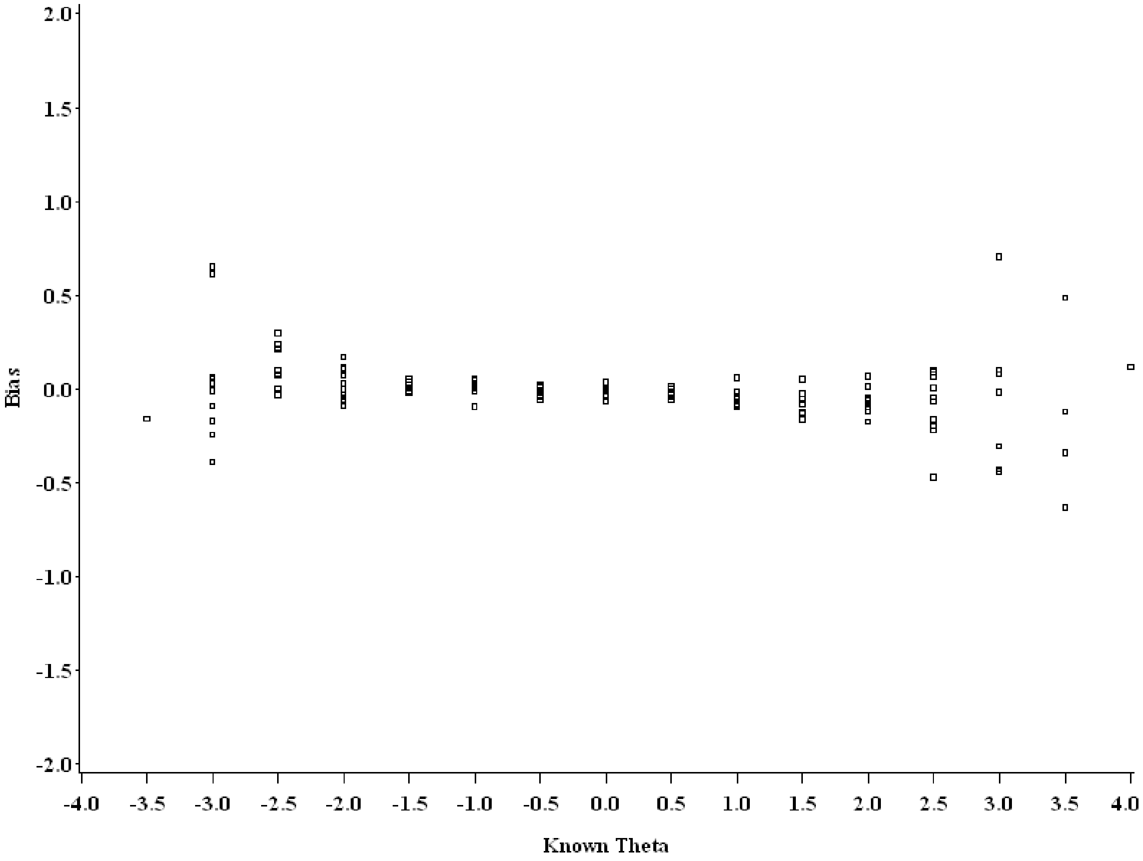


Figure A12: Conditional Bias Plot for Restricted Modified Within 0.10 logits

(Item Group Size = 3) with the 100 Item Pool

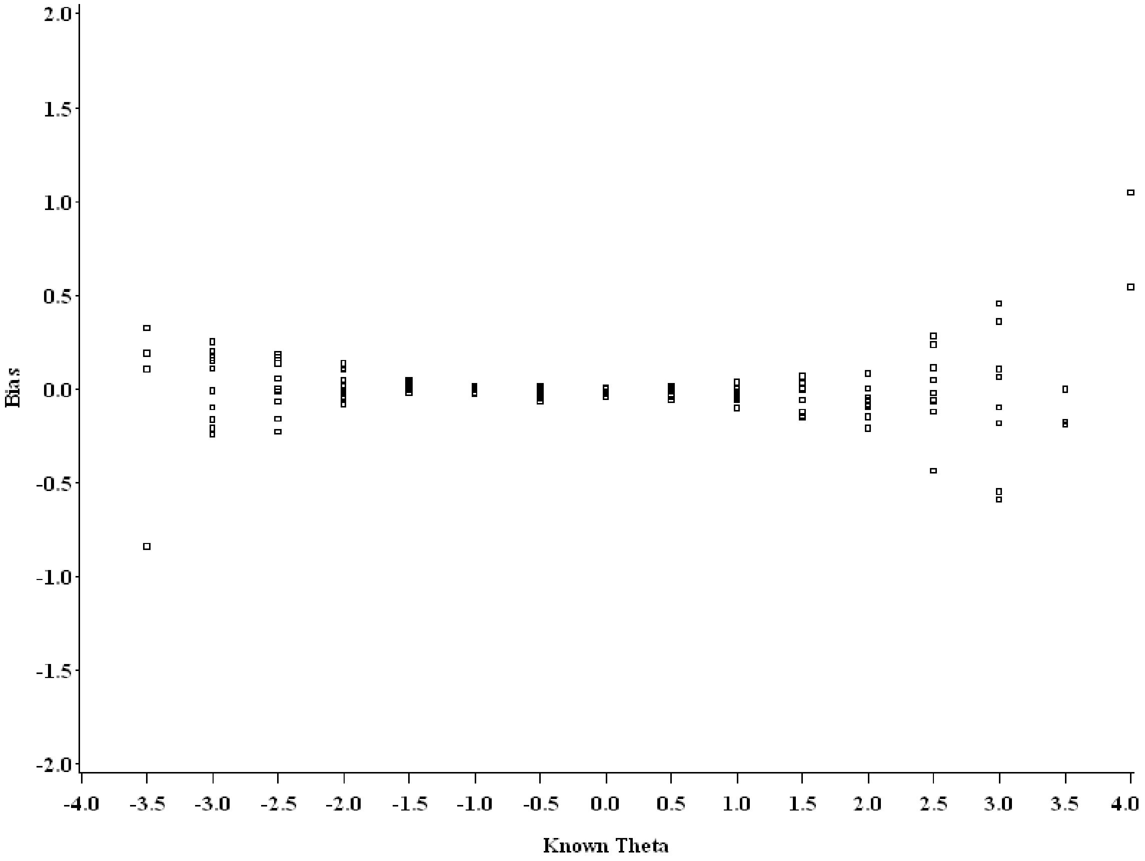


Figure A13: Conditional Bias Plot for Restricted Modified Within 0.10 logits

(Item Group Size = 6) with the 100 Item Pool

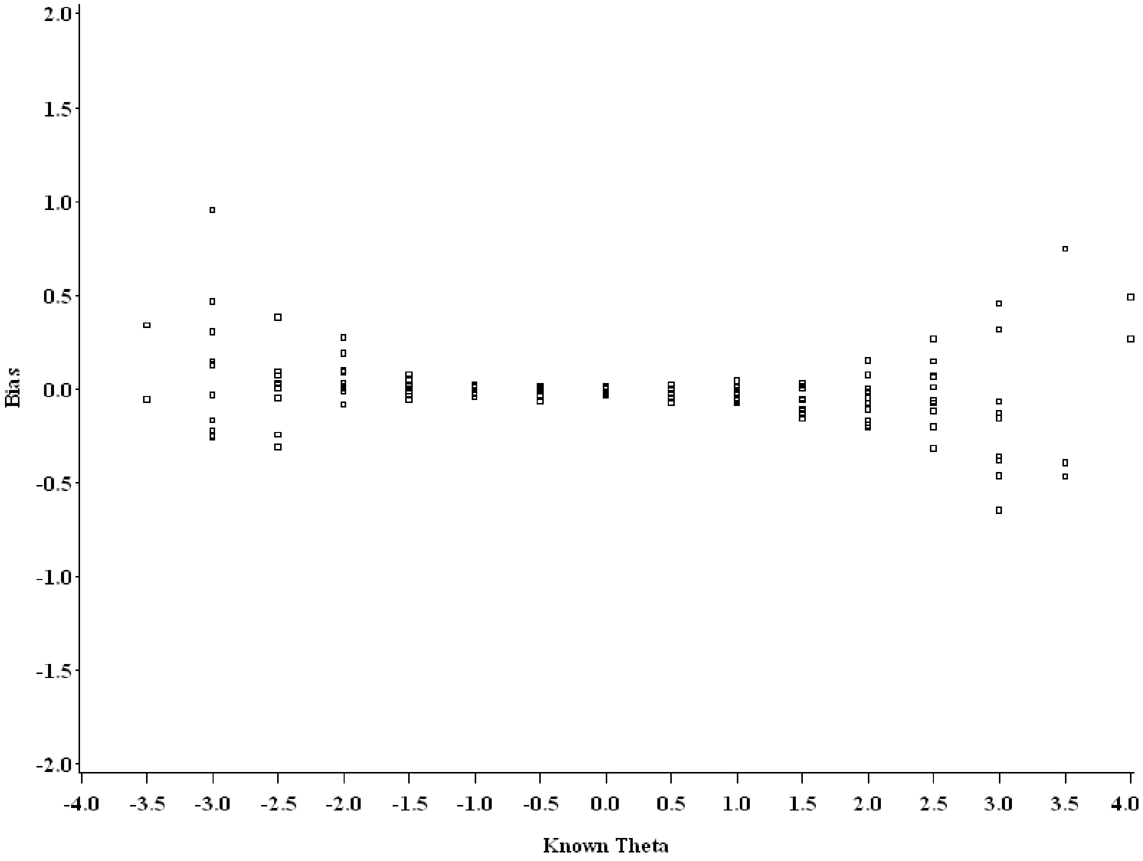


Figure A14: Conditional Bias Plot for Restricted Modified Within 0.10 logits

(Item Group Size = 9) with the 100 Item Pool

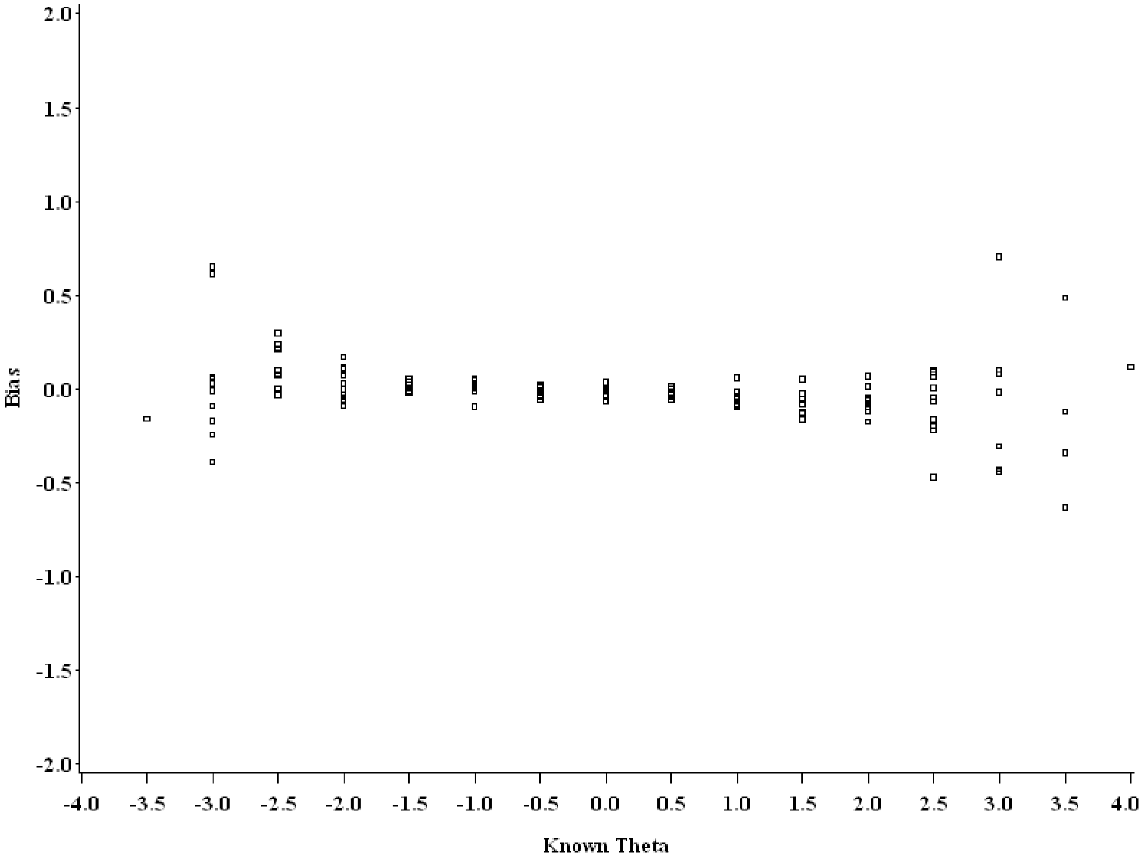


Figure A15: Conditional Bias Plot for Maximum Information (No Exposure Control)

with the 200 Item Pool

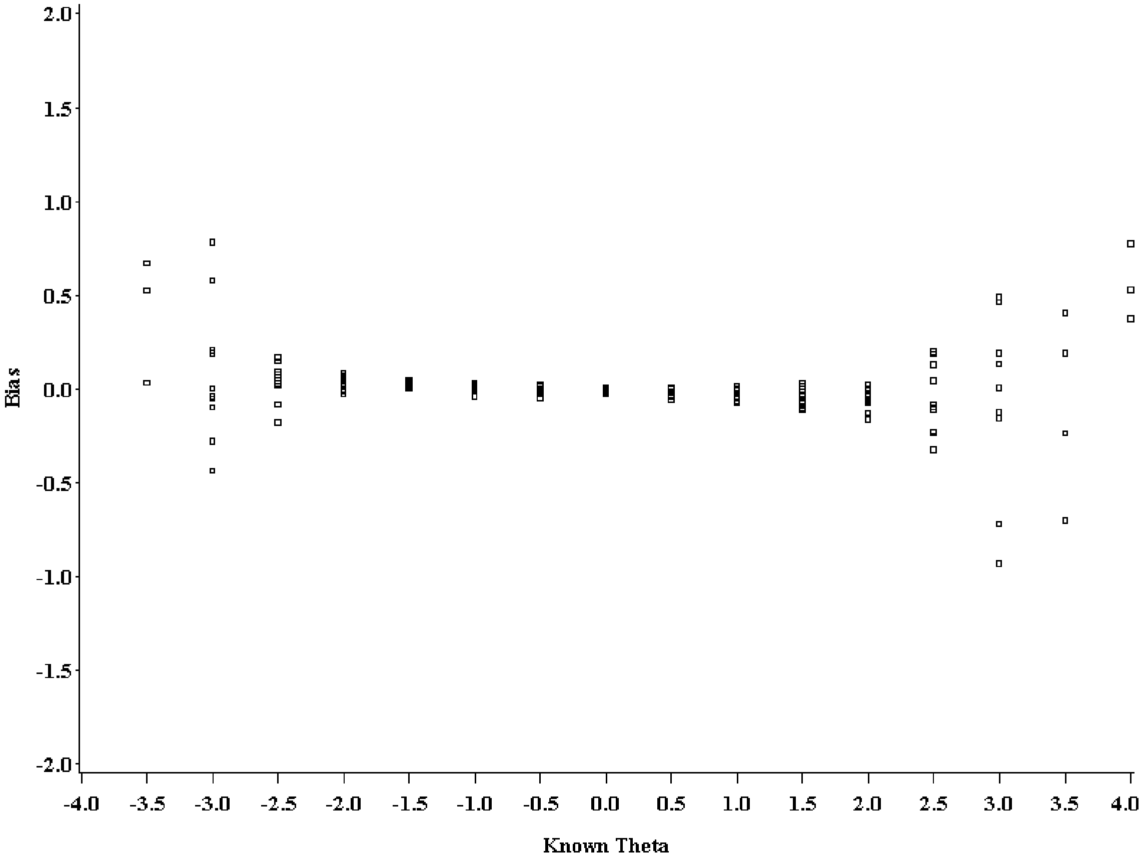


Figure A16: Conditional Bias Plot for Progressive Restricted with the 200 Item Pool

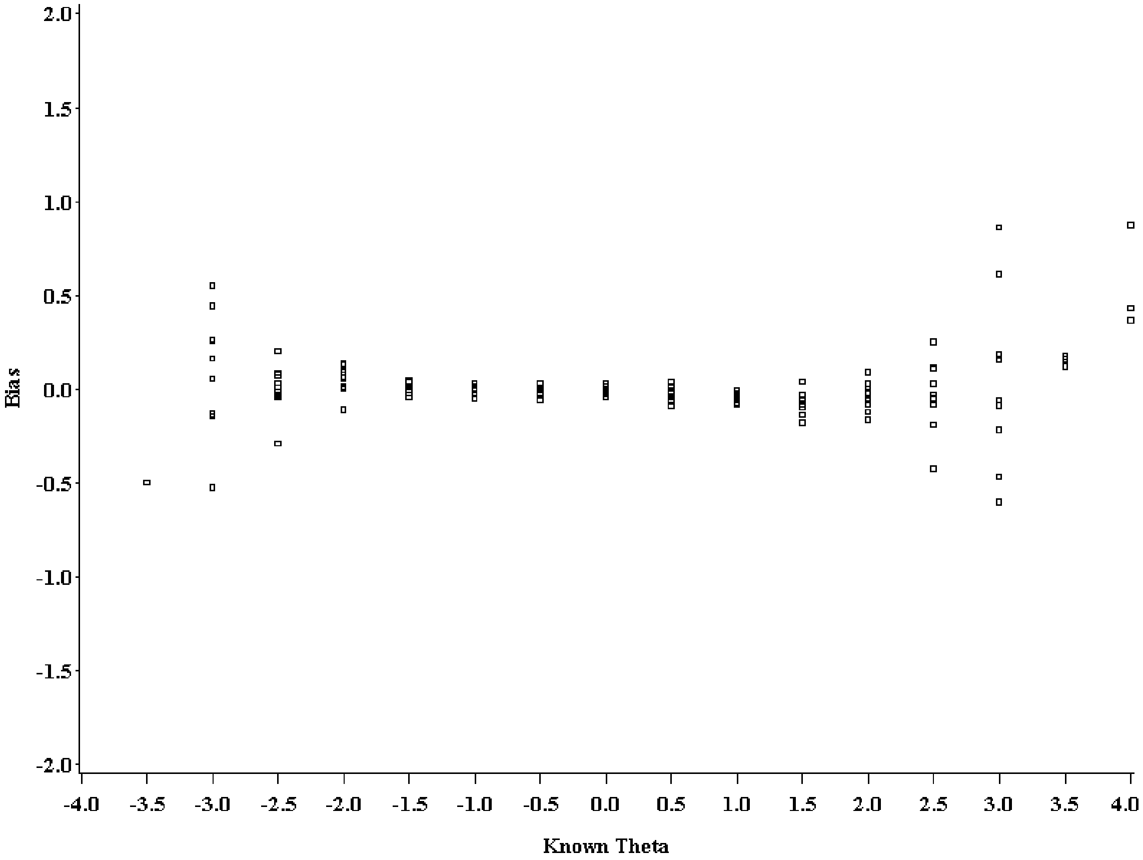


Figure A17: Conditional Bias Plot for Randomesque (Item Group Size = 3) with the 200 Item Pool

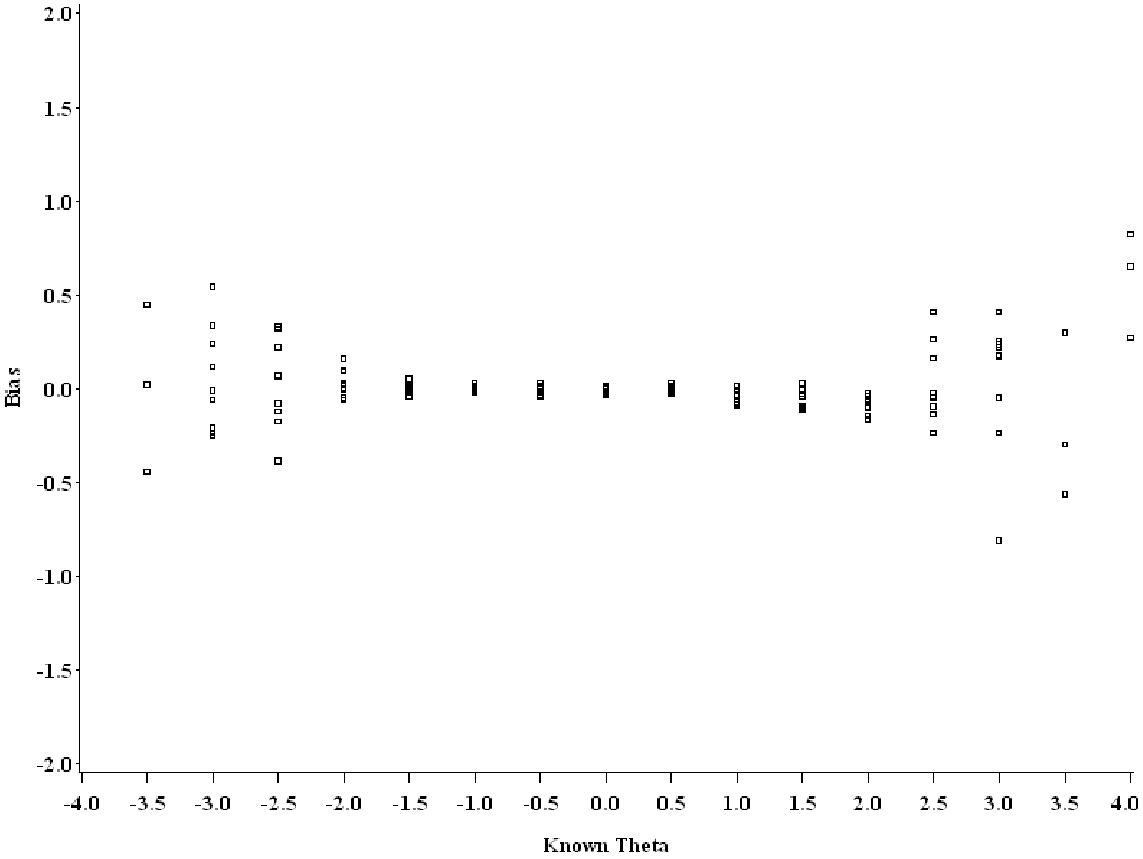


Figure A18: Conditional Bias Plot for Randomesque (Item Group Size = 6) with the 200 Item Pool

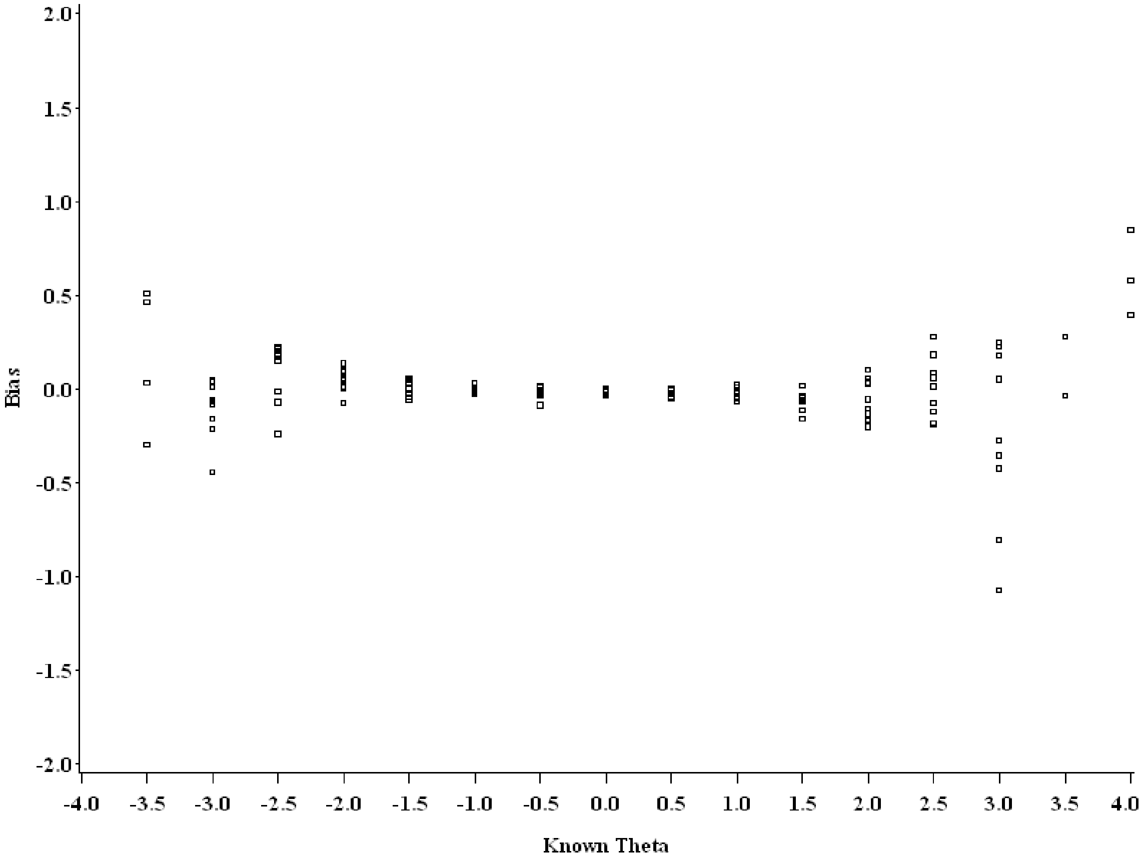


Figure A19: Conditional Bias Plot for Randomesque (Item Group Size = 9) with the 200 Item Pool

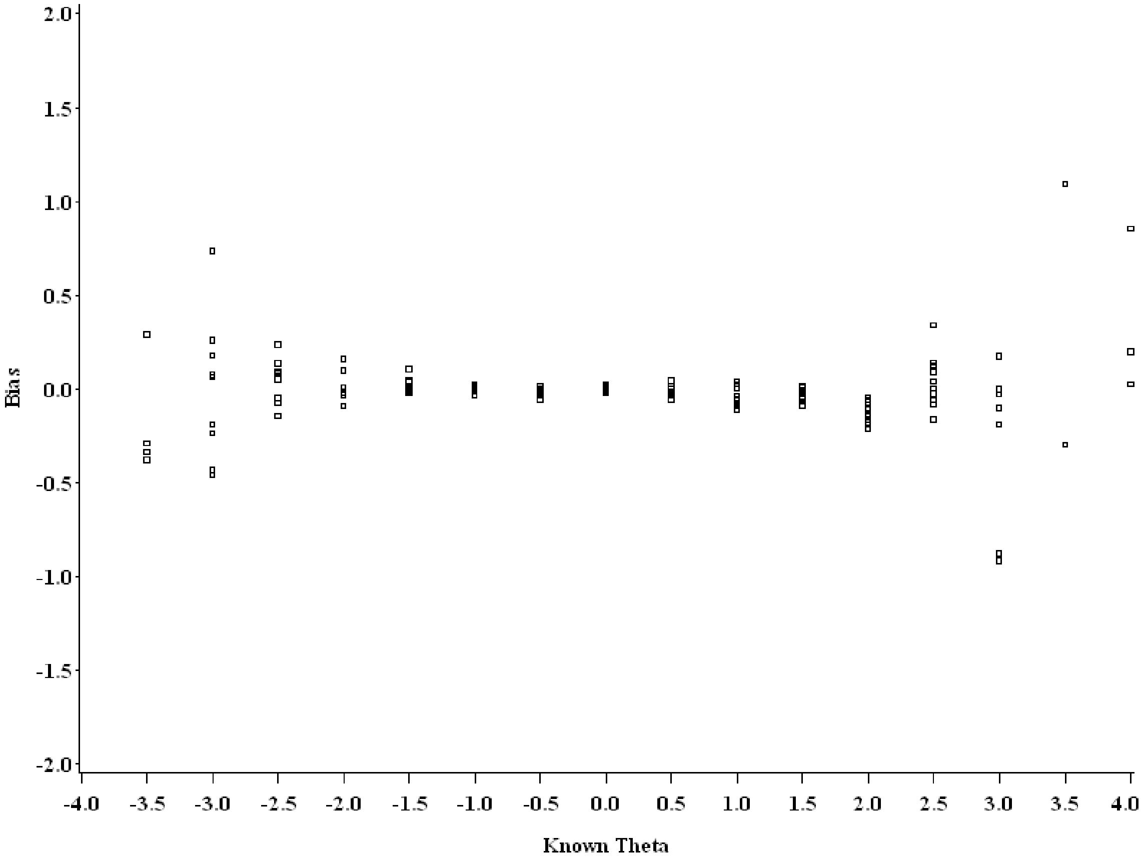


Figure A20: Conditional Bias Plot for Restricted Randomesque (Item Group Size = 3) with the 200 Item Pool

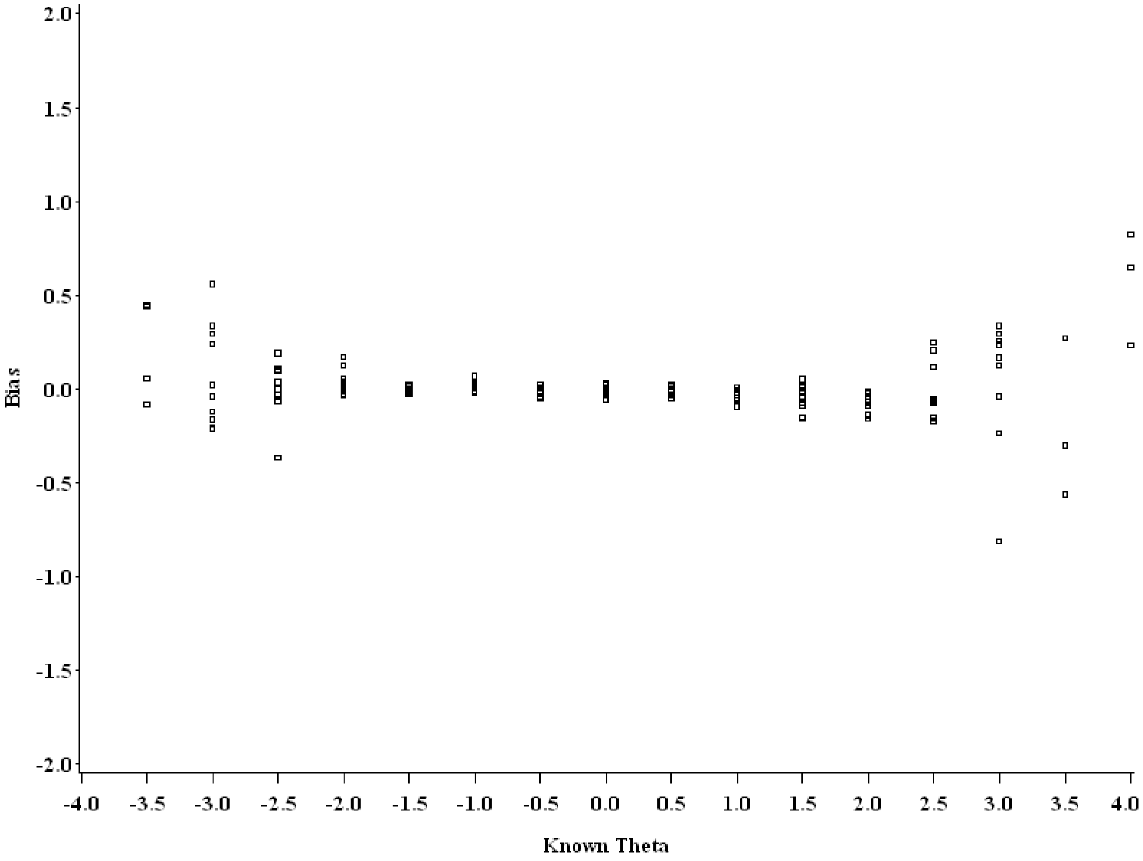


Figure A21: Conditional Bias Plot for Restricted Randomesque (Item Group Size = 6) with the 200 Item Pool

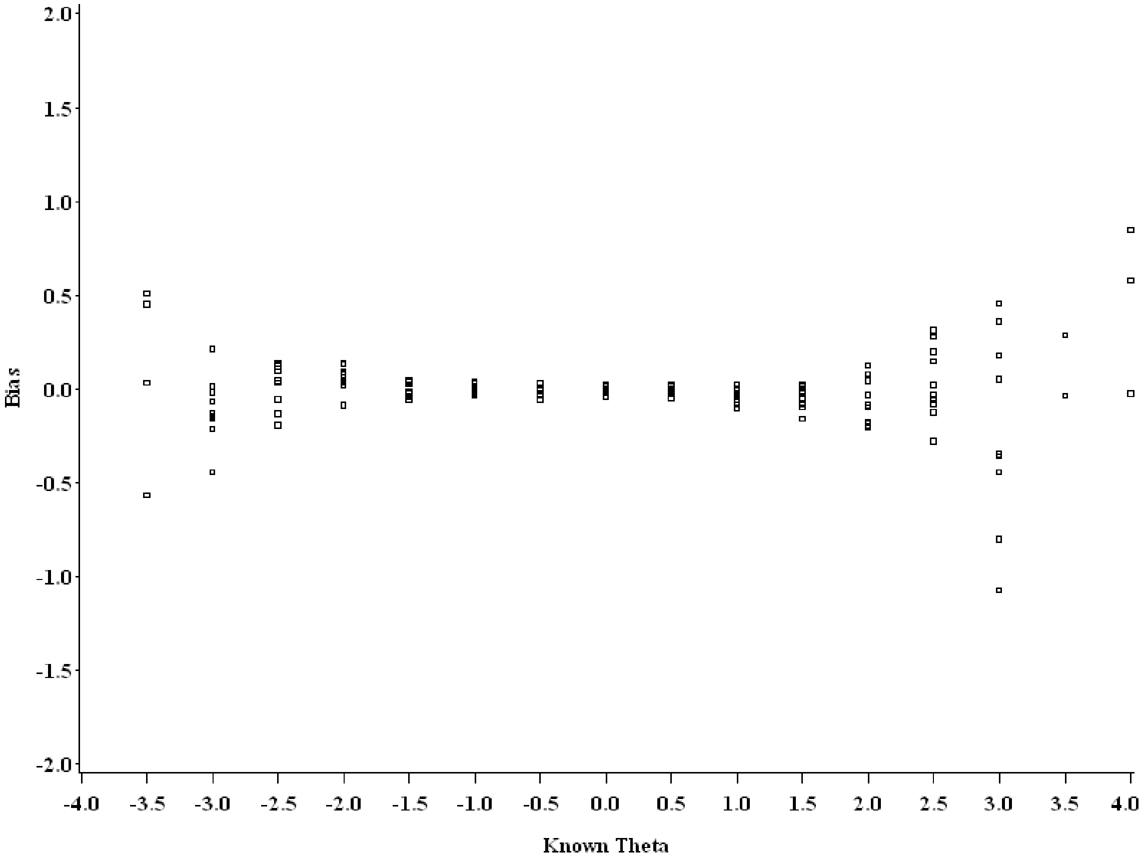


Figure A22: Conditional Bias Plot for Restricted Randomesque (Item Group Size = 9) with the 200 Item Pool

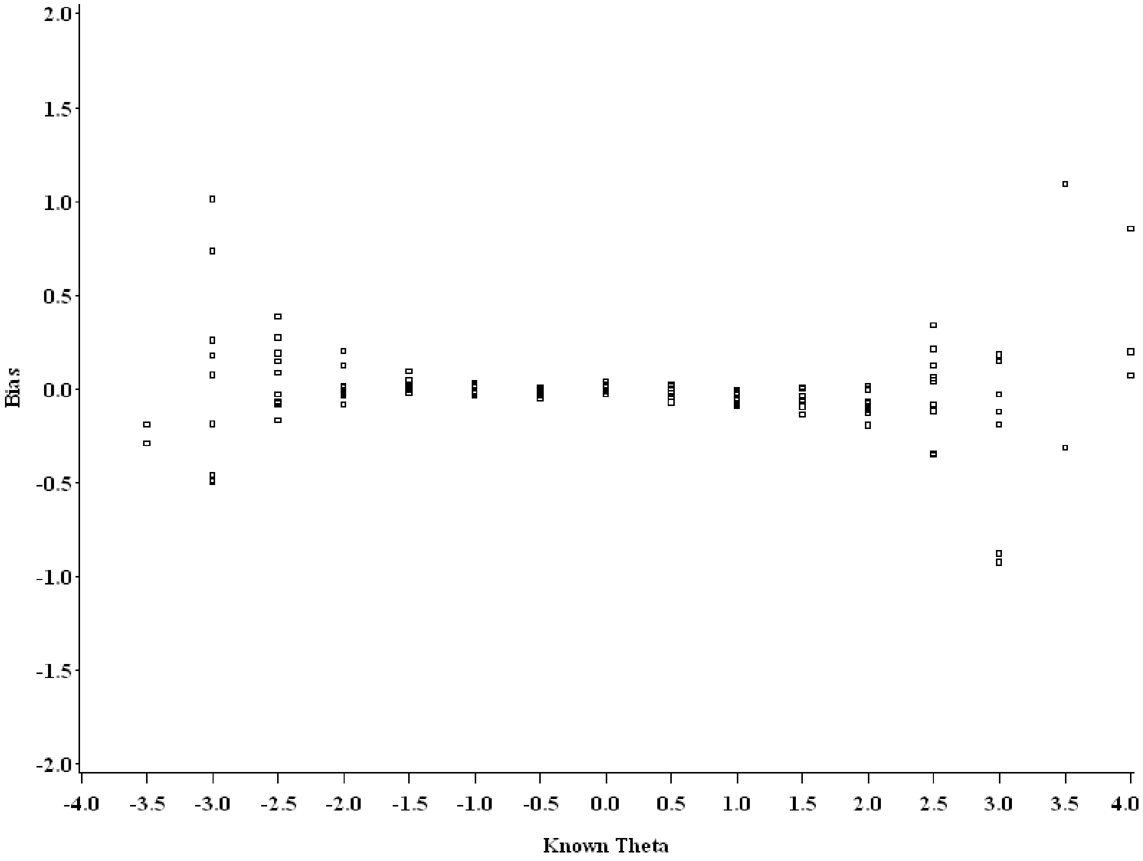


Figure A23: Conditional Bias Plot for Modified Within 0.10 logits (Item Group Size = 3) with the 200 Item Pool

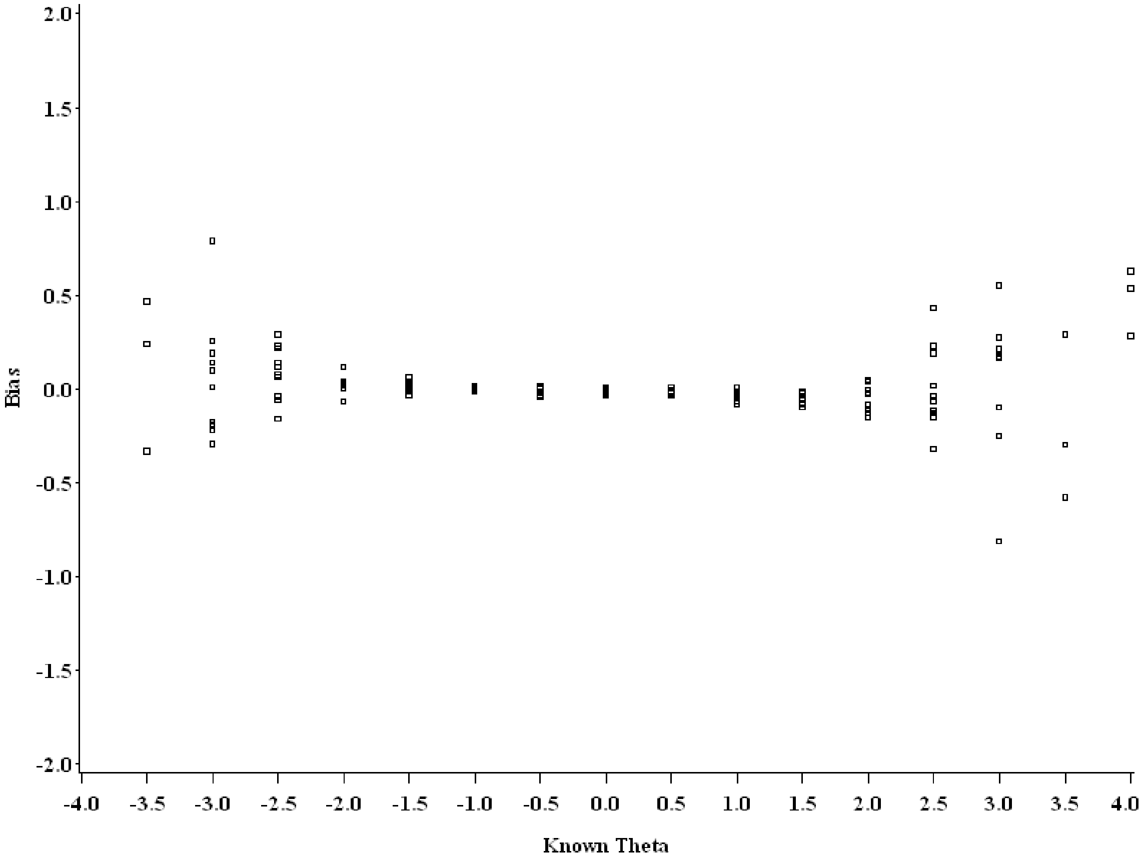


Figure A24: Conditional Bias Plot for Modified Within 0.10 logits (Item Group Size = 6)

with the 200 Item Pool

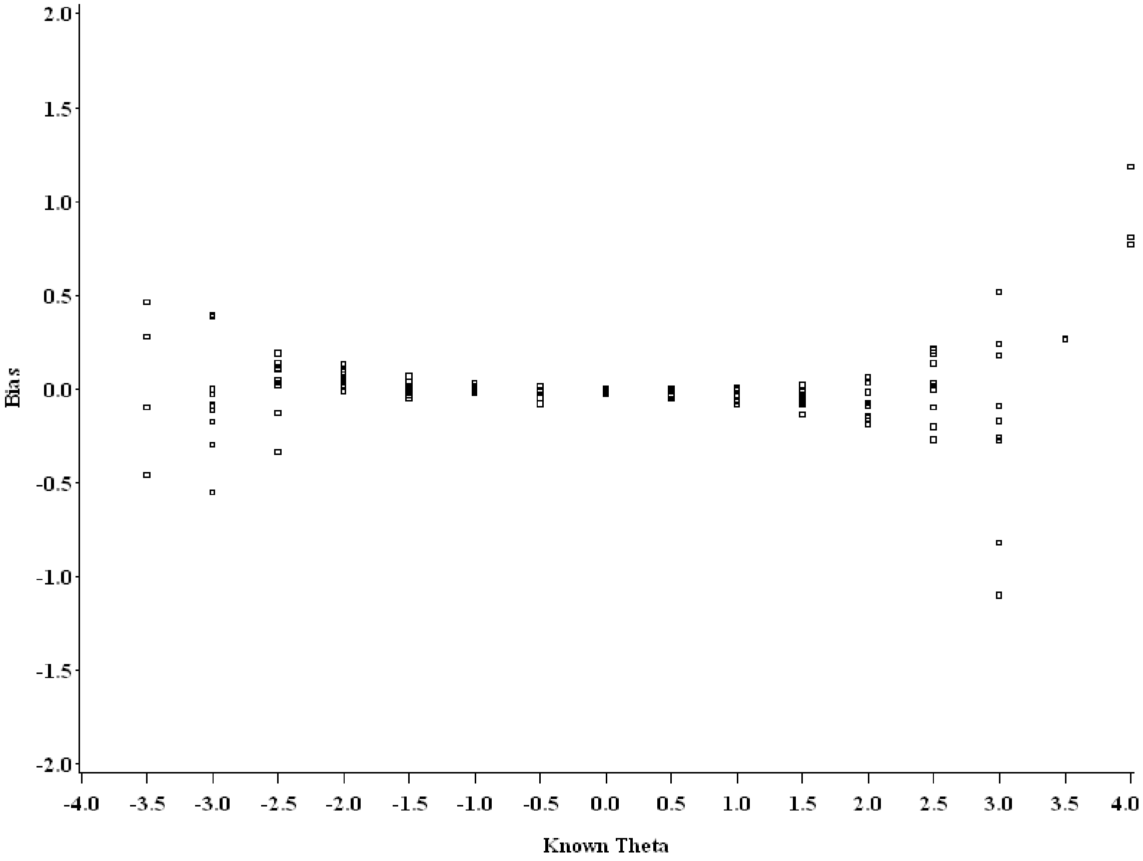


Figure A25: Conditional Bias Plot for Modified Within 0.10 logits (Item Group Size = 9)

with the 200 Item Pool

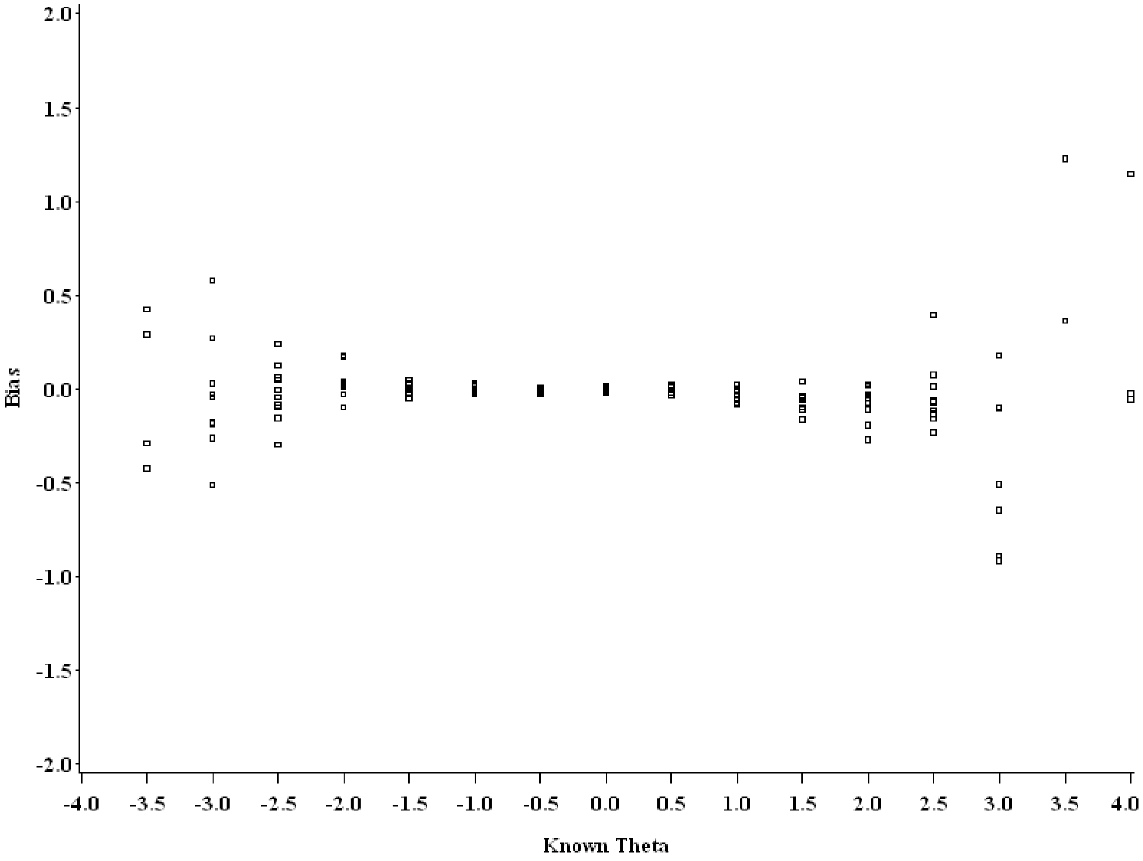


Figure A26: Conditional Bias Plot for Restricted Modified Within 0.10 logits

(Item Group Size = 3) with the 200 Item Pool

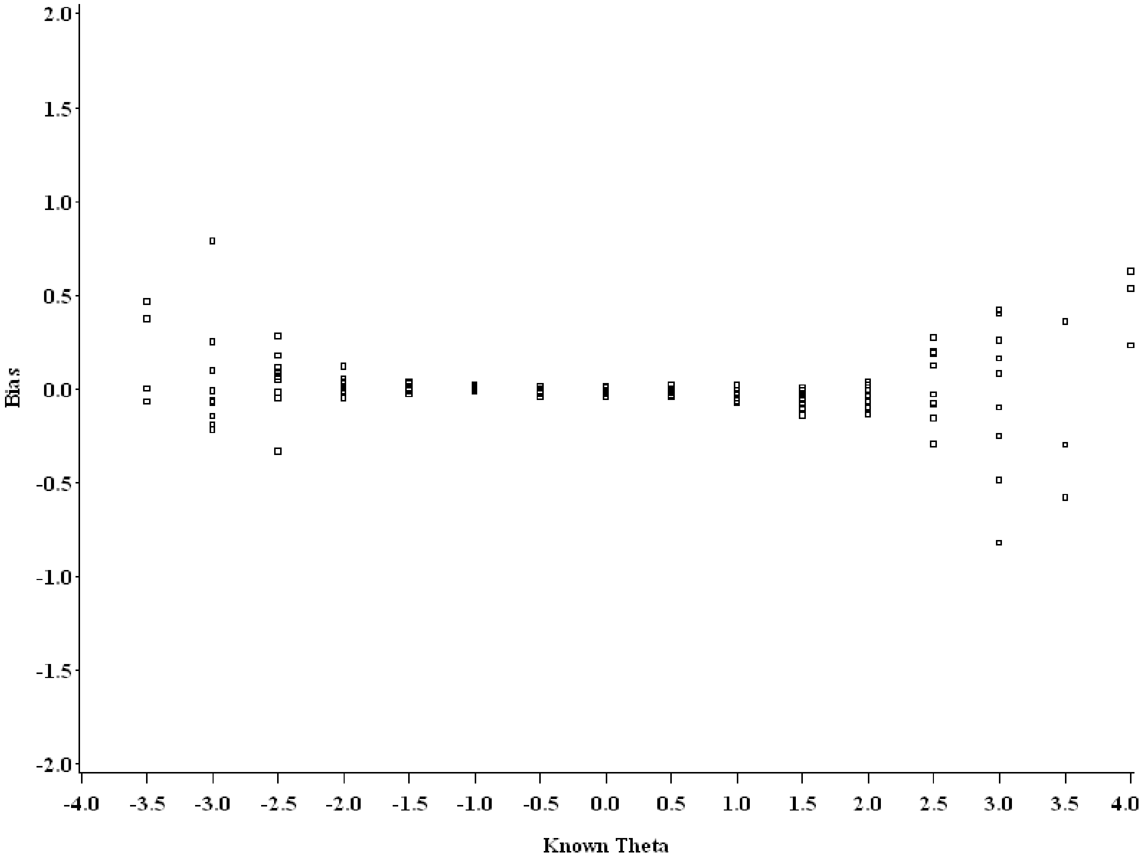


Figure A27: Conditional Bias Plot for Restricted Modified Within 0.10 logits

(Item Group Size = 6) with the 200 Item Pool

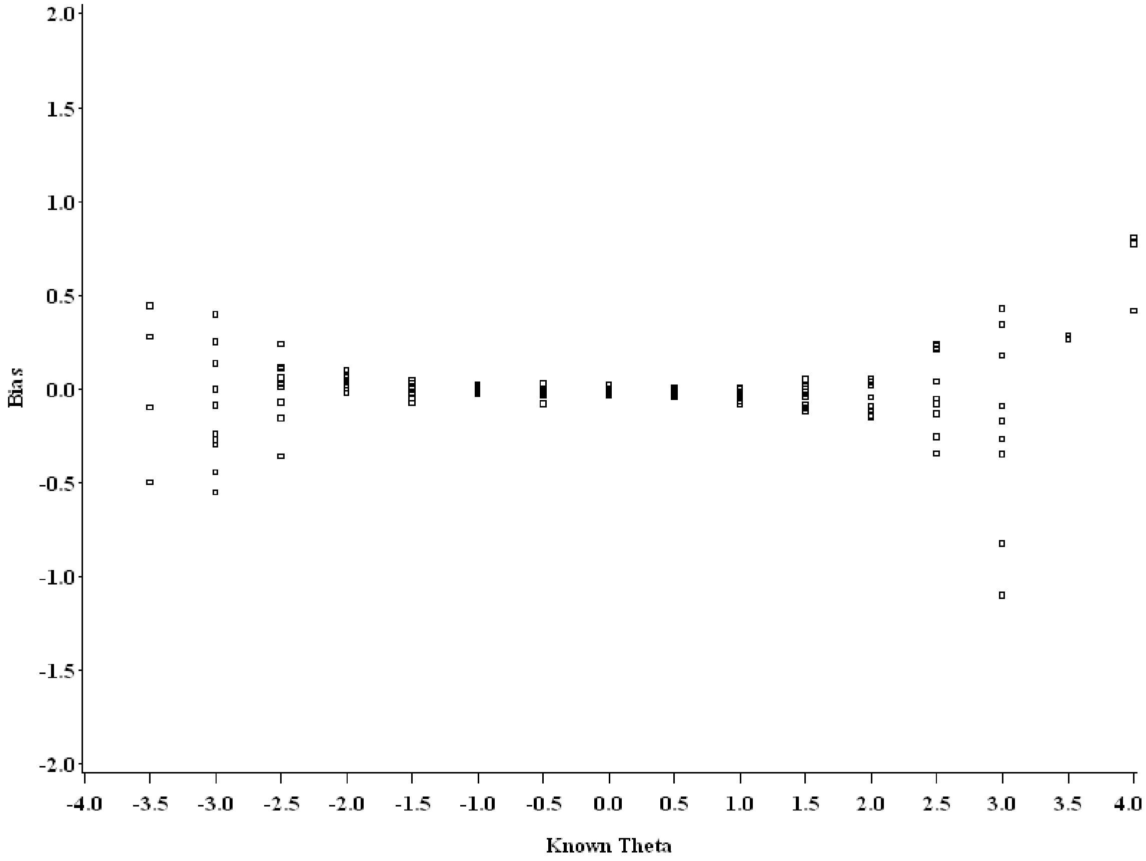
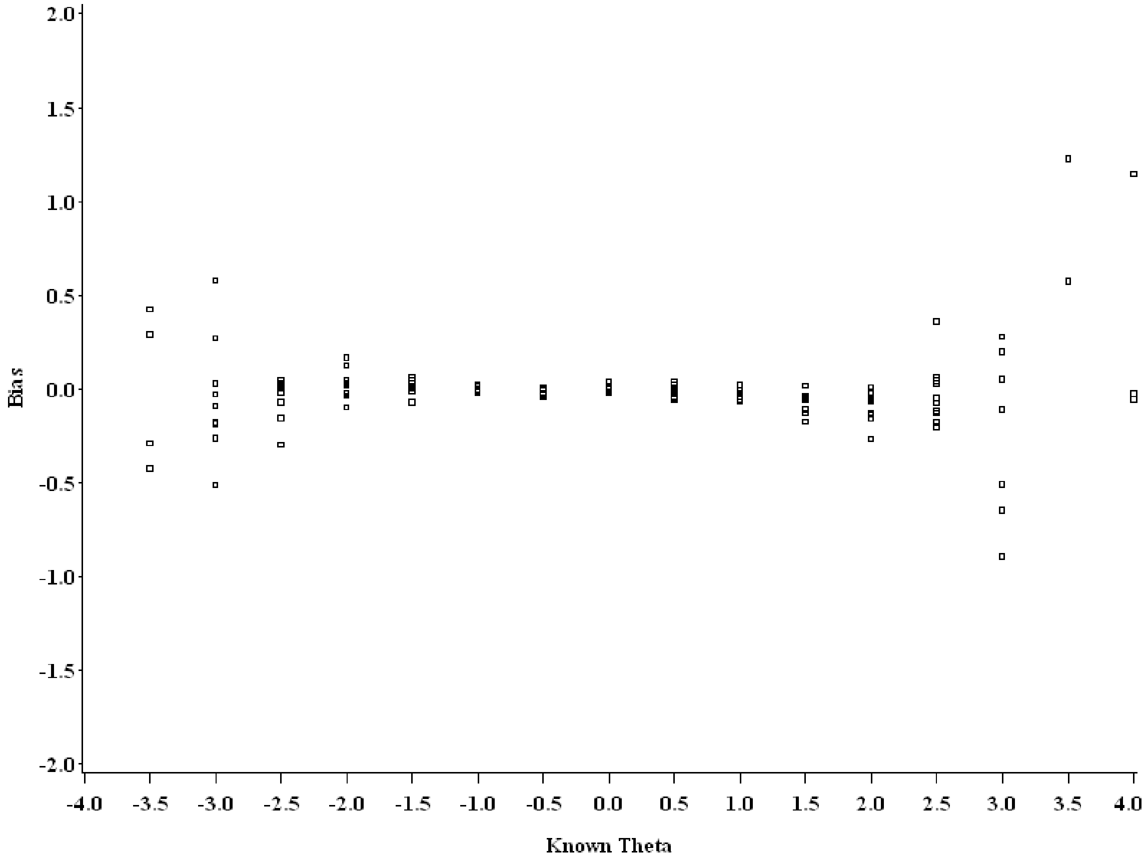


Figure A28: Conditional Bias Plot for Restricted Modified Within 0.10 logits

(Item Group Size = 9) with the 200 Item Pool



Appendix B: Conditional Standard Error Plots

Figure B1: Conditional Standard Error Plot for Maximum Information (No Exposure Control) with the 100 Item Pool

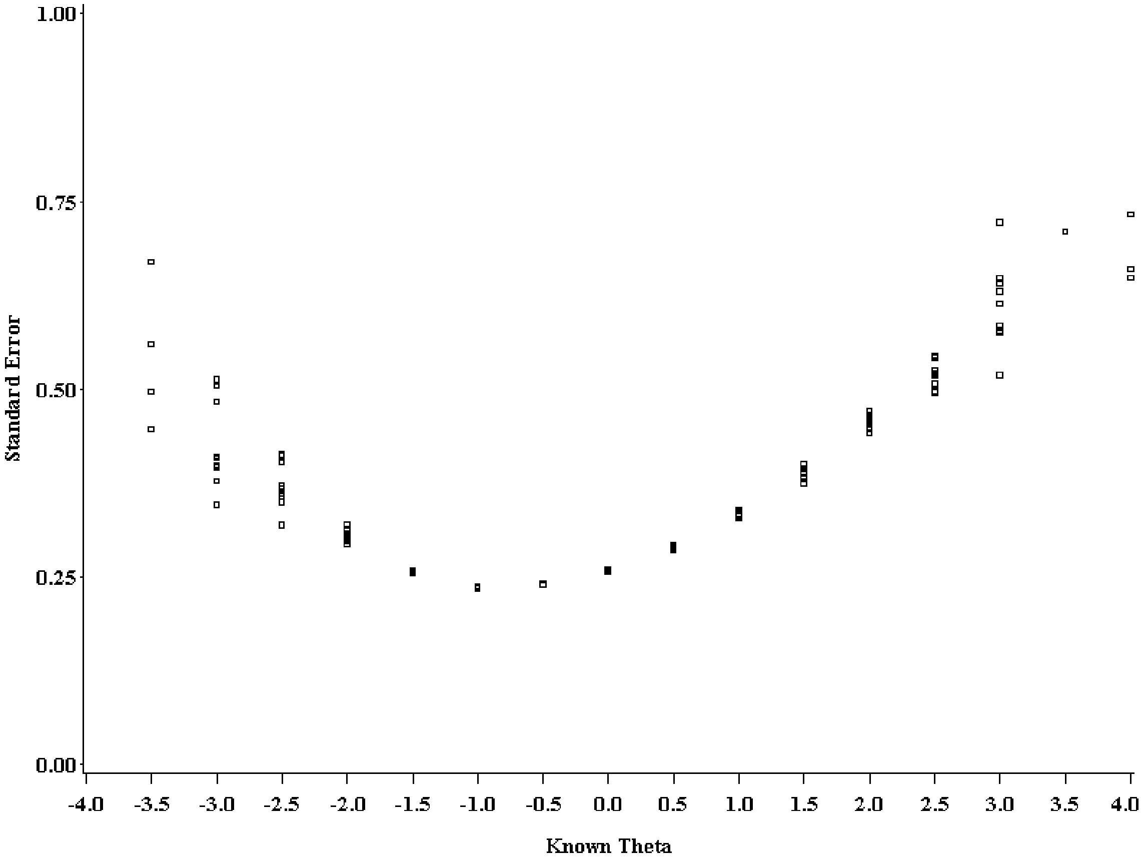


Figure B2: Conditional Standard Error Plot for Progressive Restricted with the 100 Item Pool

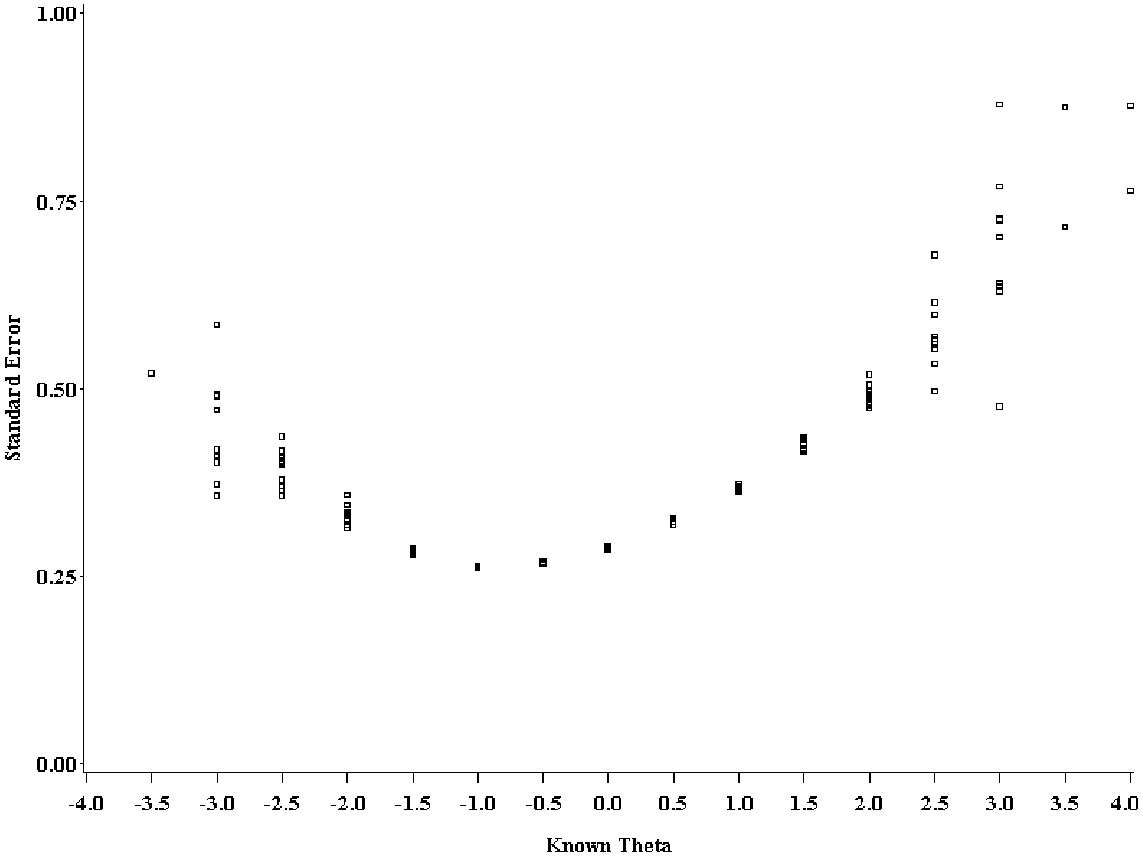


Figure B3: Conditional Standard Error Plot for Randomesque (Item Group Size = 3) with the 100 Item Pool

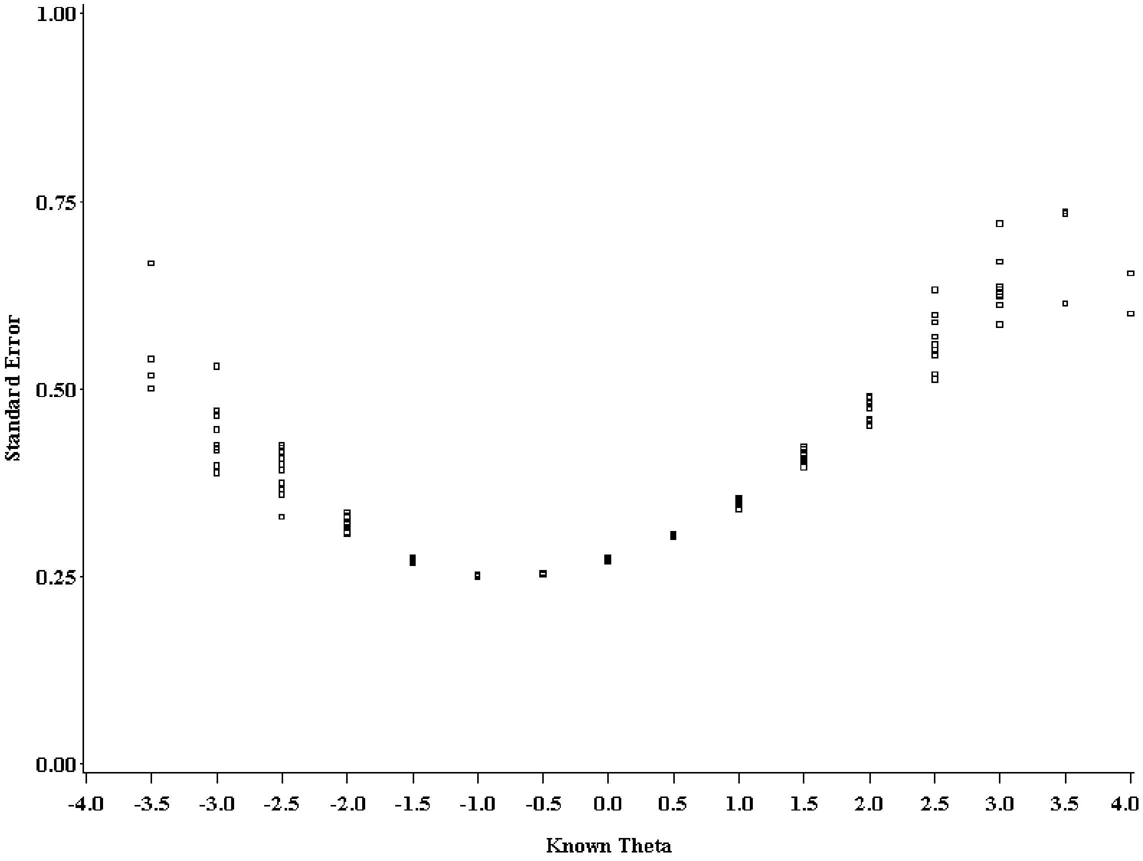


Figure B4: Conditional Standard Error Plot for Randomesque (Item Group Size = 6) with the 100 Item Pool

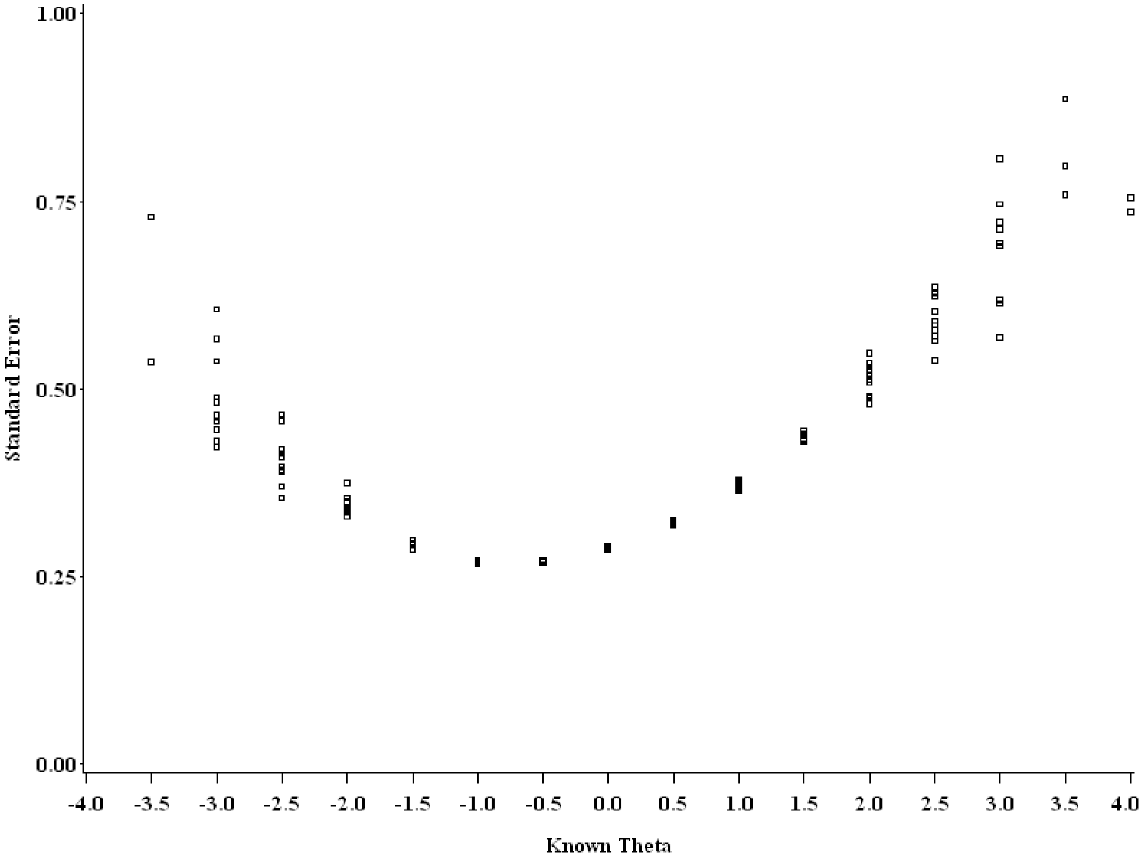


Figure B5: Conditional Standard Error Plot for Randomesque (Item Group Size = 9) with the 100 Item Pool

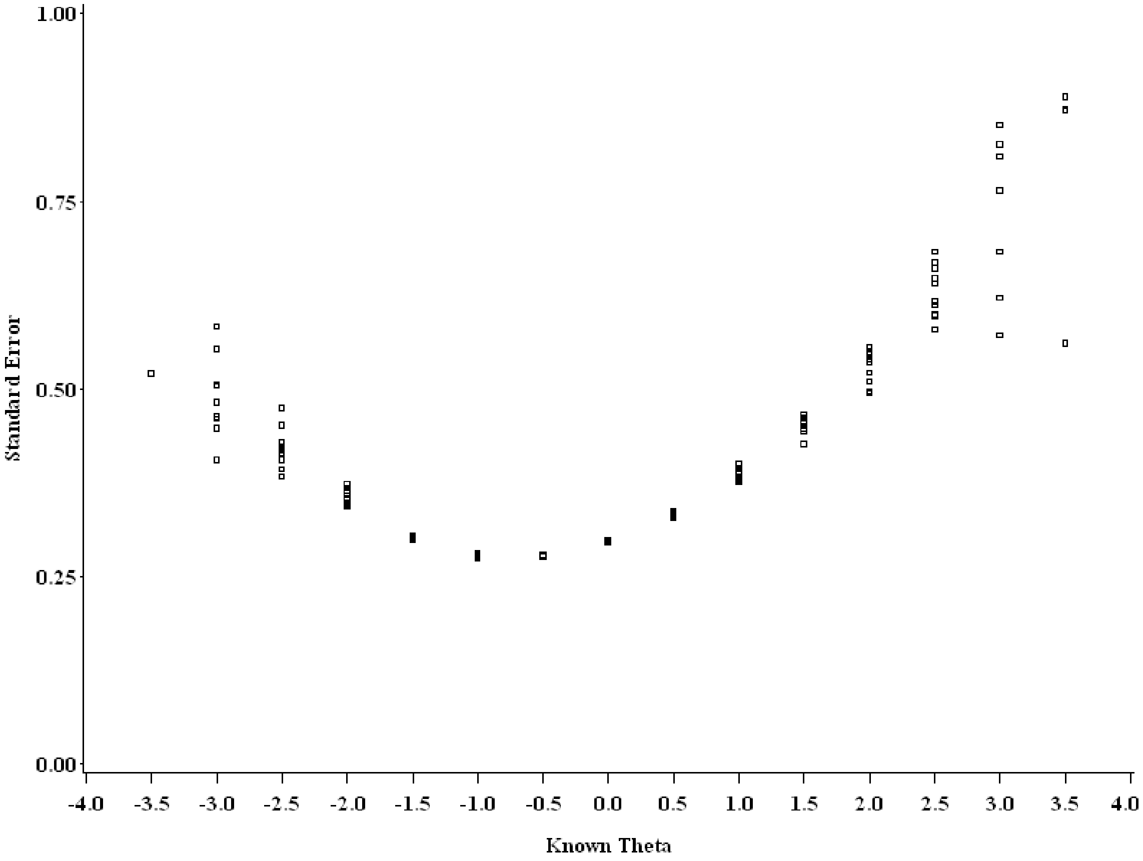


Figure B6: Conditional Standard Error Plot for Restricted Randomesque (Item Group Size = 3) with the 100 Item Pool

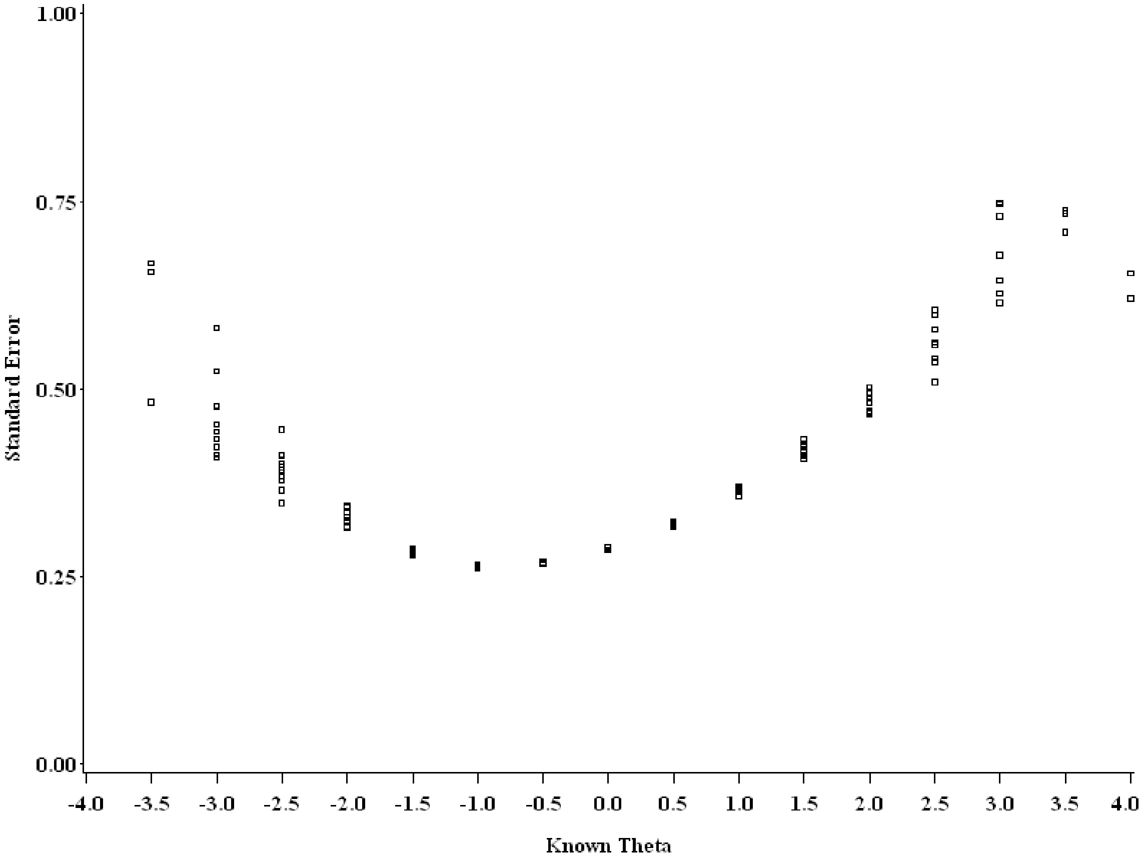


Figure B7: Conditional Standard Error Plot for Restricted Randomesque (Item Group Size = 6) with the 100 Item Pool

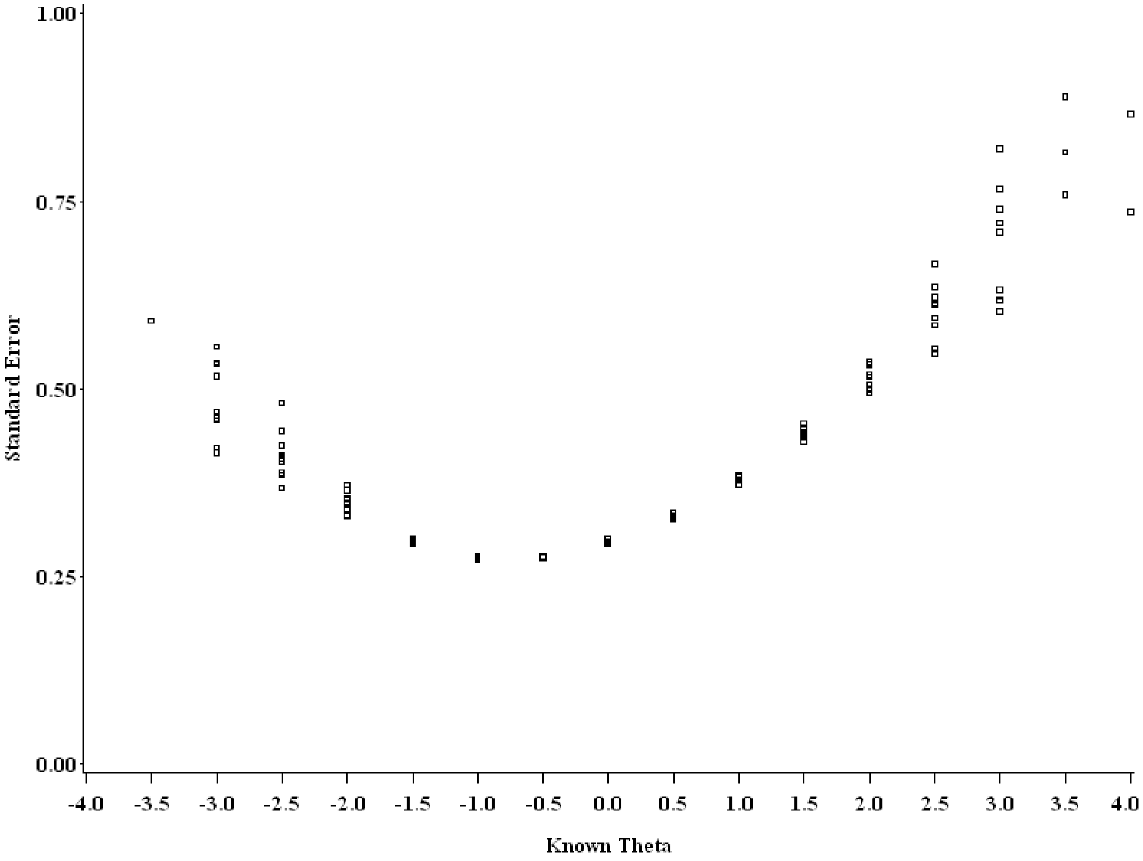


Figure B8: Conditional Standard Error Plot for Restricted Randomesque (Item Group Size = 9) with the 100 Item Pool

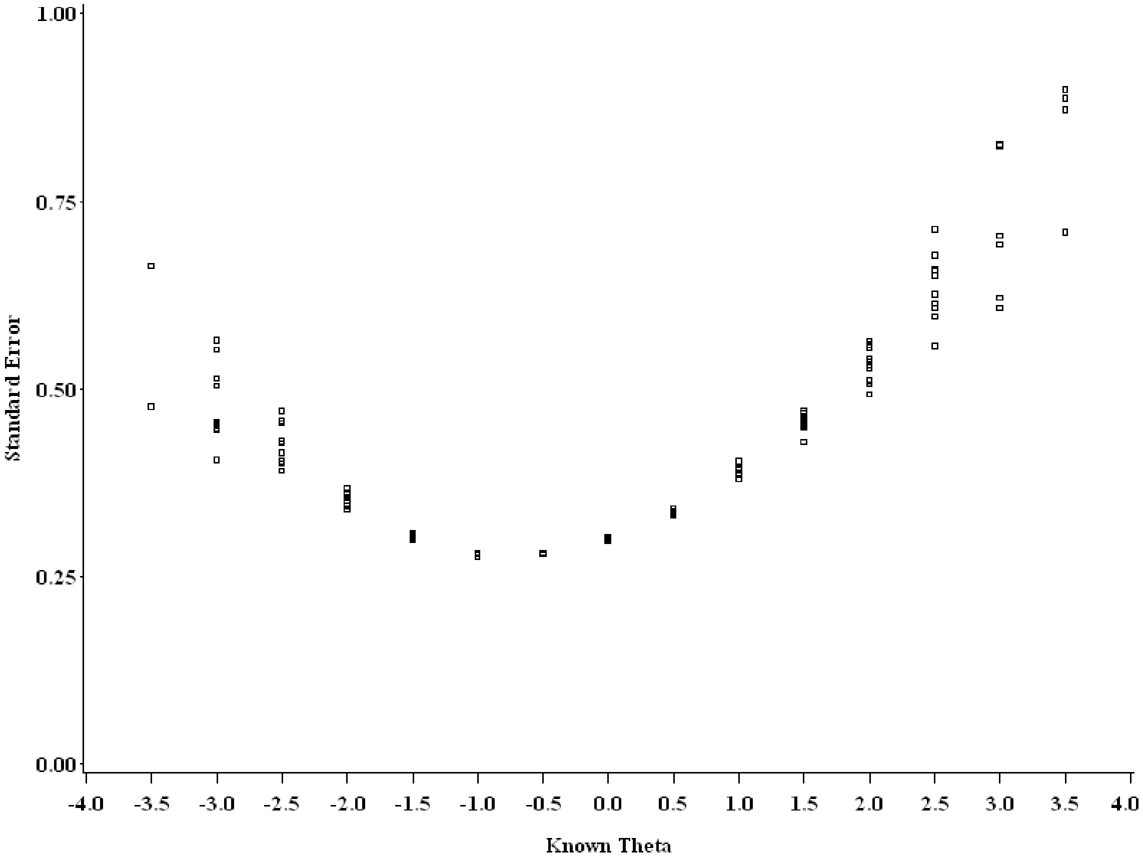


Figure B9: Conditional Standard Error Plot for Modified Within 0.10 logits

(Item Group Size = 3) with the 100 Item Pool

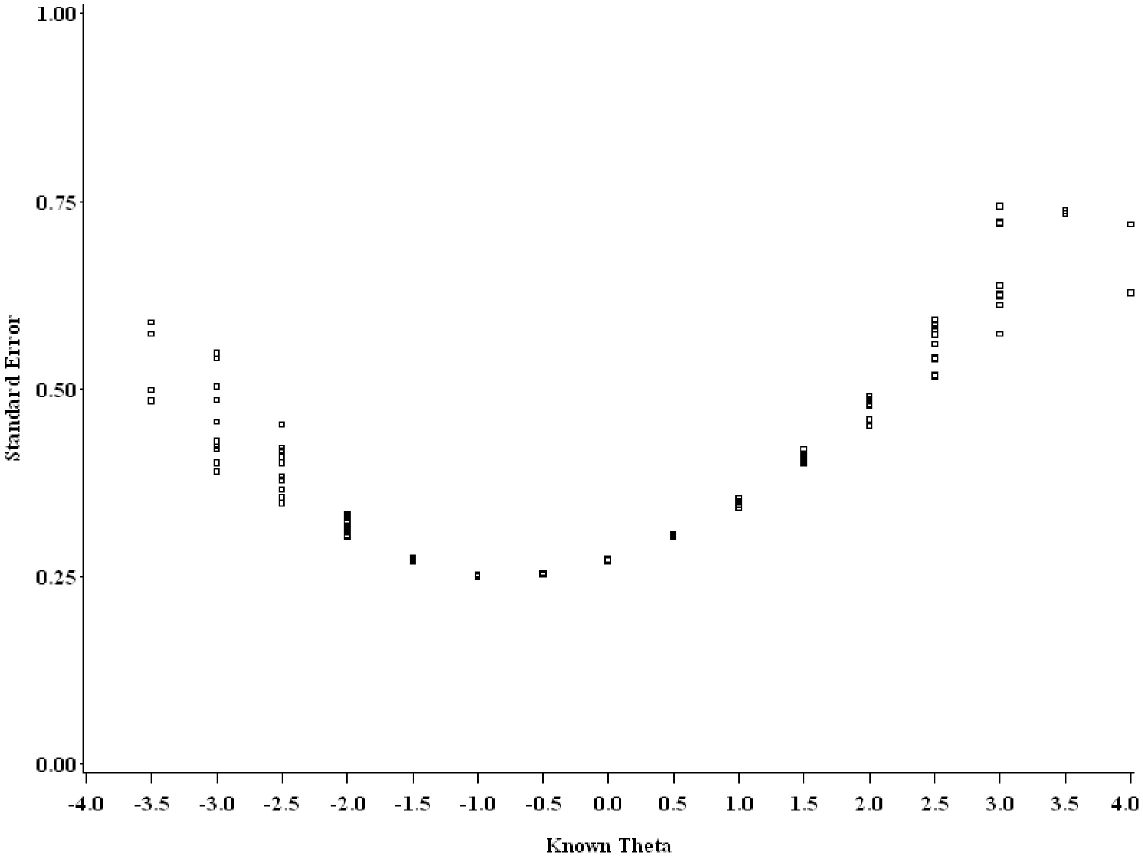


Figure B10: Conditional Standard Error Plot for Modified Within 0.10 logits

(Item Group Size = 6) with the 100 Item Pool

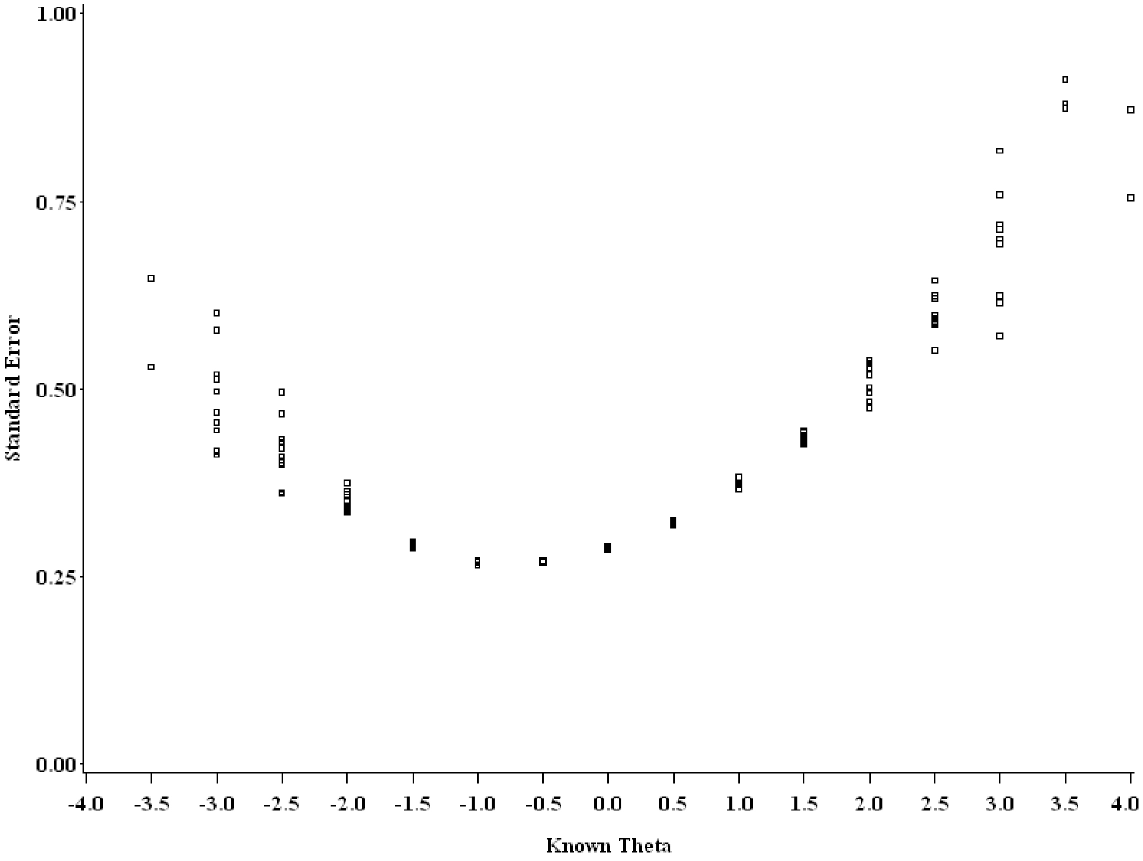


Figure B11: Conditional Standard Error Plot for Modified Within 0.10 logits

(Item Group Size = 9) with the 100 Item Pool

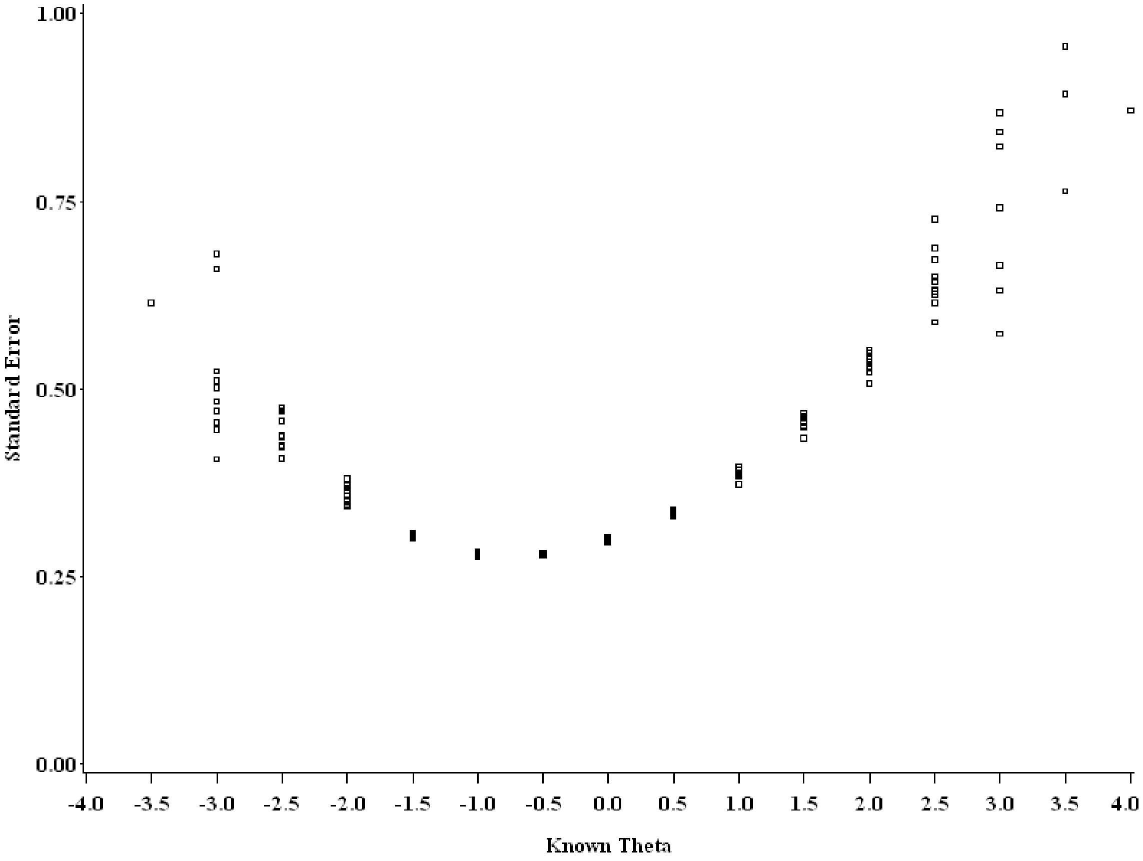


Figure B12: Conditional Standard Error Plot for Restricted Modified Within 0.10 logits

(Item Group Size = 3) with the 100 Item Pool

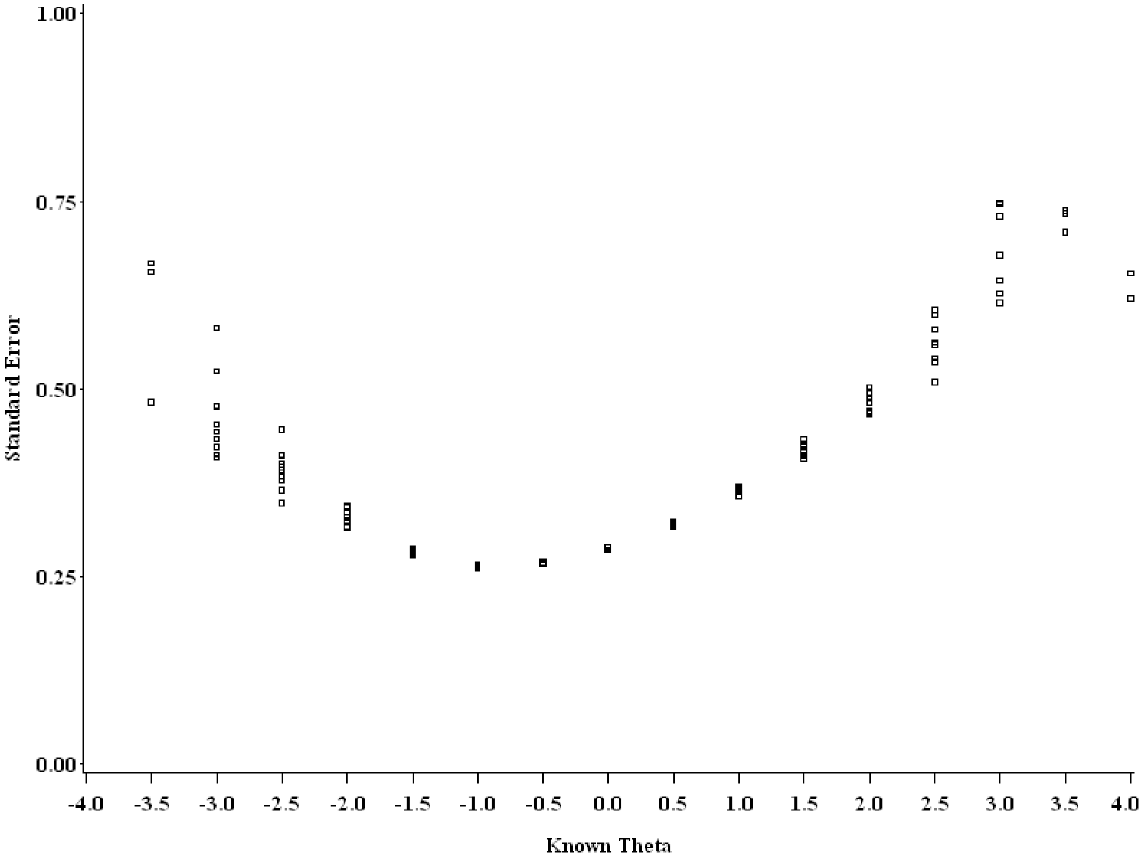


Figure B13: Conditional Standard Error Plot for Restricted Modified Within 0.10 logits

(Item Group Size = 6) with the 100 Item Pool

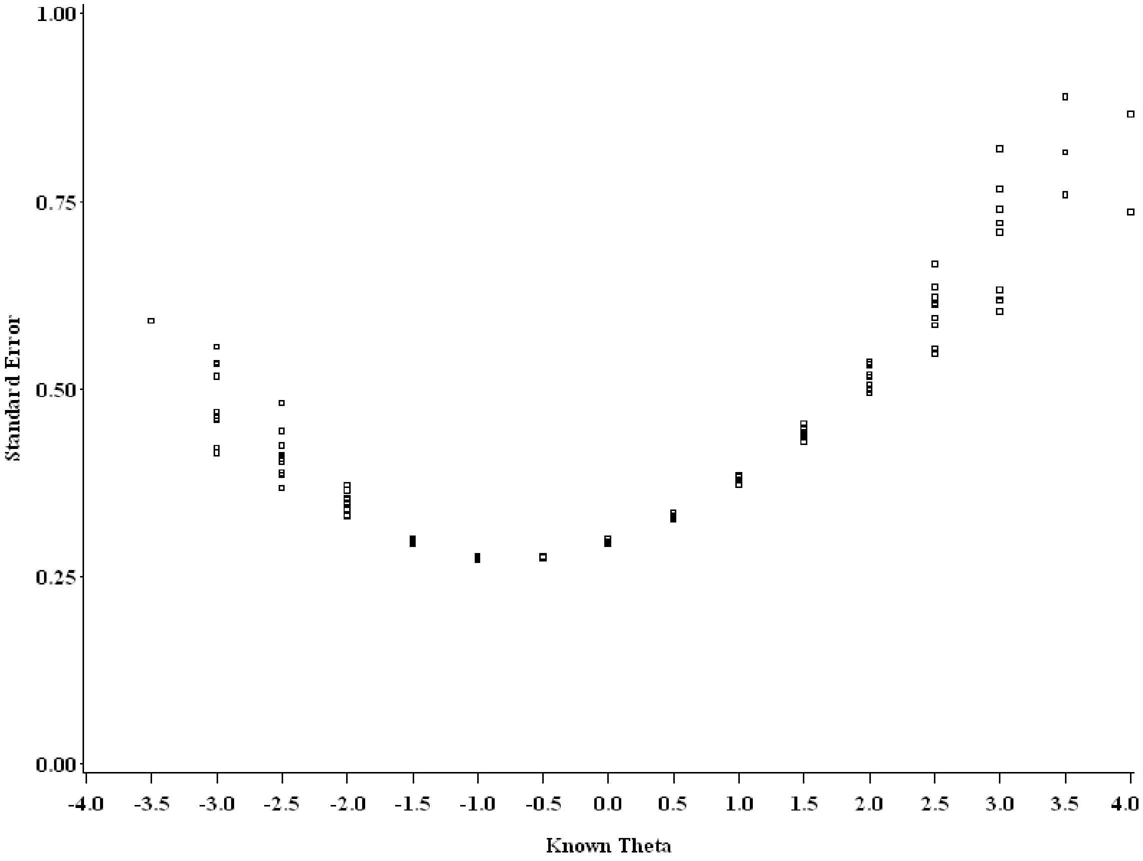


Figure B14: Conditional Standard Error Plot for Restricted Modified Within 0.10 logits

(Item Group Size = 9) with the 100 Item Pool

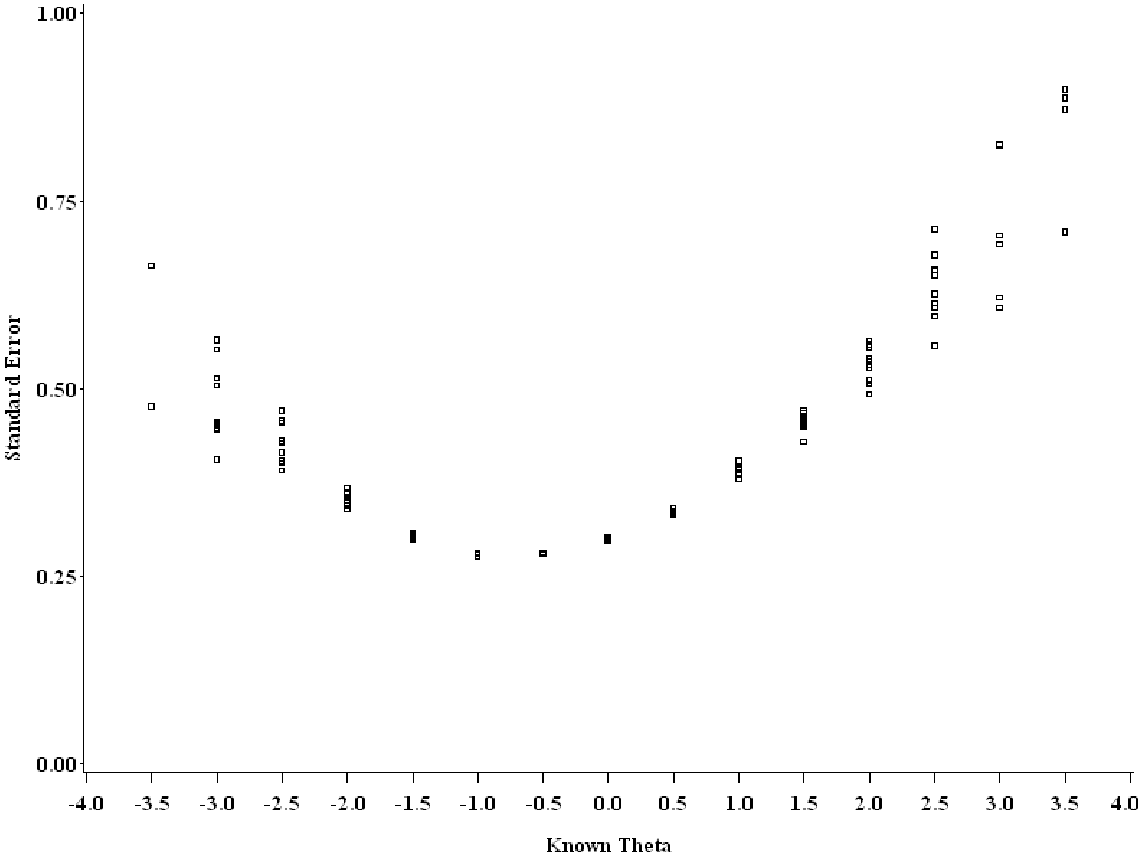


Figure B15: Conditional Standard Error Plot for Maximum Information (No Exposure Control) with the 200 Item Pool

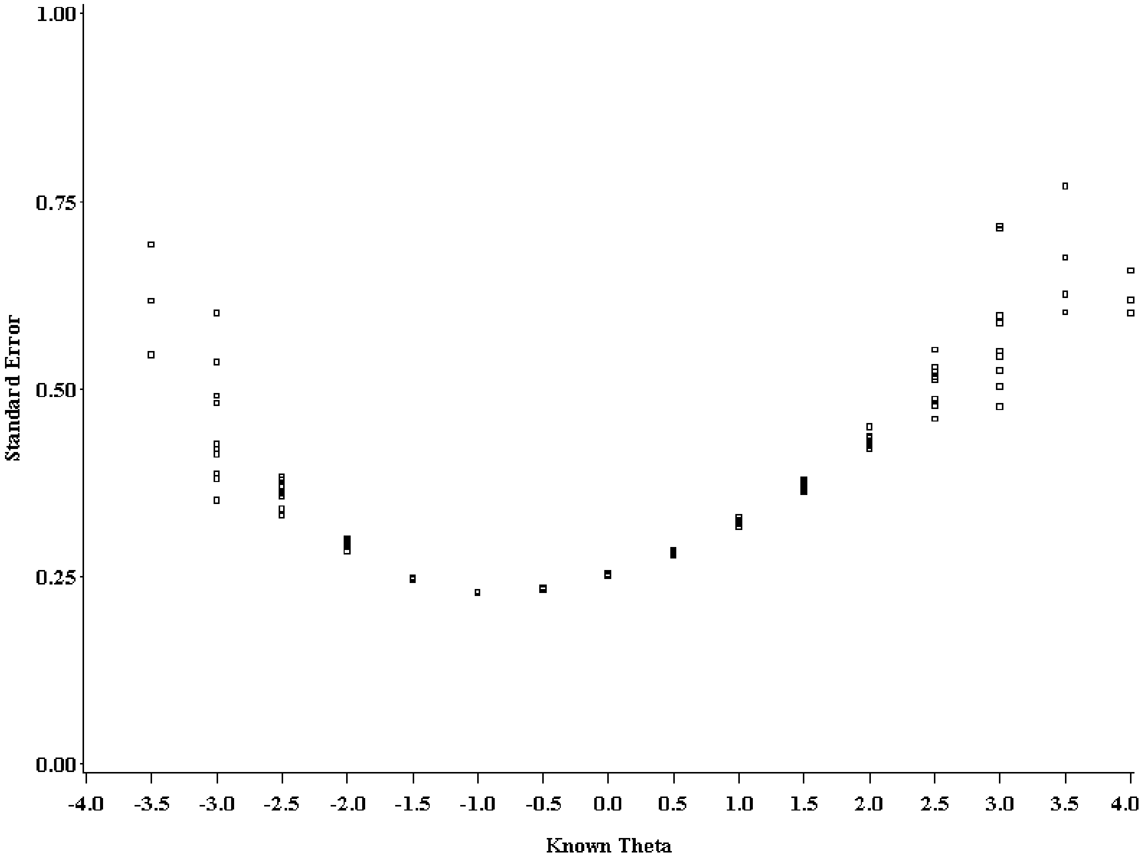


Figure B16: Conditional Standard Error Plot for Progressive Restricted with the 200 Item Pool

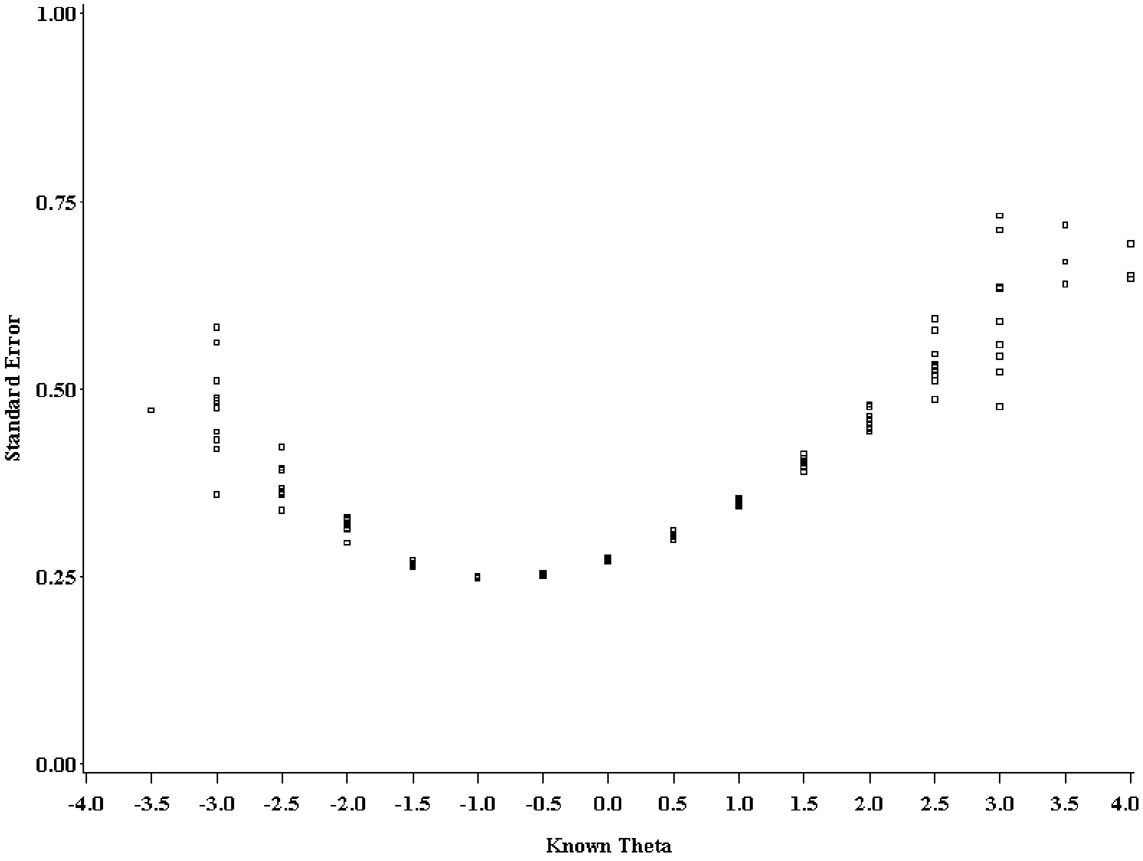


Figure B17: Conditional Standard Error Plot for Randomesque (Item Group Size = 3)

with the 200 Item Pool

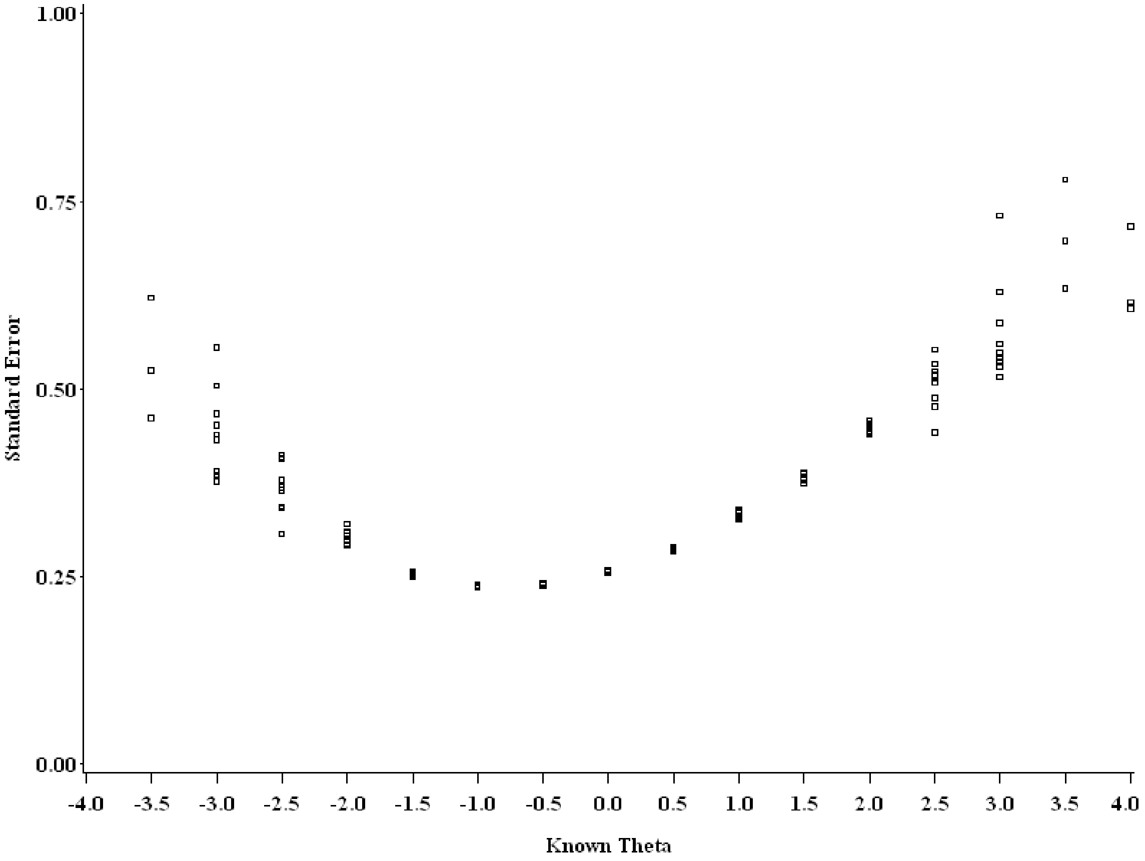


Figure B18: Conditional Standard Error Plot for Randomesque (Item Group Size = 6)

with the 200 Item Pool

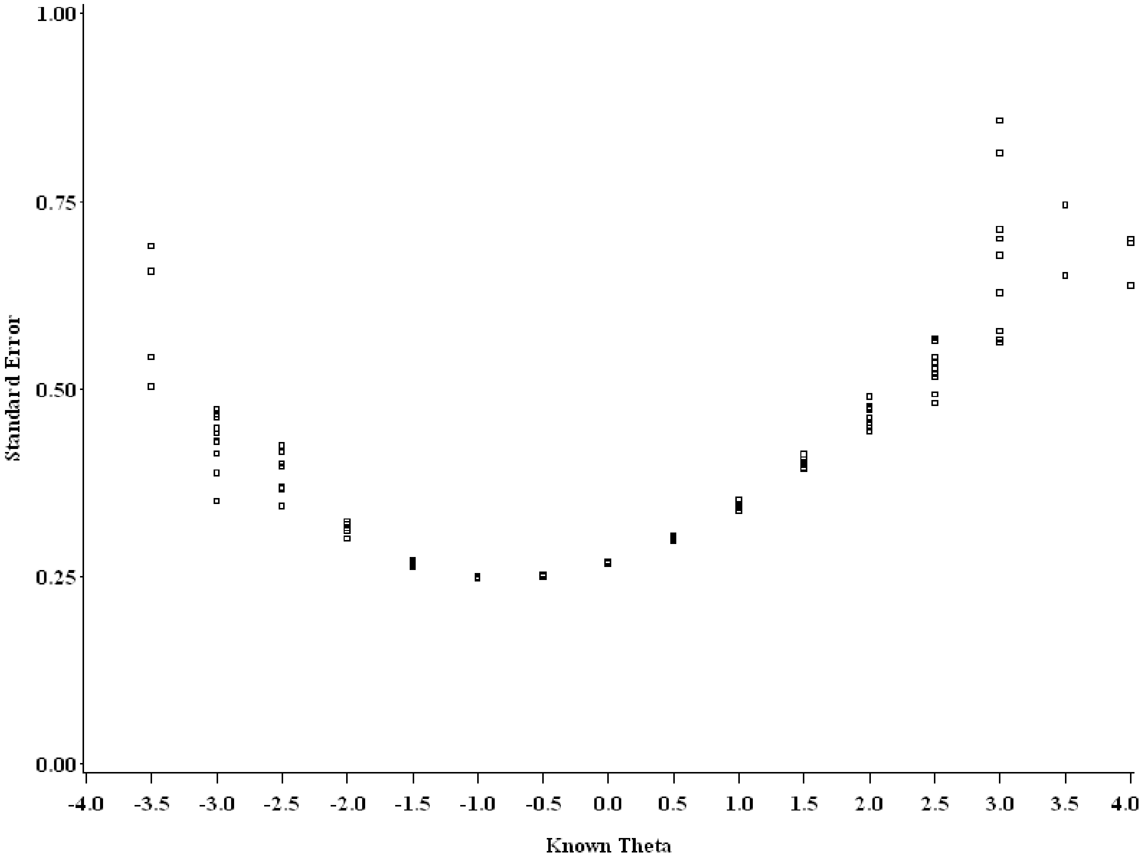


Figure B19: Conditional Standard Error Plot for Randomesque (Item Group Size = 9) with the 200 Item Pool

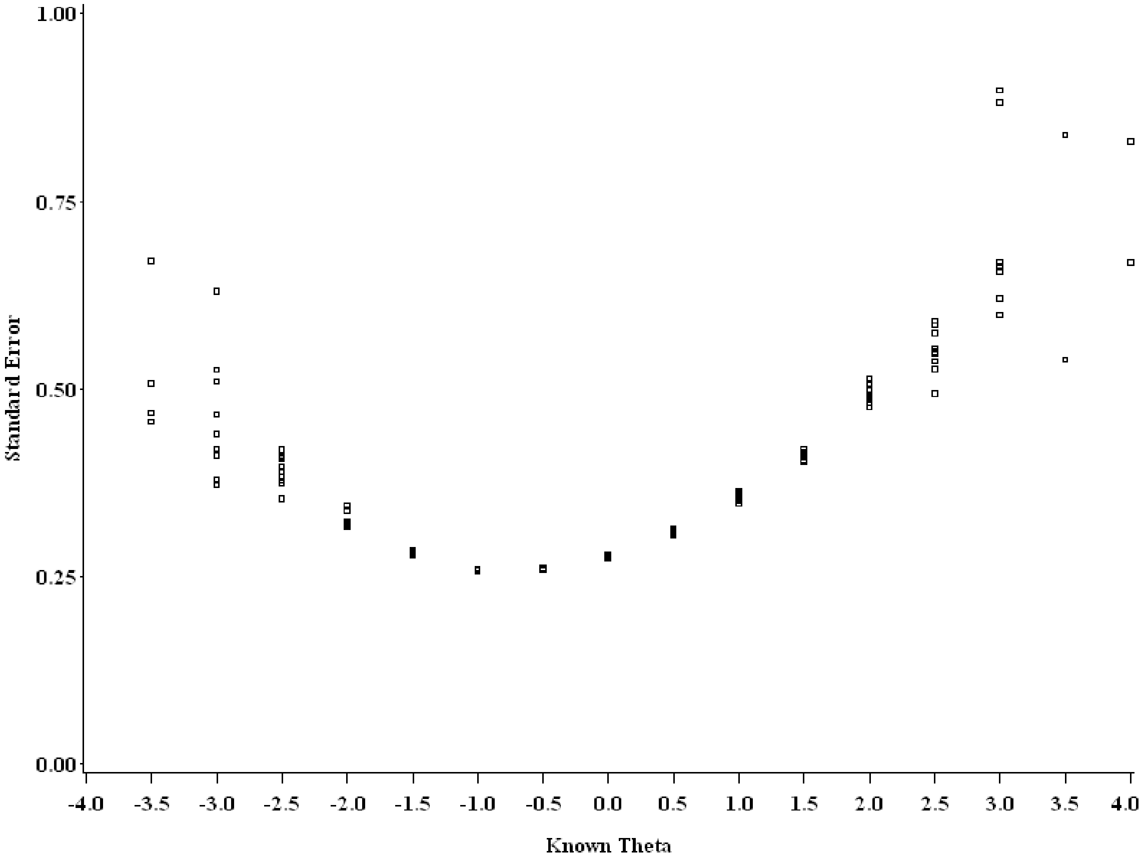


Figure B20: Conditional Standard Error Plot for Restricted Randomesque (Item Group Size = 3) with the 200 Item Pool

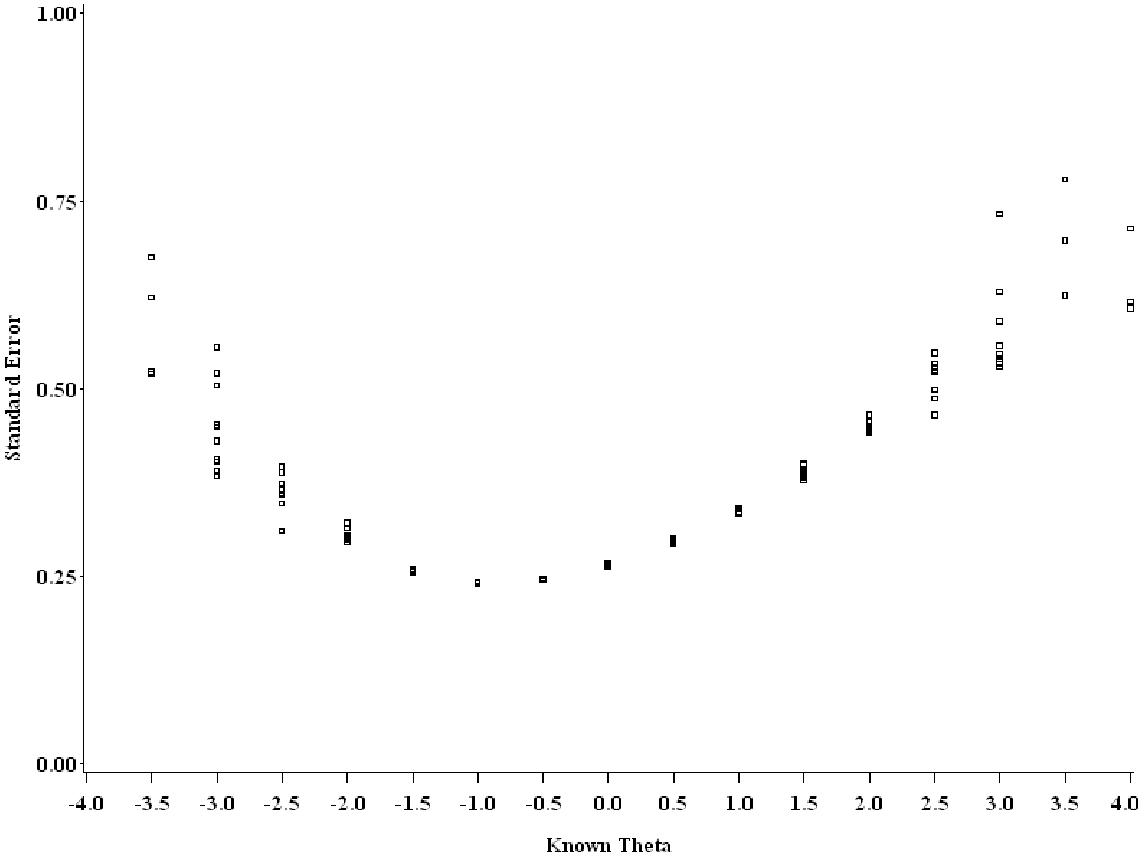


Figure B21: Conditional Standard Error Plot for Restricted Randomesque (Item Group Size = 6) with the 200 Item Pool

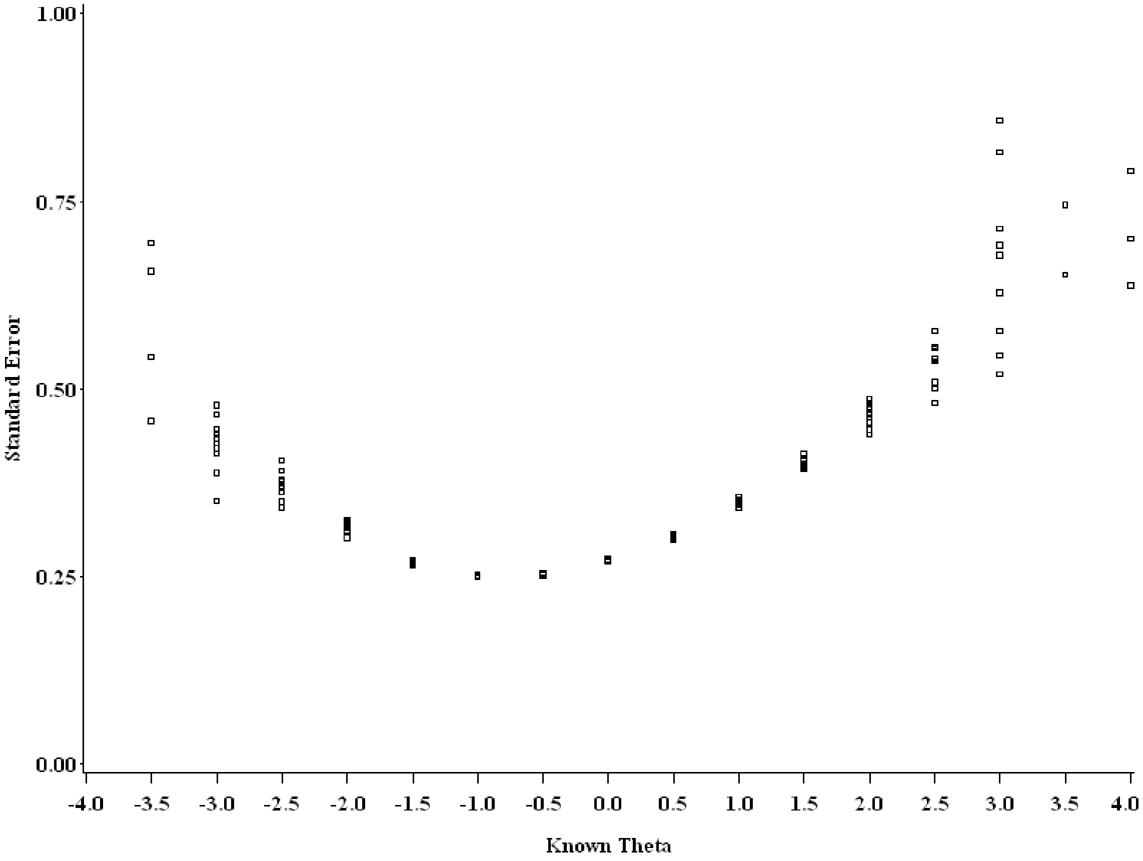


Figure B22: Conditional Standard Error Plot for Restricted Randomesque (Item Group Size = 9) with the 200 Item Pool

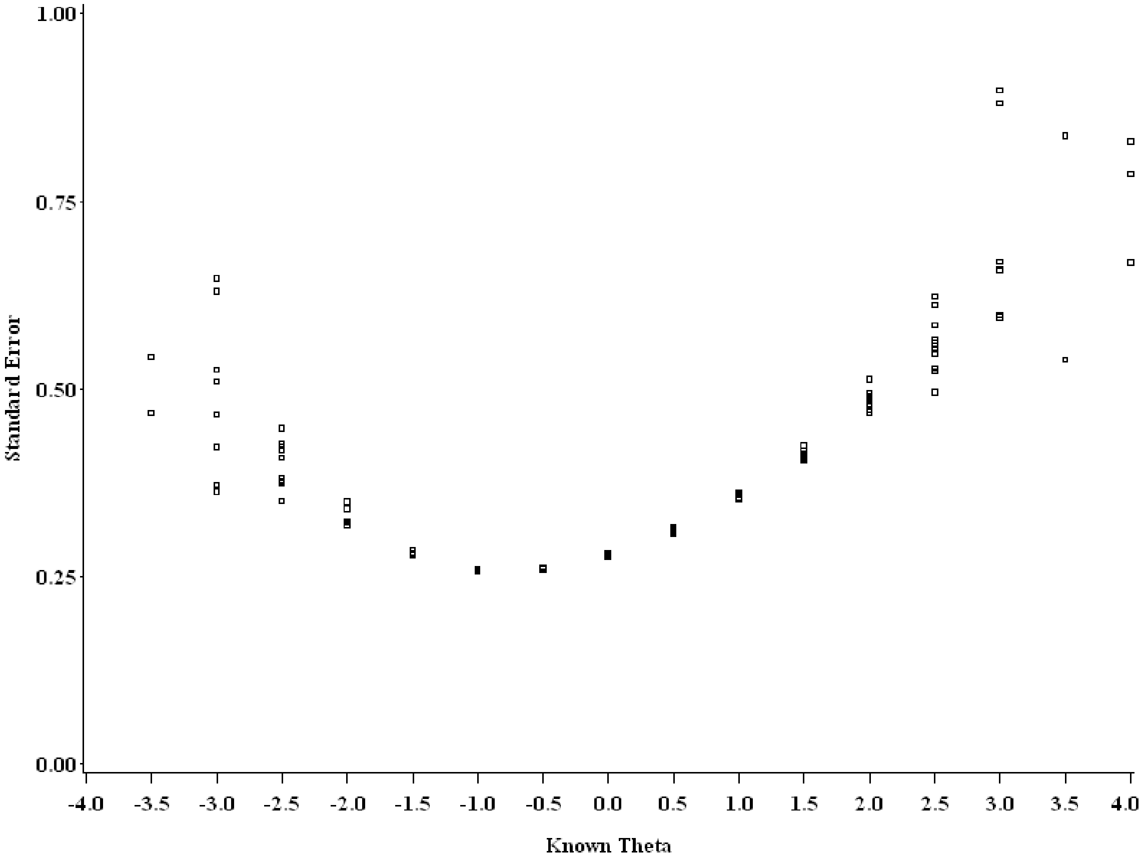


Figure B23: Conditional Standard Error Plot for Modified Within 0.10 logits

(Item Group Size = 3) with the 200 Item Pool

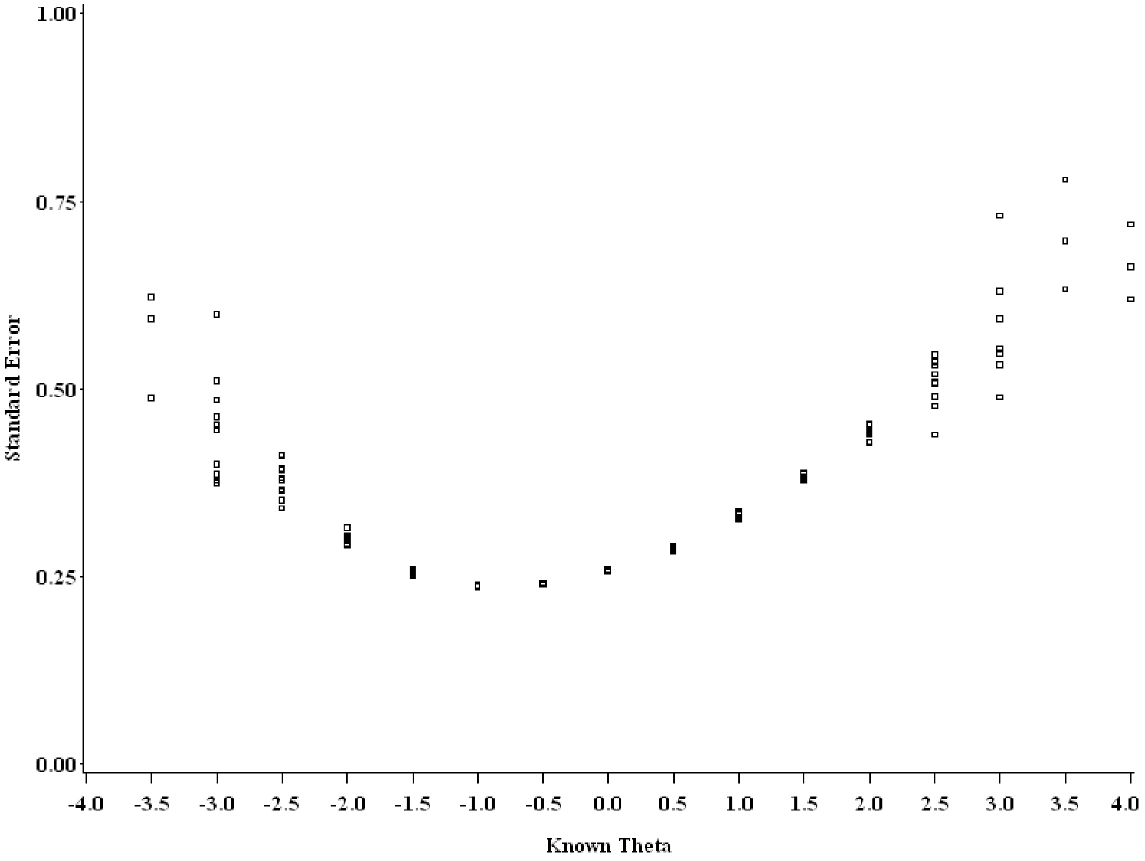


Figure B24: Conditional Standard Error Plot for Modified Within 0.10 logits

(Item Group Size = 6) with the 200 Item Pool

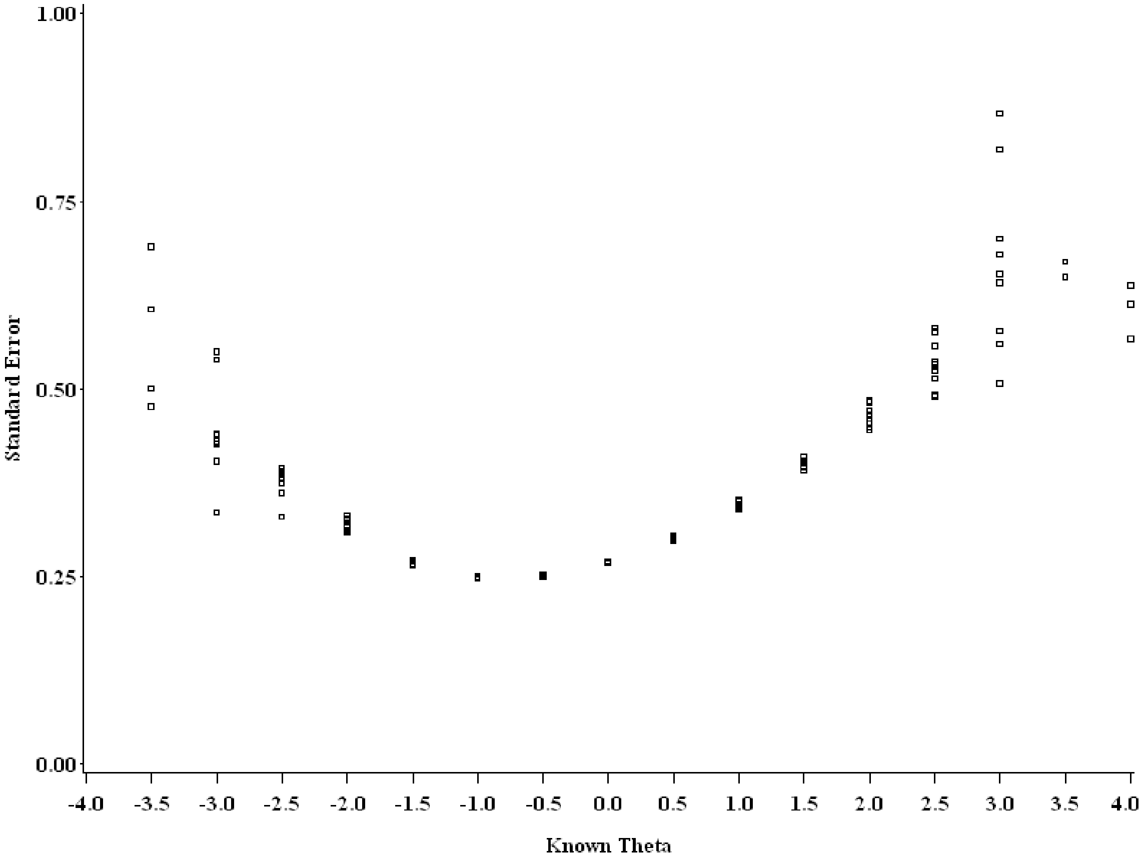


Figure B25: Conditional Standard Error Plot for Modified Within 0.10 logits

(Item Group Size = 9) with the 200 Item Pool

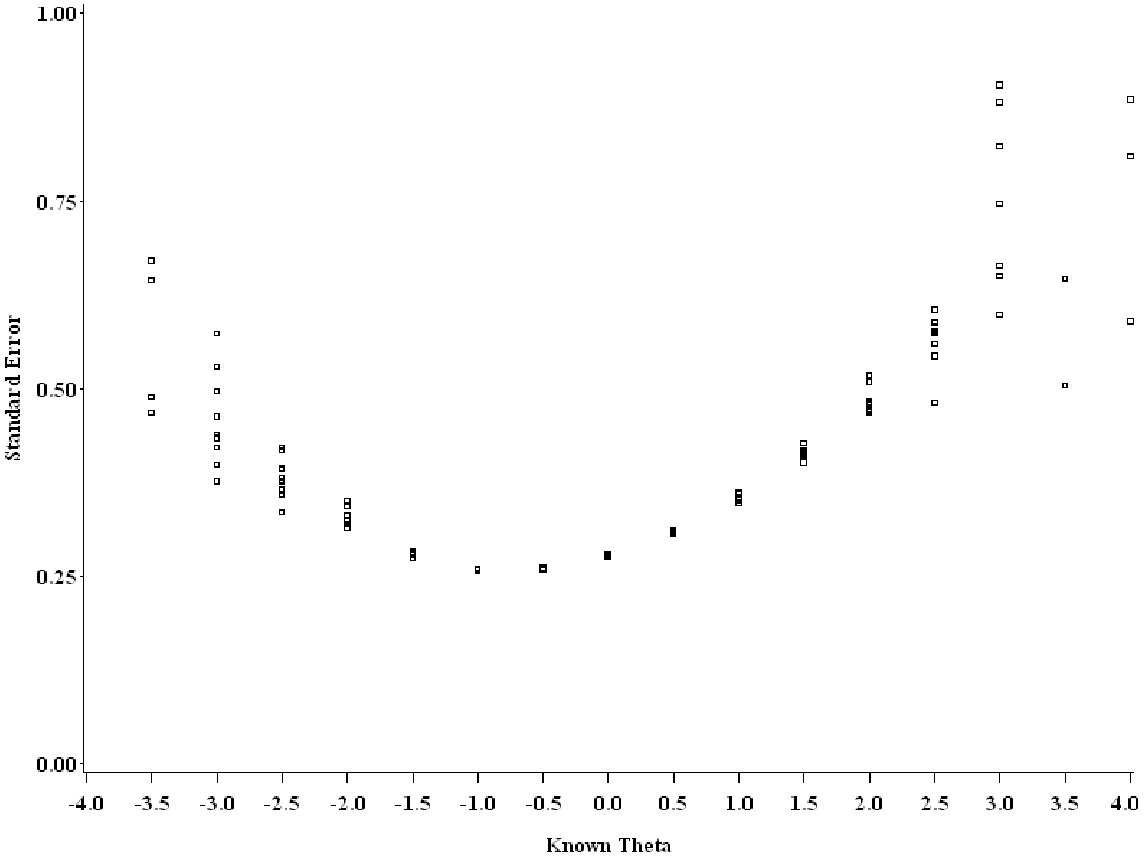


Figure B26: Conditional Standard Error Plot for Restricted Modified Within 0.10 logits

(Item Group Size = 3) with the 200 Item Pool

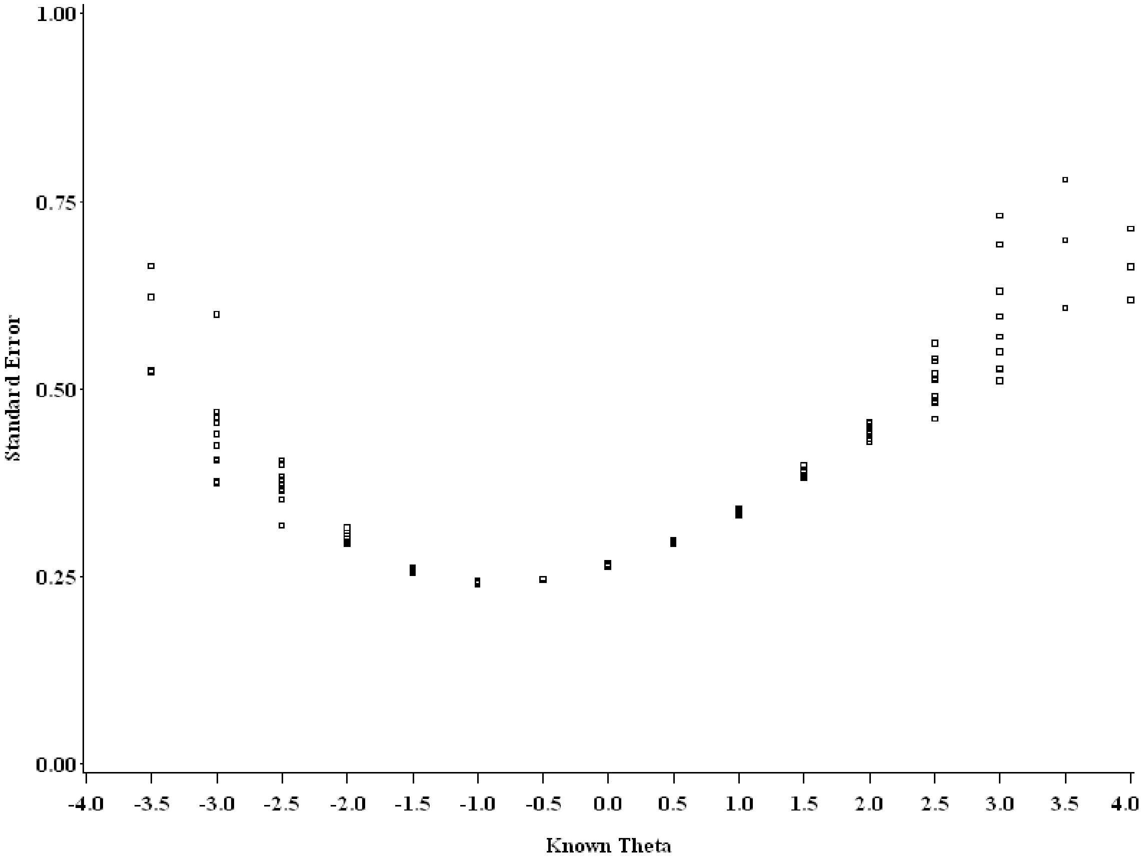


Figure B27: Conditional Standard Error Plot for Restricted Modified Within 0.10 logits

(Item Group Size = 6) with the 200 Item Pool

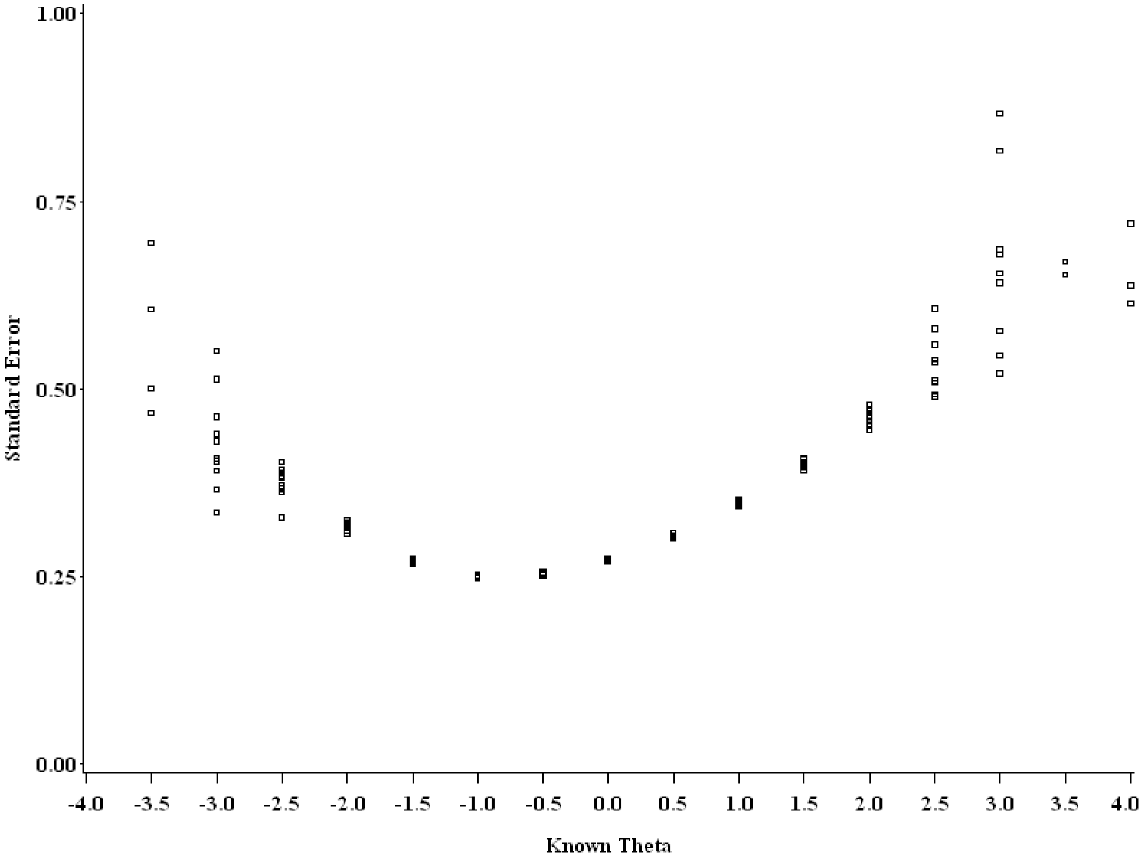
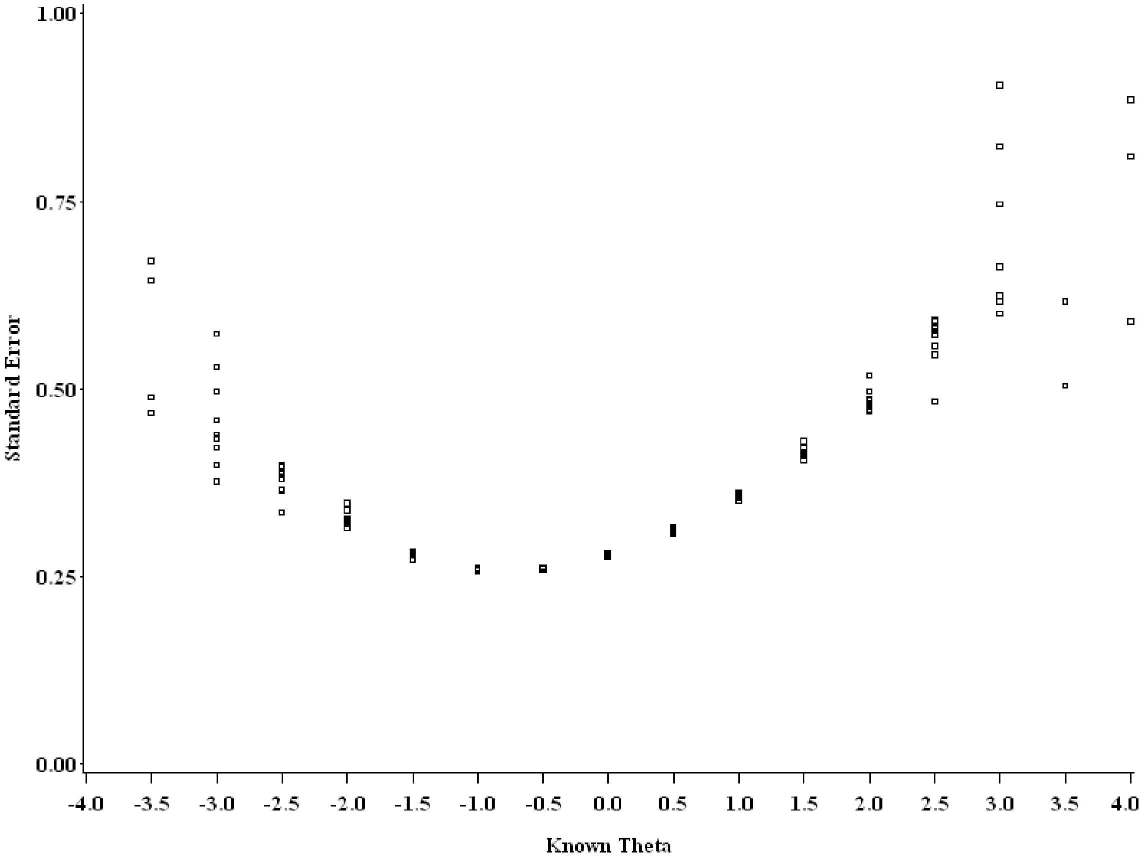


Figure B28: Conditional Standard Error Plot for Restricted Modified Within 0.10 logits

(Item Group Size = 9) with the 200 Item Pool



References

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

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*, 605–634.

Boyd, A. M. (2003). Strategies for Controlling Testlet Exposure Rates in Computerized Adaptive Testing Systems. (Doctoral Dissertation, The University of Texas at Austin, 2003) *Dissertation Abstracts International, 64*, 5835. (UMI No. 3110732)

Chang, S., Ansley, T. N. (2003). A Comparative Study of Item Exposure Control Methods in Computerized Adaptive Testing. *Journal of Educational Measurement*, *40 (1), 71–103*.

Chang, H., & Ying, Z. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211– 222.

Chang, H., Qian, J., & Ying, Z. (2001). *a*-stratified Multistage Computerized Adaptive Testing with *b*-blocking. *Applied Psychological Measurement, 25(4)*, 333-341.

Chen, S. (1997). A comparison of maximum likelihood estimation and expected a posteriori estimation in computerized adaptive testing using the generalized partial credit model. (Doctoral Dissertation, The University of Texas at Austin, 2001) *Dissertation Abstracts International, 58*, 453. (UMI No. 9710321)

Chen, S., Hou, L., Fitzpatrick, S. J., & Dodd, B. G. (1997). The effect of population distribution and method of theta estimation on computerized adaptive testing

(CAT) using the rating scale model. *Educational & Psychological Measurement, 57(3), 422-439.*

Childs, R., & Oppler, D. (2000). Implications of test dimensionality for unidimensional IRT scoring: An investigation of a high stakes testing program. *Educational and Psychological Measurement, 60(6)*, 939–955.

Cook, K. F., Dodd, B. G., & Fitzpatrick, S. J. (1999). A comparison of three polytomous item response theory models in the context of testlet scoring. *Journal of Outcome Measurement, 3(1),* 1-20.

Davis, L.L. (2004). Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. *Applied Psychological Measurement, 28(3)*, 165-185.

Davis, L.L. (2002). Strategies of Controlling Item Exposure in computerized adaptive testing with polytomously scored items. (Doctoral dissertation, The University of Texas at Austin, 2001). *Dissertation Abstract International, 64, 458.* (UMI No. 3077522)

Davis, L.L. & Dodd, B.G. (2005, March). *Strategies for controlling item exposure in computerized adaptive testing with the partial credit model*. (Research Report 0501). Retrieved January 29, 2008 from the Pearson Educational Measurement website: http://www.pearsonsolutions.com/resources/research.htm

Davis, L.L. & Dodd, B.G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement, 27(5)*, 335-356.

Davis, L. L., & Dodd, B. G. (2001). *An examination of testlet scoring and item exposure constraints in the verbal reasoning section of the MCAT.* MCAT Monograph Series: Association of American Medical Colleges.

Davis, L.L., Pastor, D.A., Dodd, B.G., Chiang, C., & Fitzpatrick, S. (2003). An examination of exposure control and content balancing restrictions on item selection in CATs using the partial credit model. *Journal of Applied Measurement, 4*, 24 -42.

Dodd, B.G., Koch, W.R., De Ayala, R.J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13*, 129-143.

Dodd, B.G., De Ayala, R.J., & Koch, W.R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19(1),* 5-22.

Flaugher, R. (2000). Item Pools. In Wainer, Howard (Ed). *Computerized adaptive testing: A primer (2nd ed.)*. pp. 271-299. Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R., & Swaminathan, H. (1985). *Item response theory: Principles and applications. B*oston MA: Kluwer-Nijhoff.

Hetter, R.D., & Sympson, J.B. (1997). Item exposure control in CAT_ASVAB. In William Sands, Brain K. Waters, and James R. McBride (Eds.), *Computerized adaptive testing-from inquiry to operation* (pp. 141-144). Washington, D.C.: American Psychological Association.

Johnson, M.A. (2006). An investigation of stratification exposure control procedures in CATs using the generalized partial credit model. (Doctoral dissertation, The University of Texas at Austin, 2006). *Dissertation Abstract International, 68, 05.* (UMI No. 3266891)

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

Leung, C., Chang, H & Hau, K. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympson-Hetter algorithm. *Applied Psychological Measurement, 26*, 376-392.

Luecht, R.M., & Nungester, R.J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*, 229-249.

Lunz, M.E., & Stahl, J.A. (1998). *Patterns of item exposure using a randomized CAT algorithm*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Muraki, E. & Bock, R. D. (1993). The PARSCALE computer program [Computer program]. Chicago, IL: Scientific Software International.

Owen, R.J. (1969). A Bayesian approach to tailored testing (RB-69-92). Princeton, NJ: Educational Testing Service.

Parshall, C.G., Davey, T., & Nering, M.L. (1998). *Test development exposure control for adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Pastor, D., Dodd, B. G., Chang, H. (2002). A comparison of item selection techniques and exposure control mechanism in CATs using the generalized partial credit model. *Applied Psychological Measurement, 26(2)*, 147-163.

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

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34* (4, Pt. 2, No 17).

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567-577.

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

Wainer, H. (2000). CATs: Whither and whence. *Psicologic, 21(1-2)*, 121-133.

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

Wang, S. (1999). The accuracy of ability estimation methods for computerized adaptive testing using the generalized partial credit model. (Doctoral dissertation, University of Pittsburgh, 1999). *Dissertation Abstracts International, 60(09)*, p.3274. (UMI No. 9945102).

Way. W. D. (1998). *Practical Questions in Introducing Computerized Adaptive Testing for K-12 Assessments*. Retrieved January 29, 2008 from the Pearson Educational Measurement website: http://www.pearsonsolutions.com/resources/research.htm

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

Whittaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement*, 27(4), 299-300.

Yi, Q., & Chang, H. (2003). *a*-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology, 56(2)*, 359-378.

Yi, Q., & Chang, H. H. (2000). Multiple stratification CAT designs with content control. Unpublished manuscript.

VITA

Edgar Isaac Sanchez was born in Eagle Pass, Texas on December 11, 1978 the son of Rodolfo Sanchez and Patricia Esther Sanchez. In June of 1995 he began his undergraduate career at Laredo Community College in Laredo, Texas. In August 1995 he entered the Texas Academy for Leadership in the Humanities, an early college enrollment program. In this program he satisfied the requirements for high school graduation by enrollment at Lamar University in Beaumont, Texas. After completing two years of college coursework there he graduated from John B Alexander High School in Laredo, Texas in 1997. In 1997 he continued his undergraduate coursework at The University of Texas at Austin. He received the degree of Bachelor of Arts from The University of Texas at Austin in 1999. He entered the Graduate School of The University of Texas at Austin in the Department of Educational Psychology studying quantitative methods and psychometrics in August of 1999. He received the degree of Master of Arts from The University of Texas at Austin in 2004.