The Dissertation Committee for Candace L. Macken-Ruiz certifies that this is the approved version of the following dissertation:

A COMPARISON OF MULTI-STAGE AND COMPUTERIZED ADAPTIVE TESTS BASED ON THE GENERALIZED PARTIAL CREDIT MODEL

Committee:

Barbara G. Dodd, Supervisor

S. Natasha Beretvas

Tiffany A. Whittaker

Linda L. Hargrove

Daniel A. Powers

A COMPARISON OF MULTI-STAGE AND COMPUTERIZED ADAPTIVE TESTS BASED ON THE GENERALIZED PARTIAL CREDIT MODEL

by

Candace L. Macken-Ruiz, B.A., M.S., 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

August 2008

A COMPARISON OF MULTI-STAGE AND COMPUTERIZED ADAPTIVE TESTS BASED ON THE GENERALIZED PARTIAL CREDIT MODEL

Candace L. Macken-Ruiz, Ph.D.

The University of Texas at Austin, 2008

Supervisor: Barbara G. Dodd

A multi-stage test (MST) design is an alternative design for the delivery of automated tests. While computerized adaptive tests (CAT) have dominated testing for the past three decades, increasing interest has been focused on the MST because it offers two advantages that CAT does not: Test sponsors and test developers can see an entire test before administration because it is pre-constructed from sets of modules of test items, and within a module examinees may skip forward and back through test items and make changes to previously answered items. Due to the dominance of CAT, little research has been devoted to differing MST designs with regard to the number of items per stage and routing rules that direct the selection of the next module after a previous module has been completed. This research used simulated response data for a large national test and the generalized partial credit model to compare a CAT to one of three MST designs that had either decreasing numbers of items per stage, increasing number of items per stage, or the same number of items per stage, and one of three routing rules, maximum information, fixed θ, or number-right routing. As anticipated, CAT had the best performance with respect to estimating proficiency and item pool use. Among the MSTs, the MST with

increasing numbers of items per stage performed the best with respect to estimating proficiency, followed by the MST with decreasing number of items per stage, and equal numbers of items per stage. By routing rule, maximum information performed the best and number-right routing performed the worst. Only one panel was constructed per MST design, so only limited comparisons of item pool use could be made. Although the MST designs did not perform as well as CAT, the differences in estimating proficiency were not large, implying that the MST design is a viable alternative to CAT.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE: INTRODUCTION

Adaptive testing has been dominated by the computerized adaptive test (CAT) paradigm since the early 1970s. In a CAT, test items are adaptively selected based on an examinee's response to a previous question. At test outset an item is selected based on a guess about the examinee's proficiency level, usually the mean of the hypothesized proficiency distribution. If the examinee obtains a correct response s/he is administered an item that is more difficult than the previous item. If the answer to the previous question is incorrect, s/he is administered an item that is easier than the previous one. The decision to administer a more difficult or an easier question is governed by a computer algorithm that computes the examinee's estimated proficiency level based on the responses to the previous question(s). This selection and administration process, with an interim re-estimate of ability prior to the selection of the next item, continues until a prespecified number of items is reached or until the precision of the ability estimate reaches a specified maximum.

Since items are dynamically and adaptively selected one-by-one in the CAT paradigm, there are theoretically a vast number of tests that can be constructed and administered. In fact, because of limitations such as the number of items in the item pool, the desire to have the test questions reflect specific content areas in a given test, and the desire to limit the reuse of items across all tests administered, the number of tests that can be constructed is somewhat less. Yet even when taking these limitations into account, CATs are very efficient tests because a stable estimate of proficiency can be reached with fewer items than a traditional paper-and-pencil test (Olsen, Maynes,

Slawson & Ho, 1986; Kingsbury & Weiss, 1983; Weiss & Kingsbury, 1984). This efficient model has a drawback, however: Due to the adaptive item selection process, test developers and test sponsors cannot review an entire test because the test is tailored to the examinee taking it. In other words, a developer or sponsor can review the individual items in the item pool before testing goes live, but s/he cannot review an entire test because the items included in it may vary from person to person.

A CAT is designed to proceed in a linear fashion: An item is presented; it is answered; proficiency is estimated; and the next item is presented. Since proficiency is estimated after a response is made to an item and the selection of the next item is conditioned on the current proficiency estimate (which in turn is based on the set of previous responses) in the CAT paradigm, an examinee cannot skip back to a previous question in order to change an answer. Since there is no set order of item presentation (as in the paper-and-pencil case), and the selection of the next item is based on the response to the item currently presented, an examinee cannot choose to skip the currently presented item and return to it later. There is no circularity of skipping and/or returning to an item allowed in a CAT in most cases, there is only a forward progression in the test. Anything other than the forward progression would invalidate the estimation of proficiency and the selection of the subsequent items. This leads to what some see as a second drawback. Clearly, while efficient, the CAT paradigm limits a test-taking strategy.

A multi-stage test (MST) is an alternate, adaptively administered design. Its history begins in the 1970s as well, but it is not in as common operational use as CAT. Rather

than a test consisting of items that are selected adaptively and administered on a one-by-one basis, an MST is a pre-constructed test panel that adaptively administers *sets* of items to examinees. The MST taxonomy comprises a panel (the test), a number of stages within a panel, a number of modules (item sets) within a stage, and a number of items within a module. Prior to any testing, multiple panels with equivalent efficiency for estimating proficiency are constructed. Within the panel are two or more stages that in turn typically comprise two or more modules, or sets of items, each. At each stage, MST modules usually are designed to represent an overall continuum of difficulty from easy to difficult. Testing begins by randomly selecting a test panel for administration. Once the panel has been selected, proficiency estimation and module (item set) selection processes are like those in CAT. The first module administered is usually designed to be of moderate difficulty, targeting examinees at the mean of the hypothesized proficiency distribution. Unlike CAT, proficiency is estimated only after the entire module has been administered. Depending on the proficiency estimate after the completion of the first stage module, the appropriate module is adaptively chosen at the next stage of the test.

The MST paradigm offers what may be seen by some as having advantages over CAT from the perspective of both the test developer and the test consumer (Hendrickson, 2007). Since an MST panel is pre-constructed, test developers or test sponsors may review an entire test as it will be administered to examinees, and adjustments to the panel in terms of content and item use may be made in advance of live test administration. Since proficiency is not estimated until after a module is completed, examinees may skip back to earlier questions within the same module and change answers or skip over

questions and return to them later. Although once a stage is completed, one is precluded from returning to a previous stage (Luecht, 2003).

For purposes of the present research, there are four salient features of MST designs. These include the measurement model, the routing rule between stages, the number of stages per panel, and the number of items per stage. Two of these features, the measurement model and the routing rule, are shared in common with the CAT paradigm. The other two, the number of stages and the number of items per stage, are unique to MST, although CAT can be thought of as a special case of an MST in which there are multiple stages (as many stages as there are items) and one item per stage.

In most MST research a dichotomously-scored, three parameter logistic (3PL) measurement model is used. Dichotomous scoring algorithms assume that items have either a correct or incorrect answer and do not award partial credit. A 3PL model is a member of a family of probabilistic models that assume based on certain item characteristics, or parameters, and examinees' response strings, an estimate of proficiency can be obtained. The three parameters associated with each item are the *a*-parameter, the *b*-parameter, and the *c*-parameter. The *b*-parameter is a measure of item difficulty; when there is no guessing assumed it marks the point at which the probability of obtaining a correct response is one-half. When there is guessing the probability of a correct response is somewhat greater; or viewed differently, because it is possible that the examinee may guess the correct answer the item may seem easier than it really is. The *a*-parameter is a measure of item discrimination. It is useful in more finely separating groups of examinees into different ability levels just above and below the difficulty level.

The *c*-parameter is the pseudoguessing parameter and accounts for the possibility that the examinee may guess and obtain a correct response. Polytomous models are less frequently used in investigational MST research. These measurement models are appropriate when items can be scored as partial credit items, as in essay or mathematics items. Depending on the model used for partial credit scoring, numerous parameters may be estimated for each item. There may be an *a*-parameter, which is similar to the *a*-parameter in the dichotomous case, and a number of parameters analogous to the *b*-parameter in the dichotomous model that indicate the level of difficulty associated with obtaining one category score relative to the others, going from a score of 3 to 4, for example. In the polytomous case, the number of *b*-parameters is one less than the number of categories. Most investigational MST research uses a number-right routing rule; this is important in the context of the 3PL model described above. The routing rule determines whether an examinee should be directed to an easy, moderate, or difficult module after the completion of a stage. A number-right rule bases the routing decision on the raw score at the previous stage. For example, if the total score possible at stage 1 is nine, examinees with score of zero to three may be routed to an easy module, those with scores of four to seven may be routed to a module of moderate difficulty, and those with scores of eight or nine may be routed to a difficult module. In the context of a 3PL model, however, the number right score is not a sufficient statistic for fixed length tests (Lord, 1980); the raw score alone does not provide sufficient information to determine relative proficiency levels among examinees. Any number of stages is possible in an MST, although most investigational MST research uses a two-stage model. This is

important in context of proficiency estimation because the number of adaption points can directly affect the efficiency of precision of the proficiency estimate in fixed length tests (Lord, 1980; Wainer, Kaplan & Lewis, 1992). Although better than a nonadaptive paper-and-pencil test, a two-stage test provides only two adaptive points. Investigational research of MSTs usually uses a fixed number of items per stage, but the number of items per stage can affect the ability of the panel to adapt to examinees at the extremes of the proficiency distribution (Lord, 1980, 1971; Luecht, Brumfield & Breithaupt, 2006; Luecht & Nungester, 1998).

Given the advantages that may be afforded to the test developer and test consumer through the use of the MST paradigm, it would do well to investigate the effectiveness of MST compared to CAT. Unlike CAT, for which there is a large body of methodological research, there is relatively little for MST. Indeed, over the past 35 years much of the research on MST has been devoted to developing programming algorithms (Boekooi-Timminga, 1987, 1990; Theunissen, 1985, 1986; van der Linden & Boekooi-Timminga, 1988, 1989) and operationalization (Adema, 1990; Luecht, 1998, 2000; Luecht & Hirsch, 1992; van der Linden, 1998; van der Linden & Adema, 1998; van der Linden & Luecht, 1994).

Comparative studies of different approaches within the context of MST have been fewer, and comparisons of MST to CAT have been fewer still. Comparative studies have included methodological inquiries on automated versus manual test assembly (Stocking, Swanson & Pearlman, 1993), score precision (Schnipke & Reese, 1997), the inclusion of content area constraints and the effect on score precision (Reese, Schnipke & Luebke,

1999), the inclusion of exposure control (item use) methods and score precision (Davis & Dodd, 2003), and master/nonmaster decision accuracy (Hambleton & Xing, 2006; Jodoin, Zenisky & Hambleton, 2006). Only three of these (Schnipke & Reese, 1997; Davis & Dodd, 2003; and Hambleton & Xing, 2006) include comparisons to CAT. Furthermore, none have compared the effect of using differing numbers of items per stage versus the same number of items per stage in an MST, and none have compared the effect of using a number-right versus a proficiency estimate routing rule in an MST.

This research centered on the comparison of three 3-stage MST designs to a CAT, all of which employed the generalized partial credit (GPC) measurement model. The CAT was used as the standard against which the MST designs were evaluated. The three MST designs differed according to the number of items per stage, with one design using a decreasing number of items per stage, one using an increasing number of items per stage, and one using the same number of items per stage. Each of these designs also used a number-right routing rule and a proficiency estimate routing rule to guide module selection at all stages.

CHAPTER TWO: LITERATURE REVIEW

This literature review provides background information on theoretical approaches and substantive issues that pertain to the current study. It begins with a brief history of MST designs describing the earliest conceptual designs studies to more contemporary design work. The next section discusses the assumptions and characteristics of item response theory (IRT), upon which this research is based. This leads to a description of the more commonly used IRT measurement models for dichotomously and polytomously-scored test items and the components of a computerized adaptive test. The next sections focus on conceptual and methodological issues that are particular to MST. The last section discusses the comparative methodological literature on MST. The chapter concludes with a statement of the problem that supplies the impetus for this study.

Approaches to Multi-Stage Designs

Among the earliest MST designs proposed were constant step size pyramidal models that require that the number of items at each stage be equal to the number of stages (Weiss, 1974). At stage one, one item is available; at stage ten, ten items are available. Examinees are routed through a test based on a "one up, one down" rule. If an item is answered correctly at stage one, the examinee is routed to a more difficult item at stage two. If it is answered incorrectly, a slightly less difficult item is administered. This process continues until the test is completed. At any stage, and especially at the higher stages of the test, item difficulties span the range of the ability

distribution. It is known as a constant step size pyramid because the difference in the difficulty level from one stage to another is the same. Other variations include variable step sizes, truncated pyramids, multiple item pyramids, and differential response option branching.

Lord (1971a) proposed the flexi-level test. The flexi-level test was not conceptualized as an automated test. Instead, it was designed to be completed by hand by the examinee, and it required answer sheets that informed him/her whether the answer was correct or not. Depending on the answer to an item s/he was directed to the next item. The test terminated when half of the available items, excluding the first item, were answered.

Lord (1971b, 1980) described a two-stage testing procedure in which an examinee is first administered a routing test, and based on the examinee's performance on the routing test s/he is administered one of several second-stage measurement tests. The difficulty level of the first stage routing test is designed to be at the mean ability level of the group to be tested, and there can be any number of second stage measurement tests depending on the size of the item pool. Each of these measurement tests comprise a set of items that are of equal difficulty, and each of them are designed to have different mean difficulty levels that overall span the ability range of the group. The design requires that an examinee respond to all items. Given that most conventional tests provide more information in the middle of the proficiency distribution, Lord asserted that there is an inherent advantage of two-stage testing was that the second stage test can better match the ability level of the examinee, especially at the extremes of the ability range.

Weiss (1974) proposed another MST design, the stradaptive (stratified adaptive) strategy, in which all items in an item pool are sorted into several difficulty strata ranging from very easy to very difficult. Like the flexi-level test, each stratum is assembled such that the items cluster around some average difficulty. Branching begins at different entry points given some prior knowledge about the examinee's ability, with lower ability examinees starting with easier items and higher ability examinees starting with more difficult items. Branching occurs between strata such that a correct answer at one stratum leads to the next item at the stratum with the next higher difficulty level. If an incorrect answer is given, the examinee is administered the next available item in the stratum at the next lower level of difficulty. Unlike the two-stage model, the stradaptive method permits a variable number of items to be administered to each examinee and can more easily avoid the routing errors due to measurement error in the routing test of the two-stage model. Testing terminates when the examinee has reached a ceiling stratum when s/he answers all, or almost all, of the items incorrectly.

Contemporary approaches to MST construction have been conceptualized as optimization solutions to linear programming problems, which until relatively recently have been hampered by the available speed and power of computers. Continuing advances in computer technology allowed for more complex solutions to the optimization problem. More sophisticated measurement methodologies, such as those based on item response theory (IRT), were also incorporated. Theunissen (1985) developed a binary

programming algorithm that used item and test information to construct tests.[1] As suggested by Lord (1980), Theunissen conceptualized test construction as the selection of items that hard-to-fill areas under a target information curve[2] by solving a linear programming problem that met the following objectives, 1) minimize the number of items from a bank while the test information is specified at one θ- point, 2) create a number of tests such that any one item can be present in only one test, 3) minimize the number of items from a bank while the test information is specified at several θ- points (thus achieving objectives 1 and 2 simultaneously). Boekkooi-Timminga (1987) developed a method by which binary linear programming could be used to simultaneously develop a number of parallel tests subject to a target information function and constraints on the number of items to be selected, whether certain items should be included or not. Later work (Adema 1990; Breithaupt, Arial & Veldkamp 2005; Du, Lewis & Pashley 1993; Glas 1998; Lewis & Sheehan 1990; Luecht 2000; Luecht, Brumfield & Breithaupt 2006; Luecht & Nungester 1998; Wu 2001) focused on the use of IRT to construct weakly parallel subtests for use in multi-stage designs.

Item Response Theory

Underlying item response theory (IRT) is the proposition that an examinee has a latent ability trait, θ, that can be estimated based on his/her responses to items on a test. Estimation of this latent trait is independent of the number of persons at the same ability

[1] Item information refers to the level of precision at which true proficiency can be measured. If the amount of information is large, then all estimates of proficiency will be close to the true level of proficiency (Baker, 2001). If the information is small, proficiency cannot be measured with precision and the estimates of proficiency will vary widely around the true proficiency level. Test information is the sum of all item information functions.

[2] The target information curve determines the accuracy of proficiency estimation at each proficiency level.

level, independent of the number of persons at all other ability levels, and independent of the set of test items administered. In other words, when a test comprises a set of items with item parameters calibrated from previous test administrations an individual with θ_0 will perform similarly to another person with θ_0 in a subsequent test administration. This is due to the fact that the regression of item score on ability is assumed to be invariant across groups of individuals as the item parameters themselves (Lord, 1980).

There are two important assumptions that must be met in order for this to hold true, however. The first is that the test is unidimensional, meaning that only one ability is measured by the set of items in a test (Hambleton, Swaminathan & Rogers, 1991). Thus, the probability of answering an item correctly is governed only by the item parameters and the examinee's ability. The second assumption, which follows from the first, is that of local independence. Local independence means that within any group of examinees at the same level of θ the item scores are independently distributed, and examinees at the same level of θ have the same error distributions over replications (Lord & Novick, 1968). If local independence were to be violated, then performance on some items would depend on some trait other than θ (Lord, 1980).

Item Response Theory with Dichotomous Items. Dichotomous items are those in which item responses are scored as either correct or incorrect. For dichotomous items the probability of a correct response given θ can be estimated using up to three item parameters. The *b*-parameter, or the location parameter, determines the difficulty of an item. It is defined as the θ-value at the point of inflection of the logistic function. The inflection point of the logistic curve is also the point on the underlying scale at which the probability of a

correct response is equal to one-half the distance between the lower asymptote of the function and 1.0. The more difficult the item, the farther the curve is shifted to the right of the ability distribution (usually defined as standardized units under a logistic distribution ranging from -3 to 3). The *a*-parameter, or item discrimination parameter, is proportional to the slope of the logistic curve at its inflection point. It is useful in separating groups of examinees into different ability levels just below and above the difficulty level under scrutiny. The steeper the slope, the better the item is able to discriminate examinees at levels of θ around the *b*-value. The *c*-parameter, the pseudo-guessing parameter, accounts for the possibility that an examinee may guess and obtain a correct response. It is equal to the lower asymptote of the item response function.

The logistic function for a dichotomous item is known as an item response function (IRF), a monotonically increasing function. Figure 2 shows the IRF for a dichotomous item and indicates the *a*-, *b*-, and *c*-parameters. The probability of a correct response is shown along the *y*-axis, and the estimated proficiency is shown along the *x*-axis.

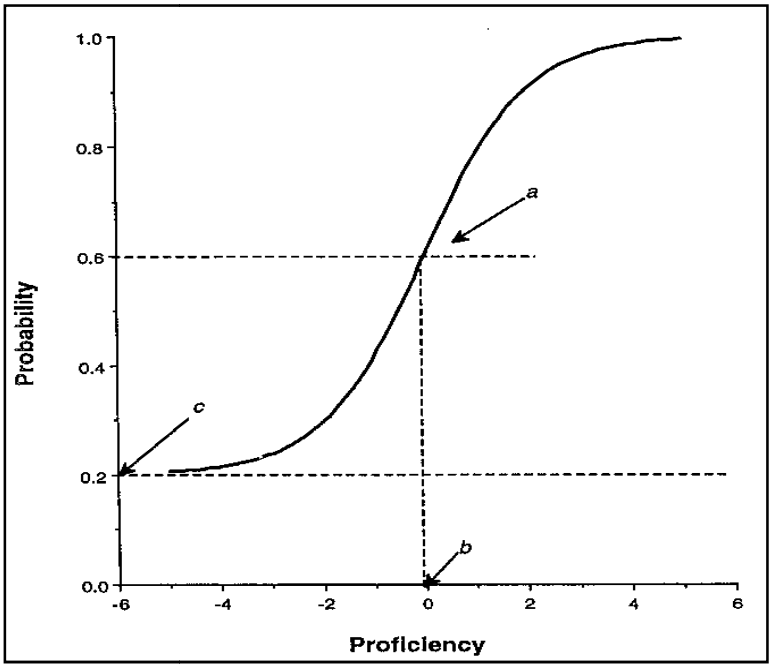

Figure 1. Item Response Function for a Dichotomously Scored Item

The most commonly used IRT models are probabilistic models based on a logistic function of item and ability parameters. The most restricted of these models is the 1-parameter logistic (1PL) model (Rasch, 1960), a function of the item and person parameters. For dichotomous items, the 1PL model is

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

where θ is the measurement of latent ability, $P_i(\theta)$ is the probability of a correct response conditional on θ, and b_i is the item difficulty. This model assumes that items vary only in the degree of difficulty; that items are equally discriminating; and that there is no guessing on the part of the examinee. Since items are equally discriminating the raw scores (or proportion-correct score) are sufficient statistics for the ability estimate

provided all examinees take the same number of items. In practical terms this means that every individual with the same raw score also has the same estimate of proficiency.

Birnbaum (1968) developed a 2-parameter logistic (2PL) model that accounts for both item difficulty and item discrimination. The 2PL model for dichotomous items is specified

$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \quad , \qquad (2)$$

where $P_i(\theta)$ and b_i are defined in the same manner as the variables described under the 1PL model and a_i is the item discrimination parameter. The *a*-parameter assumes that items have differing degrees of discrimination with some items being better at separating examinees with different levels of ability than others. The a-parameter is proportionate to the slope at the point of inflection. The sufficient statistic for the 2PL model is represented by $\Sigma a_i u_i$ where a_i represents item discrimination and u_i represents the item response.

The 3-parameter logistic (3PL) model (Birnbaum, 1968) is specified

$$P_i(\theta) = c_i + (1 - c_i)\left(\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \right) \quad , \qquad (3)$$

where $P_i(\theta)$, a_i, and b_i have been previously defined in the 2PL model. The *c*-parameter accounts for guessing by lower ability examinees. Unlike the case of the 1PL model, the raw score (or proportion-correct score) is not a sufficient statistic when items are not equally discriminating (as in the 2 and 3PL models) or when there is guessing (as in the 3PL model). This means that the final proficiency estimate is

dependent on which of the individual items are answered correctly: The same number of correctly answered items does not guarantee the same proficiency estimate.

The accuracy of the proficiency estimate resulting from these models is dependent on the item information function, which is defined as a "measure of information on the scale 'per unit separation between ability levels'" (Birnbaum, 1968, p. 449). It has the form

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \quad , \qquad (4)$$

where P_i' is the first derivative of $P_i(\theta)$ with respect to θ and $Q_i(\theta) = 1 - P_i(\theta)$. It measures how precisely an item with a given IRF contributes to the reduction of error when selecting alternative θ-values for an examinee, for example $\hat{\theta}_1$ and $\hat{\theta}_2$ (Birnbaum, 1968). More simply put, information represents the accuracy at which proficiency may be measured. Higher levels of information indicate greater precision and greater reliability, conditional on θ.

Testlet Response Theory (TRT: Wainer, Bradley & Du, 2000; Wainer & Keily, 1987; Wainer & Lewis, 1990) was suggested as an alternative to single-item measurement models to account for the local dependence that occurs when items are associated with a single stimulus, such as a reading passage. The testlet consists of the stimulus and its associated set of items. In a TRT model the unit of analysis is still a single item, and items are scored dichotomously. Since the model violates the

assumption of local independence, a random effect is added to the 3PL model to account for the violation. The TRT model is defined as

$$P_{ij}\left(x_i = 1 \middle| \theta_j\right) = c_j + \left(1 - c_j\right)\left[\frac{\exp^{\left(a_i\left(\theta_j - b_j - \gamma_{jd(i)}\right)\right)}}{1 + \exp^{\left(a_i\left(\theta_j - b_j - \gamma_{jd(i)}\right)\right)}}\right] \quad , \qquad (5)$$

where $\gamma_{id(j)}$ is the testlet effect parameter that models the dependency for person *j* on item *i* included in testlet *d(i)*.

The information function for the TRT model is given by

$$I(\theta_i) = a_j^{\,2}\left(\frac{\exp^{(t_{ij})}}{1 + \exp^{(t_{ij})}}\right)^2\left(\frac{1 - c_j}{c_{j + \exp^{(t_{ij})}}}\right) \quad , \qquad (6)$$

where $t_{ij} = a_j\left(\theta_j - b_j - \gamma_{id(j)}\right)$.

Item Response Theory with Polytomous Items. These models provide alternatives to TRT in that the groups of items associated with a given stimulus are scored as a group with the score indicating the number of items answered correctly. Polytomous models are extensions of dichotomous models and can also be used to model the probability of choosing one category over another in items that require the completion of a number of subparts or when an item, such as an essay-type item, is scored on a continuum that may reflect an inadequate response (0, say) to a completely adequate response (5). These models use multiple parameters to model the probability of a correct response. The *a*-parameter is defined as the discrimination parameter just as it is in the dichotomous case. These models include additional parameters, known depending on the model as a step difficulty, category boundary, or threshold parameter. Together these parameters model the probability of responding correctly in a given category. The relationship is specified

by a category response function (CRF) as shown in Figure 2. In this example, each of the curves models the probability of choosing one of the categories in a five response category item.

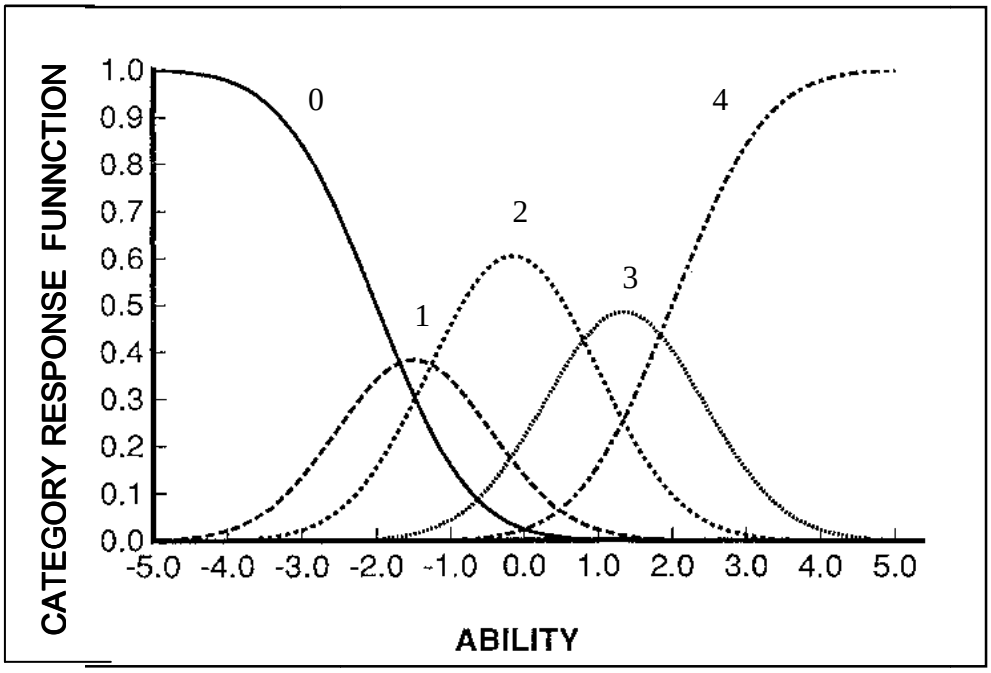

Figure 2. Category Response Function for a Polytomously Scored Item

Thissen and Steinberg (1986) categorized polytomous models into two basic types, difference models and divide-by-total models. Difference models, such as the graded response model (Samejima, 1969), model the probability of responding in a given category by subtracting the probability of responding to the given category or higher from the probability of responding to a lower category. Divide-by-total models, such as the partial credit model (Masters, 1982) and the generalized partial credit model (Muraki,

1992), calculate the probability of choosing a particular response by dividing the numerator by the sum of the numerators for all of the category the scores for that item.

The category information function (Samejima, 1969) for polytomous items can be expressed as

$$I_{ix}(\theta) = \frac{[P'_{ix}(\theta)]^2}{[P_{ix}(\theta)]^2} - \frac{P''_{ix}(\theta)}{P_{ix}(\theta)} \quad , \qquad (7)$$

where P_{ix} is the probability of obtaining category score x conditional on θ, and $P'_{ix}(\theta)$ and $P''_{ix}(\theta)$ are the first and second derivatives of P_{ix}, respectively. Samejima (1969) further defined the item information function for a polytomous item as

$$I_i(\theta) = \sum_{x=0}^{m_i} \frac{[P'_{ix}(\theta)]^2}{P_{ix}(\theta)} \quad , \qquad (8)$$

Graded Response Model

The graded response model (GRM; Samejima, 1969) is a polytomous model that deals with two or more response categories ordered in degree of the trait represented. The ordered response categories could be letter grades like A, B, C, or D; responses to Likert-type scales; or partial credit on an essay-type exam. The probability that an individual with a given θ will attain a certain category score is obtained through a two-stage process. Response categories are first artificially dichotomized (e.g., a score of 0 versus a score of 1, 2, or 3), and then the probability of responding in a given category is obtained by subtracting the probability of responding in a given category or higher from the probability of responding in an adjacent or lower category conditional on θ.

In the first stage, the probability that an individual will attain category score *x* or higher on item *i* is

$$P_{ix}^{*}(\theta) = \frac{\exp[a_i(\theta - b_{ix})]}{1 + \exp[a_i(\theta - b_{ix})]} \quad , \qquad (9)$$

where a_i is the discrimination parameter for item *i*, θ is the trait level, and b_{ix} is the category boundary for category *x* of item *i*. The b_{ix} category boundary is defined as the θ-level that corresponds to the point of inflection on the P^* function and thus requires that the m_i category boundaries to be sequentially ordered. Since the responses to item *i* are classified into *m + 1* categories, there will be m_i category boundaries.

In the second stage the probability of responding in a particular category is obtained by subtraction of cumulative probabilities for adjacent categories conditional on θ using

$$P_{ix}(\theta) = P_{ix}^{*}(\theta) - P_{i,x+1}^{*}(\theta) \quad . \qquad (10)$$

For the lowest category score, $P^*_{i0}(\theta)=1.0$ since the probability of responding in category *i0* or higher defines the entire probability space. For the highest category score, $P^*_{i,\, x=1}(\theta)=0.0$.

Partial Credit Model

Master's (1982) partial credit model (PCM) is another model used to analyze item responses with two or more ordered categories. Although the steps are sequentially ordered, subsequent steps are not necessarily more difficult than the preceding one (i.e., step 2 may be less difficult than step 1). The PCM differs from the GRM in that it

assumes all items are equally discriminating, which places it in the family of Rasch models. The probability that an examinee would score *x* on item *i* is given by

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{k=0}^{x} (\theta - b_{ik})\right]}{\sum_{h=0}^{m} \exp\left[\sum_{k=0}^{h} (\theta - b_{ik})\right]} \quad , \qquad (11)$$

where b_{ik} is the item step difficulty parameter associated with transitioning from one category to another and defined as the θ-level where the adjacent category probabilities are equal.

Generalized Partial Credit Model

Muraki (1992) extended the PCM to a generalized partial credit model (GPCM) to account for differences in discriminating power of polytomous test items. It combines some of the features of the GRM and the PCM. Like the GRM, the GPCM assumes that there may be differences in discrimination among items. Like the PCM, step difficulties are not necessarily ordered. The model assumes that the probability of choosing category *x* over category *x-1* at any given level of θ to be

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{k=0}^{x} a_i(\theta - b_{ik})\right]}{\sum_{h=0}^{m_i} \exp\left[\sum_{k=0}^{h} a_i(\theta - b_{ik})\right]} \quad , \qquad (12)$$

where $P_{ix}(\theta)$ is the probability conditional on θ of responding in category *x* of item *i* with m_i+1categories, a_i is the slope parameter, and b_{ik} is the step difficulty parameter associated with category *k* (*k*=1,…m_i). Since the model assumes that the b_{ix}, or step difficulty, parameters are not necessarily sequentially ordered, the second step in a multi-step problem may be easier than the first.

Computerized Adaptive Tests

Computerized adaptive tests (CAT) offer certain advantages over traditional paper-and-pencil or linear computerized tests. Chief among these advantages is that examinee ability can be estimated using fewer items because the items that are selected for administration closely match the examinee's estimated proficiency level. How is this accomplished? Prior to being administered in a live testing situation items are calibrated to estimate difficulty, discrimination, or pseudoguessing based on one of the appropriate measurement models. As each item is administered, proficiency is re-estimated and as it becomes more precise, items are selected only within the appropriate range of estimated proficiency. No examinee is presented an item that is too difficult or too easy allowing proficiency to be estimated in a narrower and narrower range until a tolerable level of estimation error is reached. Fewer items lead to other advantages: Examinee frustration or fatigue, and the effects of test speededness are minimized because test administration time is decreased.

A CAT refers to more than just a type of test. It is a testing paradigm best described by its major components, the item pool, item selection method, ability estimation procedure, stopping rule, content specifications, and exposure control method (Green, Bock, Humphreys, Linn, & Reckase, 1984; Reckase, 1989). Each component is described more fully in the sections that follow.

Item Pool. How appropriate it is that the batch of items from which a CAT makes a selection is known as an item pool. Like any desirable pool, it should be both wide and deep. Pool width is important because test items should cover a range of item difficulties

so that the least proficient through the most proficient examinee can be scored with precision. Pool depth is essential for item security. Since a CAT selects and administers the most informative item at the provisional ability estimate it is feasible that a few items are repeatedly selected for administration given an examinee population. If the most informative items are relatively few, and these few items are repeatedly offered in a testing program, the potential for sharing information about the item content and its correct response increases among examinees. When the potential exists for examinees to share information, the utility of the item is diminished since preknowledge of the correct response artificially inflates the proficiency estimate. Short of limiting the test taking window and the number of times a test is offered in a given time period, the only recourse to compromising item security is to include a large number of items with similar psychometric characteristics.

Another way to describe an item pool is by each item's psychometric and nonpsychometric properties. Psychometric properties include item difficulty, item discrimination, and pseudoguessing (the *a*-, *b*-, and *c*-parameters described in the previous section on IRT). Nonpsychometric item properties are content area, item format, word counts, number of response options for multiple choice items, and the sequential order in which an item might appear in a test. With the exception of item format and item sequencing, both paper-and-pencil tests and adaptive tests can manage these properties with equal efficiency. CAT makes possible the use of innovative item types that are impossible in a paper-and-pencil format, such as interactive drag-and-drop formats. Under the assumption of local independence many of the measurement models

used in CAT, with the exception of testlet response models, do not take sequencing effects into account.

Item Selection Method. The selection of the first item in a norm referenced test is usually based on the item that provides the most information at the mean of the examinee distribution or by using a Bayesian estimation algorithm. Parshall, Spray, Kalohn, and Davey (2002) describe the former as the "best guess" approach to initial item selection and the latter as the "use what you've got" approach. A third approach to initial item selection is the "start easy" approach in which relatively easy items are administered.

Once an initial item is selected continued item selection must address three, sometimes conflicting, goals: 1) Maximize test efficiency by measuring quickly and accurately as possible, 2) assure appropriate content balancing, and 3) protect the security of the item pool. Two of the more common item selection methods are maximum information and Owen's Bayesian method (Wainer, 2000). A third is maximum posterior precision (Parshall et al., 2002).

The goal of the maximum information method is to accumulate as much information about an examinee's ability as quickly as possible. Prior to the administration of a test, an information table containing all of the items ordered by the amount of information they provide at each θ-level is created. An algorithm then selects the item that provides the largest amount of information conditional on θ. A disadvantage to this approach is that estimation error may lead to the selection of items that do not match the true ability level. This is especially true early in a test and may be exacerbated when highly discriminating items are selected.

Owen's Bayesian method assumes that there is prior knowledge about an examinee's proficiency before taking a test. This prior information, operationalized as a prior density function, is used to modify the likelihood of a given response string conditional on θ. The modified likelihood function, expressed as the product of the likelihood function conditional on θ and the prior density function, is called the posterior density function. Owen's Bayesian method then applies weights obtained from an examinee's current posterior ability distribution to the tabled information values for each item at each ability level. The weighted information values are then summed. The item with the largest weighted, summed information is then selected. Owen's Bayesian method has the disadvantage that the proficiency estimate is dependent on the order in which items are administered, which has resulted in its being used less often than the maximum information method (Wainer, 2000).

The maximum posterior precision method selects items that lead to the largest decrease in the variance of the expected posterior ability distribution. Items are selected based on the entire posterior distribution rather than on a single point estimate. It has the disadvantage that the selected item may not be the most informative at the provisional ability level. Rather, the item that is selected measures well on average across the highest density region of the posterior distribution. In all methods the proficiency level is reestimated after an item has been answered, and the next item selected is based on the revised proficiency estimate.

Ability Estimation. Ability estimation can be divided into three stages, initial ability estimation, interim ability estimation, and final ability estimation (van der Linden &

Pashley, 2000). Initial ability estimation starts the item selection process. Interim ability selection guides the selection of items as the test progresses. Final ability estimation is necessary for scoring of the test and for providing examinees an assessment of their performance. Ability may be estimated using a number of techniques, among them being maximum likelihood estimation (MLE), expected a priori (EAP) estimation, maximum a posteriori (MAP) estimation and weighted likelihood estimation (WLE; Warm, 1989).

With MLE the θ value that maximizes the likelihood function of the observed responses becomes the estimate of ability. It has the drawback, though, of not finding a finite maximum value when an examinee answers all items correctly or all items incorrectly (Hambleton, Swaminathan & Rogers, 1991), which makes it unstable for short tests. In addition to the problem of ability estimation when all items are answered either correctly or incorrectly, the problem of infinite values is particularly troublesome in the early stages of a test. Since MLE cannot find a finite maximum after the first test item is administered, a variable stepsize method must be employed to estimate proficiency and to select the next item. In this method, an interim trait estimate that is equivalent to a θ-value one-half the distance between the current ability estimate and the most extreme item difficulty is assigned (Koch & Dodd, 1989). Another consideration is that MLE tends to slightly overestimate proficiency at high ability levels and slightly underestimate at low ability levels (Parshall et al., 2002). Multiple modes, or local maxima, may also be a problem in MLE.

EAP and MAP are Bayesian methods of ability estimation. EAP applies the mean of the posterior distribution of the range of true ability levels as the point estimate of

ability, and MAP applies the maximum value of the posterior distribution. Both are stable for short tests, and estimates can still be obtained when all items are answered correctly or incorrectly. It may not be possible to recover from a poor choice of a prior distribution (the unconditional probability of θ) in a short test, however. EAP and MAP tend to underestimate high ability levels and overestimate low ability levels (Parshall et al., 2002).

WLE (Warm, 1989) adjusts the maximum likelihood by a weighting function to account for bias in the MLE. In the absence of guessing (c-parameter = 0) the weight function is equal to the square root of the test information function. It has been found to be is less biased than MLE and has small variance over the entire ability range for fixed length tests (Gorin, 2005; Wang & Wang, 2001; Warm, 1989)

Stopping Rule. A CAT can be either a fixed or a variable length test. For a fixed length test, the length of the CAT is specified by the test developer or test sponsor. Once the specified number of items has been administered, the test ends. A variable length CAT is terminated by one of two stopping rules. The test can be stopped when a specified level of precision, based on the standard error of measurement, is reached or when a specified level of confidence in a pass/fail decision is reached or a maximum number of items have been administered (Bergstrom & Lunz, 1999).

Exposure Control and Content Balancing. Exposure control and content balancing procedures are subsets of item selection that deserve special mention. Both address the political realities encountered in testing. Exposure controls address the issue of test security since the most informative items may be administered to examinees more often,

potentially resulting in another examinee knowing the item and the correct response prior to taking a test. The pool is functionally limited, however; it has been shown that about 33% of the verbal portion of the SAT item pool accounts for 50% of the tests administered (Eignor, 2000). By administering some items more frequently the likelihood that pre-knowledge of a correct response increases, and the resulting proficiency estimates may be inflated. Overexposure also represents increased cost to test developers who must then develop new items to replace those that that have been compromised due to overexposure. Content balancing addresses the requirement by test sponsors and/or test consumers that the items included in a CAT reflect the table of specifications for the test. Although the selection of certain item types within a larger subject area, physical science and biological science for example, should have no effect on ability estimation (Luecht, Champlain & Nungester, 1998) content balancing assures that face validity concerns are addressed.

Exposure Control Methods

Given the security and cost concerns mentioned above, a number of strategies have been outlined to control the overexposure of some items and the underutilization of others. These methodologies can be broadly characterized by the underlying item selection approach used, either through the use of randomization or the use of a conditional selection strategy (Way, 1998). Randomization strategies choose items randomly from an optimal item set, not necessarily the most informative item. Conditional strategies constrain item selection and administration based on the frequency that an item has been administered. The first two procedures described below employ

randomization strategies. The next two employ conditional strategies. The last is a method that combines randomization and conditional strategies.

5-4-3-2-1 Procedure. This procedure, credited to McBride and Martin (1983), selects the five most informative items at the beginning of a test, and of these one is randomly selected for administration. The next item is randomly selected from the set of four most informative items given the new theta estimate; the next is randomly selected from the set of three, etc. Starting with the fifth item, item selection is based on maximum information.

Subsequent research showed that the procedure yielded high item exposure rates, high test-retest exposure rates, and high peer-to-peer exposure rates when compared to four conditional procedures, Sympson-Hetter, Davey-Parshall, the Stocking-Lewis unconditional multinomial method, and the Stocking-Lewis conditional multinomial (Chang & Ansley, 2003). The same study showed that precision of proficiency estimates was better for the 5-4-3-2-1 procedure than for the conditional procedures, however, implying that there is a price to be paid for better control of item exposure.

Although it became to be considered as an exposure control mechanism, the 5-4-3-2-1 procedure was not intended to be used as such. In fact, it was developed as a means to make comparable item selections in paper-and-pencil and CAT versions of the Armed Services Vocational Aptitude Battery (ASVAB). For the paper-and-pencil version of the ASVAB, two 30-item tests were developed. Each form was divided into six 5-item subtests. The items in each subtest were arranged in decreasing order of discrimination within a difficulty level. In the final version of the tests, item were

arranged so that items one through five represented the most discriminating items at their respective difficulty levels, item six to ten represented the next most discriminating, etc. Similar to the approach used in the CAT, the five most optimal items were selected at the beginning of the test.

Randomesque Procedure. Kingsbury and Zara (1989) suggested the randomesque procedure as an adjunct to maximum information and Bayesian item selection procedures. Rather than select the most informative item at the provisional ability estimate, the randomesque procedure selects the most informative items from a set of items ranging from two to ten and randomly selects one item from the set. The same procedure is used for each item to be administered until the test stops. The authors opined that given a deep enough item pool, few examinees even at the same trait level will be administered the same items. Subsequent research found that while the randomesque procedure yielded low maximum exposure rates, it used only about 70% of a testlet pool (Boyd, 2003).

Sympson-Hetter Procedure. The Sympson-Hetter (SH; Hetter & Sympson, 1997) procedure conditions item administration on the joint probability of item administration and item selection such that the item exposure rate falls below an upper bound. Under this procedure

$$P(A_i) = P(A_i \mid S_i)\, P(S_i) \quad \text{and} \quad (13)$$

$$P(A_i) \leq r_{max} \quad (14)$$

where $P(A_i)$ is the probability that the item is administered, $P(S_i)$ is the probability that the item is selected for administration, and r_{max} is the maximum exposure rate for an item.

Sympson-Hetter uses a series of simulations to solve for the exposure control parameters that determine whether the selected item is administered. In each iteration of the simulated CAT, the probability of administration is reestimated according to the following rule

$$P^{t+1}(A_i \mid S_i) = \begin{cases} 1 & \text{if } P^t(S_i) \leq r_{max} \\ r_{max} / P^t(S_i) & \text{if } P^t(S_i) > r_{max} \end{cases}, \qquad (15)$$

where t and $t + 1$ represent iterations of the CAT simulation. The iterations are repeated until the exposure control parameters defined in (15) above stabilize.

During test administration the most informative item for the current ability estimate is selected, and a pseudo-random number, x, from a uniform normal distribution is generated. If x is less than or equal to the exposure control parameter, the item is administered; if x is greater than the exposure control parameter, the item is not administered. Items that are selected but not administered are excluded from further use for the examinee.

The SH method can cause unexpected behavior in exposure rates for an item from one iteration to the next due to the adjustment rule. A decrease in the exposure control parameter at iteration t may increase the probability of selection, and the exposure rate, for the same item at iteration $t + 1$ (van der Linden, 2002). This in turn can result in a failure to achieve an acceptable exposure rate for some items. Furthermore, an investigation of SH with the GRM, PCM, and GPCM showed that had a higher percentage of unused items compared to two other randomization procedures (Davis, 2002).

Davey-Parshall Procedure. The Davey-Parshall procedure (DP; Davey & Parshall, 1998; Parshall, Davey & Nering, 1998) is an extension of the SH procedure that takes into account not only limiting item use but controlling test overlap rates as well. This procedure provides exposure control conditioned on the items previously presented to an examinee. Like the SH procedure a series of iterative simulations are run to determine exposure parameters. Before the simulations are run, an acceptable exposure rate is determined. An *n* by *n* exposure table is then created through the simulations. Diagonal elements represent the probability that the item is selected. Off-diagonal elements represent the probability of pairs of items being selected. Items and item pairs with lower probabilities tend to be selected more often. As an example, the conditional probability of administering a selected item given two previously administered items is calculated by

$$P(A) = e_{ii}\,[(e_{ij} + e_{ij-1})/2] \quad , \qquad (16)$$

where $P(A)$ is the probability of administration, e_{ii} is the probability of selection of the next item, e_{ij} and e_{ij-1} are the probabilities of selection of the next item given the probability of selection of the previous items.

During test administration an item is selected using the appropriate item selection procedure, and the selected item's exposure parameter is obtained. The selected item's exposure parameter, e_{ii}, and the pairwise exposure parameter from the previously administered items, e_{ij} (where j ranges from 1 to the number of items already administered) are used to find the conditional probability of administering the selected item. The mean of the set of e_{ij} values are multiplied by the e_{ii} values to determine $P(A)$,

and the selected item is administered with this probability. If the item is not administered, it is removed from the available pool of items.

The DP procedure was found to be as effective as the SH procedure in controlling exposure rates, better than SH in controlling test overlap, and similar errors in proficiency estimation (Parshall, Davey & Nering, 1998). Another study found while that DP had better exposure control than three other conditional procedures (SH, Stocking-Lewis unconditional multinomial, Stocking-Lewis conditional multinomial) and a randomization procedure (5-4-3-2-1), it had the highest standard error of measurement (Chang & Ansley, 2003).

Progressive-Restricted Procedure. The exposure control method used in this research is the progressive-restricted (PR) method (Revuelta & Ponsoda, 1998). The PR method is a combination of two exposure control methods developed by Revuelta and Ponsoda, the restricted maximum information method and the progressive method.

The PR method selects items according to the maximum information method but adds a random component to the item selection algorithm that is more influential at the beginning of a test but becomes less so as the test continues. To administer a new item the information at the ability estimated from the previous h items is computed. H is denoted as the highest information value obtained. A random value R_i from a uniform distribution is obtained. The relative serial position of the item is calculated

$$s = h/m \quad , \qquad (17)$$

where h is the number of items already administered, and m is the test length.

Each unused item is assigned a weight, and the item with the highest weight is administered according to

$$w_i = (1 - s)\, R_i + s\, I_i \quad , \qquad (18)$$

where s is the relative serial position of the item in the test, R_i is a random value drawn from a uniform distribution, and I_i is the information at the ability estimated from the previously administered items.

Revuelta and Ponsoda (1998) found that in simulation studies PR with a maximum exposure rate of 40 percent, denoted PR40, yielded lower maximum rates and lower rates of unused items than three other exposure control methods (maximum information, restricted maximum information, and Sympson-Hetter). It terms of bias and standard error, PR40 yielded estimates that were only slightly less precise than those produced by the maximum information method. In a comparison of three exposure control conditions, Boyd (2003) found that two levels of the PR procedure (.20 and30) outperformed two levels of the Sympson-Hetter procedure (.20 and30), the randomesque and the modified within-.10 logits procedures with respect to item exposure rates and item pool utilization.

Content Balancing

Broadly speaking, content balancing addresses concerns about how well the content of the items in a test reflects test specifications. More precisely, it addresses concerns over content validity. From a purely psychometric aspect, a well calibrated item bank will produce nearly exact estimates of proficiency whether all of the items measure behavioral science or biochemistry, for example. However, if the content of

some items tends to be more difficult than others (chemistry versus physics, for example) examinees of lower ability will see items that are quite different from the items that examinees of higher ability will.

To illustrate this, Luecht, Champlain, and Nungester (1998) performed a simulation study in which items were selected for a 100 item adaptive test by choosing the most informative item from one of two content areas, A and B. Items in content area A were targeted to higher scoring examinees, and those in content area B were better matched to lower scoring examinees. It was found that in terms of mean scores, standard deviations of scores, and reliability of score estimates there were virtually no differences between groups. When examining scores and item content areas under the condition of maximum information, however, it was found that 80% of the items were selected from content area B for lower scoring simulees, and 80% of the items were selected from content area A for higher scoring simulees. Neither group saw the same blend of items from the two content areas potentially calling test validity into question.

From this perspective, content balancing is no trivial consideration in the construction of an adaptive test. Two types of content balancing designs are described in the following sections: The first is the constrained CAT and the second is the weighted deviation model.

Constrained CAT. Kingsbury and Zara (1989) proposed a constrained CAT procedure that would satisfy the needs of test sponsors who want to balance the content of items administered to match test blueprints while simultaneously avoiding the overuse of any content area in a test. The procedure they outlined was 1) calculate provisional

proficiency, 2) calculate the proportion of items already administered in each content area, 3) compare this proportion to a desired, pre-specified proportion of items for each content area and identify the content area with the largest discrepancy, 4) for the content area with the largest discrepancy administer the item that provides the most information at the provisional proficiency.

In a subsequent study Kingsbury and Zara (1991) found that when compared to a CAT, a constrained CAT, and a testlet design[3], the former two yielded similar results when evaluated on the basis of mean absolute error and mean information compared to the latter model. Mean absolute error represents the average deviation of the estimated proficiency, $\hat{\theta}$, from true proficiency, θ, not accounting for positive or negative deviation from θ. Mean information represents the average level of precision at which $\hat{\theta}$ is estimated. All performed equally well when evaluated on the basis of mean bias, which is the average deviation of $\hat{\theta}$ from θ taking positive and negative deviations into account. The increase in test length for a constrained CAT to achieve the same error level as the traditional CAT was moderate (5% to 11%) compared to the testlet design (43% to 104%).

Weighted Deviations Model. Only one of the purposes of the weighted deviations model (WDM) is content balancing. Indeed, it serves a larger purpose is the context of both CAT and MST design as constraints on the psychometric properties of items as well as

[3] In this context, testlets are multi-item preconstructed modules. Testlets were formed using two items from Content Area A, two from Content Area B, one from Content Area C, and one from Content Area D. Items were assigned to testlets by identifying the peak of information for each item and assigning to the testlet the items from each content area that peaked at the lowest value of theta for the testlet. Eight 6-item testlets were adaptively selected for administration to simulees.

item selection, exposure control, and content balancing in test construction. In terms of content balancing, the WDM assumes that there is more to consider than subject matter. It also includes considerations on how many of the items included in a test meet certain upper and lower boundary constraints with respect to certain properties other than subject matter. Other item properties include format (e.g., multiple choice or completion), word counts, and number of answer categories; these, too, are subject to constraints. Other constraints with respect to item content include item overlap and item sets. Item overlap refers to situations in which a previous item might provide a clue to the correct response to a subsequent item. Item set constraints are important to ensure that items refer to a common stimulus, such as a reading passage, are selected together and are not intermixed with other items.

The model employs a linear programming approach that seeks to minimize the objective function *z* subject to

$$\sum_{i=1}^{N} x_i = n \quad , \quad (19)$$

$$\sum_{i=1}^{N} a_{ij} x_i + d_{Lj} - e_{Lj} = L_j \quad , \quad (20)$$

$$\sum^{N} a_{ij} x_i - d_{Uj} + e_{Uj} = U_j \quad , \quad (21)$$

where x_i is a binary variable that determines whether item *i* is included in the test; *j* indexes nonpsychometric constraints associated with the item; a_{ij} is a binary variable that indicates whether the item has property *j* or not; d_{Lj} are the variations from the lower bounds when the lower bounds are not met; e_{Lj} are the variations from the lower bounds when the lower bounds have been exceeded; d_{Uj} are the variations from the upper bounds when the upper bounds have been exceeded; e_{Uj} are the variations from the upper bounds

when the upper bounds are not met. The last four variables are slack variables (Stocking & Swanson, 1993; Swanson & Stocking, 1993) that accommodate the differences between the desired and obtained number of items with specific properties contained in the test. L_j and U_j are the lower and upper bounds on the number of items in the test with each of the desired properties.

When a large number of constraints are included in a test, all constraints cannot always be met. One way of dealing with this is to prioritize the importance of the constraints by weighting them. The goal of the model then becomes one of minimizing the sum of the weighted deviations from the constraints. Given the reformulated goal of minimizing the weighted deviations, the objective function now becomes

$$\sum_{i=1}^{J} w_j d_{Lj} + \sum_{j=1}^{J} w_j d_{Uj} \quad , \qquad (22)$$

where w_j is the weight assigned to constraint j.

In an adaptive setting where conformance to a target test information function (TIF) is also considered a constraint the model becomes

Minimize

$$\sum_{i=1}^{J} w_j d_{Lj} + \sum_{j=1}^{J} w_j d_{Uj} + w_\theta d_\theta \quad , \qquad (23)$$

where w_θ is the weight assigned to the TIF constraint.

Multi-Stage Tests

A multi-stage test (MST) can be viewed as an intermediate test design, with a nonadaptive, computerized test at one end of the continuum and a fully adaptive, computerized test at the other. With some variations, an MST is very much like a

computerized adaptive test (CAT) with regard to its components: It requires an item pool, item selection and ability estimation algorithms, and a stopping rule. Unlike CAT, which selects items on an item-by-item basis, an MST adaptively selects modules, or sets of items, based on a current proficiency estimate. While an MST also incorporates exposure controls and content balancing, it does so outside the actual test administration process, however. As currently operationalized in most research and operational testing programs, both of these components are incorporated during the test panel construction phase prior to actual test administration. Like a CAT, ability estimation is also adaptive during test administration. Most MSTs use a fixed stopping rule, although a variable length test is possible. An illustrative example of an MST design is shown in Figure 3. In the MST framework, sets items are grouped into modules of varying degrees of difficulty. Modules are grouped into stages with a set of modules at each level of difficulty per stage. Stages are then grouped to form a panel. The panel is the test.

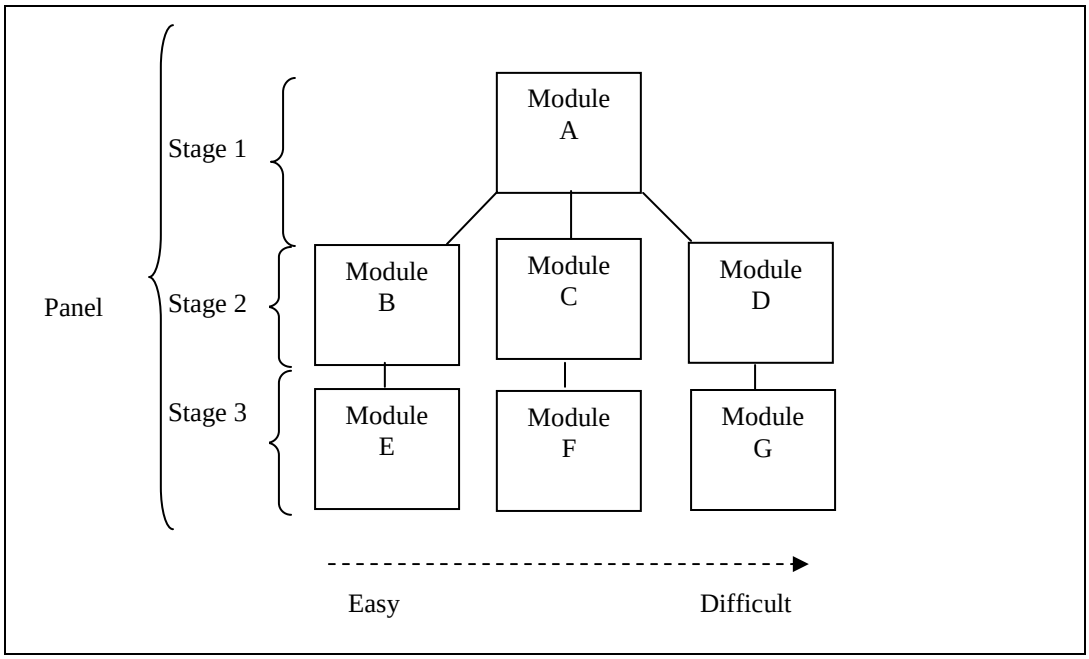

Figure 3. A 1-3-3 Panel Design

At Stage 1 Module A, usually designed to include an item set of moderate difficulty, is administered. Depending on the Stage 1 outcome, an examinee is administered either an easy, moderate, or difficult module at Stage 2. The Stage 2 outcome then determines which of the Stage 3 modules are to be administered. Given a large enough item bank, multiple panels can be generated and randomly assigned to examinees. As can be seen from this design framework, a CAT can be considered an MST with one item per module.

In terms of overall design, MST is believed to have an advantage over CAT in that the test developer can review the contents of the entire panel before it is administered. With CAT the developer can review the entire item bank, but not an entire test, prior to administration since the specific items that are administered are dependent

on the examinee's provisional ability estimate. From an examinee's vantage point, MST has an advantage because it provides the capability to skip items within a module and return to them later, and to review responses to earlier items since proficiency is estimated after the completion of an entire module. Whereas in a CAT proficiency is estimated after the completion of each item (a one-item module); skipping and/or returning to items are not allowed in most cases. Despite the capability of skipping and review that an MST provides, CAT performs better in terms of efficiency in estimating proficiency, however. Correlations between true and estimated proficiency tend to be higher for CAT than MST (Armstrong, Jones, Koppel & Pashley, 2004; Hambleton & Xing, 2006).

Multi-Stage Test Assembly

In an MST design tests are assembled through the use of automated test assembly (ATA) software. Most ATA models are based on linear programming techniques that seek to optimize an objective function subject to a set of linear constraints imposed by test developers. Constraints may include content area, word counts, item types, etc. The objective of the ATA process, then, is to build test panels by selecting items from an item pool to create modules that satisfy the constraints.

There are two approaches to designing test panels in an MST, either by a "bottom-up" or a "top-down" strategy (Luecht & Nungester, 1998). The bottom-up strategy requires module-level specifications for the target test information functions (TIF) and for all constraints. The top-down strategy only requires test-level specifications. The advantage of the former is that modules can be assembled to create

many combinations of panels; modules are not necessarily exchangeable under the latter. Most applications use a top-down approach since computer software using a bottom-up strategy has not been completed (Luecht, Brumfield & Breithaupt, 2006).

The process for constructing panels is 1) generate the target TIFs for the modules across stages, 2) partition the set of constraints across stages, and 3) create multiple panels as needed (Luecht & Nungester, 1998). Target TIFs are generated using IRT item information functions to specify the amount of measurement precision in various regions of the proficiency scale. The ATA software then selects items such that

$$T(\theta) - \sum_{i=1}^{n} I(\theta) \cong 0 \quad , \qquad (24)$$

where $T(\theta)$ is the target TIF and $\sum_{i=1}^{n} I(\theta)$ is the sum of the item information for the items selected subject to the constraints. In the top-down approach, target TIFs would be generated for each of the major pathways since it is assumed that an examinee would be administered all easy, all moderate, or all difficult modules. All modules are usually assumed to have the same content specifications (Huitzing, Veldkamp, Verschoor, 2005; Luecht & Nungester, 1998). Other constraints may be partitioned across stages.

Sample Size for Multi-Stage Test Parameter Estimation

Chuah, Drasgow, and Luecht (2006) designed a study that sought to determine the adequate sample size for parameter estimation in MST. Two measures of item parameter estimation error, correlations between estimated and true ability and decision accuracy, were evaluated. Four hundred fifty dichotomous items from an operational item bank were calibrated to fit a 3PL model based on 20,000 examinees who actually took the

exam to estimate the "true item parameters." Item responses were generated using these item parameters. Proficiency estimates were sampled from a normal distribution. Item responses were generated by first computing the probability of a correct response and then sampling uniform random numbers. If the uniform random number was less than or equal to the probability of correctly answering an item, the item was scored as correct; otherwise, it was scored as incorrect.

Four sets of item parameters based on estimates from 300, 500, 1,000, and 20,000 simulees per item were submitted to build a 3-stage MST with 20 items per module. Two MST panels were created for each set of item parameters, yielding a total of eight panels. Exposure control was operationalized by setting cut scores at each stage such that equal numbers of persons were routed through each of the modules at stages two and three. Responses were generated for 5,000 simulees in each of eight panels using the estimates calibrated from the "true" item parameters. Proficiency was estimated using MLE.

When compared to the TIF values generated by the true item parameters, TIF values for the estimated conditions were inflated. TIF values for the 300 item condition showed the greatest inflation suggesting that using small sample sizes for pretesting items may lead to overly optimistic beliefs about the precision of ability estimates.

One measure of item parameter estimation error was the correlation between estimated ability and the true simulation abilities. Correlations between true and estimated proficiency were highest for the 20,000 simulees per item condition (.965), but there was little difference among the three other simulation conditions (.955,956, and952 for the 1,000, 500, and 300 simulees per item conditions, respectively). It was concluded

that using the estimated item parameters generated from as few as 300 responses per item had little effect on ability estimation.

For licensing and credentialing purposes, the decision accuracy of classifying examinees as masters and nonmasters is of greater importance than the correlation of true and estimated ability. Decision accuracy was the highest for the 20,000 responses per item condition (94.07), and there was little difference among the three other conditions (93.73, 93.20, and 93.57 for the 1,000, 500, and 300 simulees per item conditions, respectively). Another aspect of decision accuracy, the false negative rate, was also examined. False negatives occur when true masters are classified as nonmasters based on the results of the test. The false negative rate was lowest for the 20,000 responses per item condition (2.87). The rate was similar for the 300 and 1,000 responses per item conditions (3.52 and 3.29, respectively) and highest for the 500 responses per item condition (4.12).

Multi-Stage Test Item Selection Procedures

Normalized Weighted Absolute Deviation Heuristic. Luecht (1998) proposed the normalized weighted absolute deviation heuristic (NWADH) as an approach to solving complex item selection problems. The NWADH is a series of optimization models that can be solved item-by-item or for sets of items. It endeavors to solve an optimization problem that minimizes the absolute difference between a target TIF and the test information represented by the sum of the item information functions for items with certain attributes. This objective function is normalized by dividing the coefficients for the current value of the target TIF by their sum over all unselected, eligible items in the set transforming the

absolute difference function into a proportion. These normalized coefficients are subtracted from one. Larger values indicate closer fit to the interim TIF.

Using this heuristic, four 300-item test forms were created using data from the United States Medical Licensing Examination Step 2. One form was designed to be easy, two parallel forms were designed to be moderately difficult, and one form was designed to be difficult. There were eight groups of constraints specified for a total of 3,064 constraints. Items were calibrated using a Rasch model. The four tests were assembled simultaneously, and no item overlap between tests was allowed. The NWADH satisfied all but seven of the constraints. The seven violations were due to exceeding the maximums. Item difficulties varied as expected from form to form. Standard deviations for the easiest and most difficult tests were smaller, consistent with the narrower target TIFs used for those forms. Expected number correct scores were used to confirm the differences in mean difficulties for the forms.

This study showed that using a NWADH resulted in simultaneous assembly of test forms that met most of the constraints and showed desirable properties both from an IRT and a classical test theory perspective. No attempt was made to compare the performance of NWADH with other heuristics, the weighted deviations model, for example, nor was any attempt made at comparing variations of the heuristic for the proposes of evaluation of different models, including a CAT.

ITEMSEL Heuristic. Luecht and Hirsch (1992) proposed an item selection heuristic that used an item selection composite to select test items from an item pool. A target TIF is specified, and an item with the smallest item fit selection composite is selected. The item

fit selection composite represents the least overall error weighted by information importance and is defined by

$$S_m = \sum_{k=1}^{K}[1 - w(\theta_k)]\xi_m(\theta_k) \quad \text{and} \quad (25)$$

$$\xi_m(\theta_k) = |I_m(\theta_k) - \delta(\theta_k)| \quad , \quad (26)$$

where $w(\theta_k)$ is the normalized moving average of the distance between the target TIF and an estimated TIF conditional on θ_k , $I_m(\theta_k)$ is the information of the *m*th item conditional on θ_k , and $\delta(\theta_k)$ is the moving average of the positive distance between a targeted TIF and an estimated information function conditional on θ. Luecht and Hirsch noted that this algorithm assumes that items are fairly homogeneous. When subsets of items are to be selected, the fitting process involves two stages where the set of items is first matched to a subtarget then to the overall TIF.

Six hundred items from the ACT mathematics tests were selected to evaluate the two-stage item selection heuristic as operationalized by the ITEMSEL software program. Items were calibrated using a 3PL model. Six parallel tests were assembled simultaneously. The target TIF was set based on an operational 40-item math test. A comparison of the item parameters between the operational test and the six tests assembled using ITEMSEL showed that the program tended to overfit the average *a*-parameters and to select items with somewhat higher mean *b*-parameters. Overall the program seemed to underfit slightly at the peak of the TIF and compensate at other points along the function. The results showed that the tests assembled using ITEMSEL were comparable to the operational test with respect to item parameters and actual information functions. No comparisons were made to other item selection methodologies.

Weighted Deviations Model. Stocking, Swanson and Pearlman (1993) used the weighted deviations model (WDM) to compare automated test assembly to manual test assembly using verbal and quantitative tests. Both item pools were calibrated using a 3PL model. Both were matched to target TIFs based on previous editions of the same tests. The same group of test specialists performed the manual and automated assembly operations. It took approximately two work days to manually assemble the verbal test and approximately one work day for the quantitative test. It took 10 minutes and 8.5 minutes, respectively, to assemble the test using the WDM heuristic. The resulting TIFs fell between the upper and lower bounds of the desirable range of the target TIFs. The manually assembled verbal and quantitative tests targeted slightly higher ability levels than the test that used automated assembly. The manually assembled test verbal test provided more information, while the automated assembly quantitative test provided more information.

Multi-Stage Test Exposure Control Methods

Bundled Multi-Stage Adaptive Testing. Luecht (2003) proposed a methodology, bundled multi-stage adaptive testing (BMAT), to handle exposure control in an MST. Exposure control is usually operationalized in a MST through the use of constraints or by visual examination of modules and the substitution of items with similar properties for overexposed items. In the BMAT methodology, a bin of modules with similar psychometric and nonstatistical properties are preconstructed using automated test assembly (ATA) software prior to the administration of a test. Once these modules have been constructed, module exposure is controlled through setting expected route

proportions, statistical targets, and the number of modules produced per bin. Proportional routing is accomplished by specifying the proportion of the population that will be routed to each bin. Statistical targets are met through specifying a target TIF that will be maximally informative for the proportions of the population routed through the bins. The final exposure control component is the number of unique modules per bin, with the exposure probability of module *i* given bin *j* being equal to the probability of bin *j* divided by the total number of bins. Thus for a 1-3-4-4 design, 40 modules would be necessary to keep maximum exposure at 0.1. During actual test administration, a module is randomly selected from a bin. The BMAT format eliminates the need for exposure controls while the test is running. No comparisons to other designs were made.

Preconstructed Testlets. Reese, Schnipke, and Luebke (1999) evaluated whether content constraints could be met by preconstructing testlets in a two-stage design that would achieve the precision equivalent to a paper-and-pencil version of the Law School Admission Test (LSAT). For the two-stage design, testlets were assigned to the stage 1, routing, test or stage 2, measurement, tests. Stage 2 tests were classified as low, medium, or high difficulty. Number-correct score was used to route simulees. The total test length was 25 items. The results obtained from the two-stage content balanced design were compared to simulated paper-and-pencil tests. One of the tests used 25 items, and the other used 51 items.

A target TIF was defined for stage 1 and stage 2 testlets. Item parameters were generated through simulation assuming a normal distribution for the *a*- and *b*- parameters and a uniform distribution for the *c*-parameter one testlet at a time, beginning with the *b*-

parameter. Testlets were centered on the specified mean *b*-parameter and spanned a range from 1.5 to 2.0 values from the mean. Next, the *a*- and *c*-parameters were generated. Testlet information was summed over the item information in the testlet. Information functions for each stage were then averaged and were treated as the lower bound for the target TIF. The upper bound was derived by increasing the lower bound by 22.2%, which is the difference between the upper and lower bounds of the paper-and-pencil forms of the LSAT

Proficiency values (θ) were simulated for 1,000 simulees at each level of θ from -3 to 3 in increments of 0.25, resulting in 25,000 simulees. Item parameters were taken from the logical reasoning items for the LSAT. Five-item testlets of each content type were assembled for ten testlet types for a total of 101 testlets. Each test comprised five testlets. Three testlet-type schemes were employed: One had two sets of testlet-level constraints, one had three sets of testlet-level constraints, and one had five sets. Testlet overlap or item exposure was not controlled.

An analysis of the testlet information functions for all three testlet types showed that a number of them were unacceptable in that either the minimum TIF was not met or that the peak of the information function occurred in the desired region of the scale. The simulation was based only on the two testlet-type content balancing scheme. Root mean square error (RMSE) and bias were calculated to evaluate the results of the two testlet-type scheme and the paper-and-pencil test. The RMSE indicated that the content balanced two-stage test was more precise than the paper-and-pencil tests in the middle of the ability scale. The paper-and-pencil tests were more precise at the extremes of the

ability scale. The two-stage design and the 25 item paper-and-pencil test performed similarly with respect to bias. The 51 item paper-and-pencil test was more biased at the lower extreme of the ability scale and less biased at the upper extreme. Testlet item exposure rates were highest at the low and high ability levels.

Comparisons of CAT and Multi-Stage Test Designs

This section begins by summarizing the literature on studies that compare CAT and MST designs. Each study is described in sufficient detail to highlight unique study characteristics and study findings. It ends with a summary of all the MST-related literature included in this chapter, emphasizing the commonalities and unique aspects of the studies.

Schnipke and Reese (1997) compared two two-stage tests, an MST, and two forms of a paper-and-pencil test to two CAT designs that used a 3PL model to evaluate the precision of score estimation. Items were taken from two sections of the Law School Admission Test (LSAT). Two groups of simulated test takers were generated. One group of 50,000 simulees was used to establish cutoffs for routing decisions for the two-stage and multi-stage designs. Another group of 25,000 simulees was use for the simulations of the test designs. The test consisted of twenty-five items. Modules consisting of five items apiece were created for the two-stage and multistage designs. Separate *b*-parameters were generated first for the stage one, stage two, stage three, and stage four modules. Stage one had one level of difficulty; stage two had three levels of difficulty; stage three had four levels; and stage four had five levels. After the *b*-parameters were generated, *a*- and *c*-parameters were generated.

For the two-stage designs, two modules were randomly assigned; the number-right score was used to route simulees to the next stage. Number-right cutoffs were determined by calculating the mean square error (MSE) of ability at each number-right score for simulees administered the low, medium, and high difficulty modules. The cut score was determined by the point at which the low and medium, and medium and high MSE lines crossed. The items that comprised the modules were used for the CAT, also.

A variation on the two-stage design involved rerouting simulees through the second stage to determine if they were misclassified. After completing stage one, the MSE at each number-right score was again calculated, and the same analysis was performed to determine number-right cutoffs for reclassification.

The MST was a four-stage design. As in the previous two designs, the number right-score on the previous stage was used to route simulees through subsequent stages. Number-right cutoffs were determined as they were for the two-stage designs.

Two CAT designs based on a 3PL model were also simulated: The first was based on single item selection, and the second was based on testlets. Items for the first CAT simulation were selected using maximum information. Item information was calculated at 37 θ-values from -2.25 to 2.25 in increments of 0.125. For the testlet-based design, testlets were selected based on maximum information by summing across items in the testlet. Exposure control in both designs was operationalized using the randomesque method (Kingsbury & Zara, 1989).

The paper-and-pencil designs were taken from two intact sections of the LSAT, which were designed to provide the best measurement in the middle of the ability

distribution. Responses were simulated for a 25-item section and for two sections combined for a total of 51 items.

The efficiency of the designs was evaluated by calculating root mean square error (RMSE) and bias. The single-item CAT yielded the smallest RMSE and the least bias, particularly in the tails of the proficiency distribution. The 25-item paper-and-pencil test yielded the largest RMSE and the most bias. The two-stage and multi-stage designs led to less error and less bias than the 25-item paper-and-pencil test. The two-stage, multi-stage, and testlet-based CAT had estimates that were similar to the 51-item paper-and-pencil test in terms of RMSE and bias for θ-values less than 1.5. The 51-item test was more efficient than the two-stage and multi-stage tests for θ-values greater than 1.5. The testlet-based CAT led to θ-values that had somewhat less error and bias than the 51-item paper-and-pencil test, especially in the tails of the distribution.

Davis and Dodd (2003) compared a 3-2-2 MST to a polytomous CAT using one of three exposure control methodologies, maximum information, randomesque, and within-.10 logits. The MST design had easy, moderate, and difficult modules at each stage of the test panel. Using data from a large, national examination, the data were calibrated, and responses were generated for 1,000 simulees. The MST had no nonconvergent cases, defined as a final θ estimate that was greater than or equal to 4.0 or less than or equal to -4.0 or if a maximum likelihood estimate was unsuccessful. Random item selection had the most nonconvergent cases with maximum information and within-.10 logits falling in between. Maximum information had the lowest standard error, and random item selection had the highest. The MST and within-.10 logits fell between these

two. Correlations between known and estimated θ were similarly high for the MST, maximum information, and within-.10 logits. The randomesque method produced the lowest correlations. Both the MST and the randomesque method administered all of the items in the pool. Since the tests were all fixed length, the average exposure rate did not differ among them. The MST produced the lowest standard deviation for the exposure rate and maximum information produced the highest. Within-.10 logits had the next largest percentage of items not administered (21%), and maximum information had the highest percentage (61%).

With respect to MST, the between stage routing decision was made by summing the information of passages within the modules and selecting the next module that provided the most information at the current θ estimate. The distribution of simulees across paths was skewed toward the extremes with greater than 50% of the simulees routed to the difficult passages at both stages, and nearly 20% routed to the easy modules at both stages. This result suggested that the moderate modules did not provide as much information as the easy and difficult modules did.

Hambleton and Xing (2006) compared optimal and nonoptimal MST designs, linear parallel-form test (LPFT; parallel forms of a computerized, non-adaptive test) designs and CATs based on decision accuracy and decision consistency of pass-fail decisions. Six hundred dichotomous items from an operational item bank were selected. Items were calibrated using a 3PL model. The items were assigned to five content areas and equal numbers of items were drawn from each content area. Five thousand simulees

drawn from a normal distribution were used. Passing scores were set at -0.5, 0.0, and 0.5 to correspond to passing rates of 70%, 50%, and 30%, respectively.

Five nonoverlapping 60-item LFPTs were assembled and centered at each of the three passing scores. Each of the five forms was randomly assigned to 1,000 examinees maintaining an exposure rate of 20%.

A three-stage MST was used with five forms of a 20-item first stage module, two versions of a 3-module second stage with 20 items, and two versions of a 3-module third stage with 20 items in each module. When the MST TIF was centered at 0.0, the design was termed optimal because the mean of the proficiency distribution was 0.0. It would also be optimal if the passing scores were set at 0.0. It would be a nonoptimal design if the passing scores were set at -0.5 or 0.5.

Two routing strategies were used. When the MST TIF was centered at 0.0, cut scores were determined to route approximately equal proportions of candidates to each second and third stage module. When the TIF was centered at 0.5, examinees with proficiencies near 0.5 ± 2 standard errors were assigned the middle difficulty module. Examinees below this were assigned the easy modules; those above were assigned the difficult modules.

A 60-item fixed length CAT was constructed with content balancing. Maximum likelihood estimation (MLE) was used to select items subject to content constraints. Conditional item exposure was held to 30%, and the overall item exposure level was 20%. No further information about the exposure control method was provided.

Decision consistency and decision accuracy were used to evaluate the results. For the LPFT designs, decisions were optimal when the TIF matched the mean of the proficiency distribution. Decisions were less optimal when the TIF was matched to the passing score. CAT had the highest percentage of correct and consistent decisions. For all designs the results were poorest when the passing rate was 70%, the mean of the proficiency distribution was 0.0, and the TIF centered at 0.5. The authors noted that this scenario is not uncommon in practice. MST designs produced better results than the LPFT when the TIF was matched to the passing score.

Correlations between true and estimated proficiency showed that CAT performed best followed by MST and LPFT. MST recovered the true scores better than the LPFT regardless of placement of the TIF so that if a decision is made to match a TIF to a passing score, MST is the preferred design.

The authors further concluded that matching the TIF in LPFT to the mean of the proficiency distribution would result in improved precision for failing examinees that could lead to improvements in diagnostic reporting. If the decision is made to match the TIF to the passing score, thus optimizing the proficiency estimates around the region of the passing score, an MST design would further allow for better use of the item bank.

Jodoin, Zenisky, and Hambleton (2006) compared four operational 60-item examinations to 1) three new linear fixed length tests (LFT; a computerized, non-adaptive test), 2) two 60-item three-stage tests, and 3) a 40-item two-stage test. Sixty dichotomously scored multiple choice items were classified into three content areas.

Items were calibrated using a 3PL model to develop an item bank that consisted of 238 items.

In the first three-stage design, target TIFs were established using the mean test information from the operational forms. TIFs for the medium difficulty modules at stages one, two, and three were set to one-third the values of the mean test information. Easy modules at stages two and three were set to one-third of the mean test information with a negative horizontal shift of one-half standard deviation. Hard modules at stages two and three were set to one-third of the mean test information with a positive horizontal shift of one-half standard deviation.

Exposure controls were operationalized by constructing three medium difficulty stage one modules to create three MST panels. Each panel consisted of one of three unique stage one panels and the same six modules at the second and third stages, resulting in nine 20-item modules.

In the second three-stage design, stage one TIFs were reduced to one-quarter, and stages two and three TIFs were increased to three-eighths of the mean test information to put more discriminating items at the later stages. Easy and hard modules were shifted one-half standard deviation to the left or right. Three 20-item stage one and six 20-item stages two and three modules were created.

The two two-stage designs were created by dropping the stage 3 modules resulting in a 40 item panel. This was added to the original study design because preliminary analysis showed that there were only modest increases in the recovery of true proficiency between the second and third stages.

Responses for 5,000 simulees from a normal distribution were generated, and the same examinee responses were used for each test form. There were two replications per panel. MLE was used to compute ability estimates. Simulees were assigned to the easy and hard modules if ability estimates were less than -0.43 or greater than 0.43, respectively.

Accuracy of ability estimates were evaluated by correlating true and final ability estimates, test-retest reliability for the two replications, and alternate forms reliability. Since this study was designed in a certification examination context, levels of decision accuracy and decision consistency were evaluated assuming pass rates of 30%, 40%, and 50%.

Correlations between true and estimated ability were highest across the operational tests, the LFTs, and the three-stage MSTs. The two-stage MST design had somewhat lower correlations. No notable differences were obtained between designs where the TIFs were targeted at one-third of the mean information at all stages and designs where TIFs were targeted at one-fourth of the mean information at stage one and three-eighths at stages two and three.

Test-retest and alternate forms reliabilities were similar for the LFT and the three-stage MST. Reliabilities were somewhat less for the two-stage MSTs. No notable differences were obtained for the two TIF designs.

Decision accuracy exceeded 90% on the operational test, LFT, and 3-stage MST. The MST had slightly higher decision accuracy than the LFT and slightly lower decision accuracy than the operational form. The 2-stage MST had the lowest decision accuracy.

Decision accuracy was highest for the operational and LFT forms. It was somewhat lower for the 3-stage MST and lowest for the 2-stage MST. It was notable that decision accuracy was somewhat higher when more discriminating items were used in stage 3 than when items of equal discrimination were at all stages in the 3-stage MST design, and decision accuracy was somewhat less under the same design in the 2-stage MST.

Table 1 summarizes the studies included in this chapter by salient study characteristics. The first four studies are included in this review do not focus on MST designs, per se, but they are seminal studies in the development of optimization heuristics and methods for building the target TIFs necessary to an MST design. Luecht (2003) endeavors to build exposure controls into an MST, but the purpose of this study is to explore the feasibility of incorporating exposure controls into the optimization algorithm not to explore variations in methods. It is only with the last six studies, once many of the developmental issues have been resolved, that methodological concerns are investigated. Common to five of the studies was the use of dichotomous items, parameter estimation using a 3PL model, and the same number of items per stage. Three of the studies (Reese, Schnipke & Luebke, 1999; Schnipke & Reese, 1997; Chuah, Drasgow & Luecht 2006) used the number-right score to determine the cut points for routing to later modules. Lord (1980, 1971), however, cautioned number-right is only a sufficient statistic for Rasch (1PL) fixed length tests. Routing rules in the remaining two studies (Jodoin, Zenisky & Hambleton, 2006; Hambleton and Xing, 2006) routed equal proportions to easy, moderate, and difficult modules. The latter of the two also used set levels of θ to determine cut points under the nonoptimal MST condition. Only one of the six (Davis &

Dodd, 2003) studied polytomous items, used θ for the routing rule, and varied the number of items per stage. This study did not, however, investigate fewer items at earlier stages to more items at earlier stages on the precision of the proficiency estimate.

Table 1. Summary of Current Literature Review Studies

Author (Date)	Study Purpose	Measurement Model	Comparisons to Other Methodologies	Variations on Same Design	Routing Rule	Number of Stages (MST)	Items by Stage (MST)
Chuah, Drasgow & Luecht (2006)	Necessary Sample Size, Parameter Estimation	3PL	No	No	Number-right	2	5, all stages
Luecht (1998)	Automated Item Selection	1PL	No	4 different test forms	N/A	1	N/A
Luecht (2003)	Exposure Control & Score Precision	3PL	No	No	N/A	4	Not specified
Reese, Schnipke & Luebke (1999)	Content Constraints & Score Precision	3PL	MST & paper-&-pencil	3 sets of content constraints	Number-right	2	5, both stages
Luecht (2003)	Exposure Control & Score Precision	3PL	No	No	N/A	4	Not specified
Reese, Schnipke & Luebke (1999)	Content Constraints & Score Precision	3PL	MST & paper-&-pencil	3 sets of content constraints	Number-right	2	5, both stages
Schnipke & Reese (1997)	Precision of Score Estimation	3PL	MST, paper-&-pencil, CAT	2, 3 & 5 sets of constraints	Number-right	2 & 4	5, all stages
Davis & Dodd (2003)	Precision of Score Estimation, Exposure Control, Item Pool Usage	Partial Credit Model	MST & CAT with 3 exposure control methods	8 MST panels	θ	3	6 to 10-item passages, 1st stage; 6 to 7-item passages, 2nd & 3rd stages

Table 1, con't. Summary of Current Literature Review Studies

Author (Date)	Study Purpose	Measurement Model	Comparisons to Other Methodologies	Variations on the Same Design	Ability Distribution	Routing Rule	Items per Stage (MSTs)
Hambleton & Xing (2006)	Decision Accuracy & Decision Consis-tency	3PL	Yes, linear parallel forms, MST, CAT	Yes, five linear parallel forms	Normal	Equal proportions or $\theta \pm 2$ se	Twenty at all stages
Jodoin, Zenisky & Hambleton (2006)	Decision Accuracy and Score Precision	3PL	Yes, linear fixed tests, MSTs	Yes, three linear fixed tests and two MSTs	Normal	Proportion-al	Twenty at all stages

Statement of Problem

Traditional CAT procedures using maximum information item selection provide the most efficient estimates of proficiency; however, the method results in overuse of the most informative items, compromising test security, and does not guarantee that the items administered will exactly match test specifications. Exposure control and content balancing procedures can address these concerns, but at a cost to efficiency and precision of proficiency estimation because the most informative items are not always administered. For some test sponsors face validity is a concern since items are selected on an item-by-item basis in a traditional CAT, and it is not possible for the test developer or test sponsor to preview an entire test, only the items that make up the test. An MST addresses this concern because panels can be constructed and previewed by key stakeholders in advance, but this is also likely to be at the cost of efficiency and precision of proficiency estimation.

Given that the concerns of item overexposure, matching test specifications, the ability to preview an entire test are legitimate concerns on the part of test developers and sponsors, it behooves researchers to examine designs that respond to the concerns and yet attain efficiency and precision that is close to the traditional CAT. MSTs have been offered as a viable solution, but the majority of the literature has focused on developing heuristics for the implementation and algorithms for operationalization rather than evaluating alternative designs.

Using CAT as a baseline, this dissertation will seek to answer the following questions:

1. How do MST designs compare to CAT in the recovery of proficiency estimates, item pool utilization, and item exposure assuming a normal proficiency distribution?
2. To what extent are proficiency estimates affected by a number-right routing rule versus maximum information routing rule in an MST? MST designs typically use a number-right score as the routing rule for moving to different modules across stages. However, the number-right routing rule is a sufficient statistic only when a Rasch model is used for a fixed length test, not when item discrimination parameters differ.
3. Which is the more optimal design for an MST, the same number of items per stage or varying numbers of items per stage? Most operational MST tests use the same number of items per stage, but the number of items per stage can affect the ability of the panel to adapt to examinees at the extremes of the proficiency distribution (Lord, 1980, 1971; Luecht, Brumfield & Breithaupt, 2006; Luecht & Nungester, 1998).

CHAPTER THREE: METHOD

Overview

Three MST designs were evaluated in comparison to a CAT. For the MSTs, the routing rules, the algorithms used to determine the selection of a module at the next stage, and the number of items per stage was examined in the context of score estimation. The routing rules used were maximum information, fixed θ, and number-right. Thus, a total of ten conditions will be studied, a 3x3 MST condition and a CAT condition.

The measurement model was the generalized partial credit (GPC) model. The CAT and the three MSTs were built assuming a normal proficiency distribution. The CAT system used maximum information item selection with content balancing (Kingsbury & Zara, 1989) and exposure control (Revuelta & Ponsoda, 1998). The MSTs were assembled based on the step difficulties of the items and were separated into easy, moderate, or difficult modules stratified by the *a*-parameter. Maximum likelihood estimation (MLE) was used for the interim and final ability estimates. Prior to MLE, the variable step size method was used to estimate proficiency. Until there was at least one correct and one incorrect response, a proficiency estimate equal to one-half the distance between the current proficiency estimate and the highest *b*-value for items in the cell category was assigned. The CATs and two of the MST designs will be 20 items in length. The third MST design, with equal numbers of items per stage, will be 18 items in length.

Item Pool. The item pool for this research was based on a large, national testing program. The item pool consists of three, four, or five-category items for a total pool size of 157 items (Davis, 2002). The items cover three content areas. Based on content area, 39% represent content area I, 37.5% represent content area II, and 23.5% represent content area III. Sixty-three percent are 3-category items, 18.5% are 4-category items, and 18.5% are 5-category items. Table 2 provides the breakdown of the items by content area and category length for the item pool. Table 3 shows the joint probability of the items by content area and number of response categories.

Table 2. Classification of Items by Content Area and Number of Response Categories

	Content Area I (Row Percent) (Column Percent)	Content Area II (Row Percent) (Column Percent)	Content Area III (Row Percent) (Column Percent)	Total (Percent)
3 Categories	42 (0.42) (0.69)	42 (0.42) (0.71)	15 (0.15) (0.41)	99 (0.63)
4 Categories	10 (0.34) (0.16)	6 (0.21) (0.10)	13 (0.45) (0.35)	29 (0.18)
5 Categories	9 (0.31) (0.15)	11 (0.38) (0.19)	9 (0.31) (0.24)	29 (0.18)
Total (Percent)	61 (0.39)	59 (0.38)	37 (0.24)	157 (100.0)

Table 3. Joint Probabilities by Content Area and Number of Response Categories

	Content Area I	Content Area II	Content Area III
3 Categories	0.25	0.24	0.15
4 Categories	0.07	0.07	0.04
5 Categories	0.07	0.07	0.04

Parameter Estimation. The estimated parameters for the GPC model were obtained from the Davis (2002) study. In that study, the data were calibrated using PARSCALE (Muraki & Bock, 1993) for the generalized partial credit model. PARSCALE uses a two step marginal maximum likelihood EM algorithm to estimate the item parameters. The two step process is iterative until the parameters stabilize. Provisional expected frequency and sample size are calculated in the first step, and marginal maximum likelihood is estimated in the second step. For each item the number of step difficulties plus one is equal to the number of categories associated with the item.

Data Generation. Data were generated using IRTGEN SAS (Whittaker, Fitzpatrick, Williams & Dodd, 2003) for ten samples of 1,000 simulees. A random number was drawn from a normal distribution (0,1) to represent the known trait level of the simulee. Based on the item parameter estimates and the simulee's known θ-value, the probability of responding in each category was estimated. The category probabilities for an item were then summed to create cumulative subtotal probabilities for each response category. A random number was selected from a uniform distribution that ranges from 0 to 1 and compared to the cumulative response probabilities. If the random number was less than the subtotal probability for a given category, the simulee's response was that category score.

CAT Simulations. A SAS computer program developed by Chen (1997) was used to simulate a CAT according to the GPC model. An initial trait level was set to zero using MLE for proficiency estimation. Content balancing (Kingsbury & Zara, 1989) was used to select the items from the content areas based on the proportions in the original item

pool. The progressive-restricted procedure (Revuelta & Ponsoda, 1998) was used to maintain an item exposure rate of thirty percent. The test ended after 20 items had been administered.

For the administration of the first item, the type of content and the number of items was randomly selected for each simulee. The remaining items were selected using the specified content balancing subject to the exposure control procedure.

MST Simulations. The MST simulations were built by hand without the use of automated test assembly software. The MST panels were assembled from the 157 items into a 1-3-3 panel design with a total test length of 20 items in two of the MST designs (the third design consisted of 18 items). At the first stage, one module represented items of moderate difficulty, and the second and third stages there was one module at the easy, moderate, and difficult levels. Decisions on where items fall into the easy to difficult continuum were made through examination of the step difficulty parameters. Items with step difficulties that were predominantly negative were classified as easy; those with step difficulties that have a mix of negative and positive step difficulties were classified as medium; and items with predominantly positive step difficulties were classified as difficult. Within each of these module classifications, items were assigned so that there is a mix of low to high item discrimination parameters. The *a*-parameters ranged from 0.54 to 1.52. Cuts for the range of discrimination were based on tertile rankings, with items with discrimination parameters in the first tertile ranked as low and items in the third tertile ranked as high. Table 4 shows the items by content area, tertile ranking, and item

difficulty. Module-level information functions was generated and examined to ensure that the modules will have similar levels of information.

Table 4. Classification of Study Items by Discrimination and Difficulty

	CONTENT AREA I								
Discrimination	Easy Items								
Low	V0024800	V0015600	V0020900	V0008100	V0002400				
Medium	V0007700	V0026800	V0020500	V0001500					
High	V0026100	V0015200	V1044200	V0007300	V0020100	V0024700			
	Moderate Items								
Low	V0011800	V0019500	V1300900	V0014800	V1012400				
Medium	V1001600	V1229900	V1077600	V1074800	V1297800	V1162100	V1012300	V0015300	V1144000
High	V1022400	V1099800	V1137400	V0016100	V1021500	V0021000	V0014800	V1300000	
	Difficult Items								
Low	V1297900	V1012500	V1018800	V1344800	V1184000	V1160400	V0015400	V1301200	V1183800
Medium	V1103200	V1144900	V1098900	V1015600	V1158500	V0025600	V0021000	V1420200	V1300700
High	V1099100	V1300800	V1100900	V1163100	V1302900	V1132200			
	CONTENT AREA II								
	Easy Items								
Low	V1295300	V0027500	V0023700	V0008200	V1010100				
Medium	V1008400	V1299200	V0011400	V0002300					
High	V0027000	V1298700	V0027300	V0026400	V0011300	V0007900			
	Moderate Items								
Low	V0016900	V1400500	V0021300	V1078100	V1122400				
Medium	V1296000	V1127000	V1001500	V1013000					
High	V1296000	V1127000	V1001500	V1013000	V0025300	V1100400	V1377500	V1328400	
	Difficult Items								
Low	V1352700	V1340300	V1410000	V1008200	V1184100	V1352700	V1340300	V1410000	V1008200
Medium	V1122500	V1044400	V1329800	V1229100	V1195300	V1344900	V1159400	V1077100	
High	V1058400	V1137500	V0028300	V1008200	V1058400	V1077500			

Table 4, con't. Classification of Study Items by Discrimination and Difficulty

	CONTENT AREA III
Discrimination	Easy Items
Low	V0013500 V0010300 V1070500 V0009900 V1008300 V1076000
Medium	V0025000 V1101400 V1296800 V0020000 V0011700
High	V1352400 V1096800 V0009900
	Moderate Items
Low	V1022800 V0018000 V0002200 V0023200
Medium	V1402300 V1342600 V0013700
High	V1195200 V1298900 V1011100 V0017400 V1334700
	Difficult Items
Low	V1313600 V1013300 V1076900 V1299700 V1066900 V1367200 V0013200
Medium	V1184600 V0007600 V1302500
High	V1299700

By module, content balancing was achieved by manual assignment of item to content type. In order to maintain parallelism with the original item pool, items were selected according to the joint probabilities shown in Table 3 previously.

Figure 4 shows the design for the MST condition in which the number of items per stage decreases. Along with content considerations, the number of categories was considered. In the first-stage module 10 items of moderate difficulty were available. Five 3-category items were taken from the content area I and five 3-category items were from content area II. Three second-stage modules, easy to difficult, contained items from each of the content areas. There were three 3-category, one 4-category, and one 5-category items from content area I, one 4-category item from content area II, and one 4-category item from content area III. Three third-stage modules ranging from easy to difficult contained two 5-category items from content area II and one 5-category item from content area III.

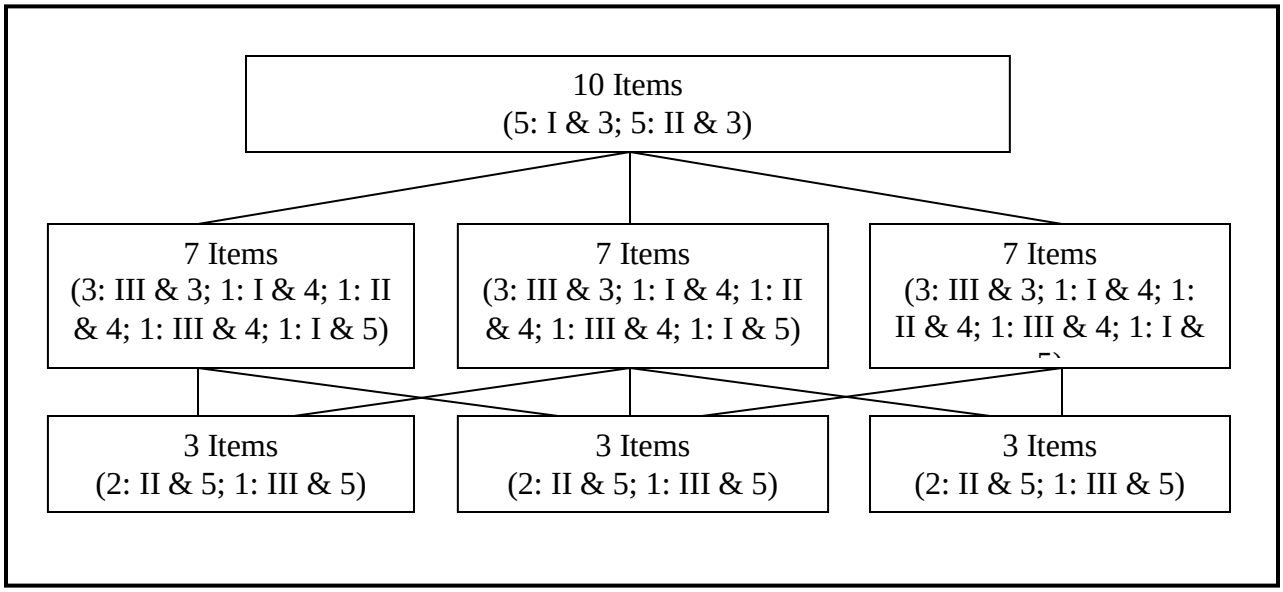

Figure 4. A 1-3-3 Panel Design with More Items at the First Stage

Figure 5 shows the design for the MST condition in which the number of items increases from stage to stage. As in the previous design, the first-stage module was be of moderate difficulty, and the second and third stage modules ranged from easy to difficult. The first stage contained one 3-category item from each content area. At the second stage there were two 3-category items from content area I, two 3-category items from the content area II, and one 3-category items from content area III. At the third stage there were two 3-category, one 4-category, and one 5-category item from content area I; two 3-category, one 4-category, and two 5-category item from content area II; and one 3-category, one 4-category, and one 5-category item from content area III.

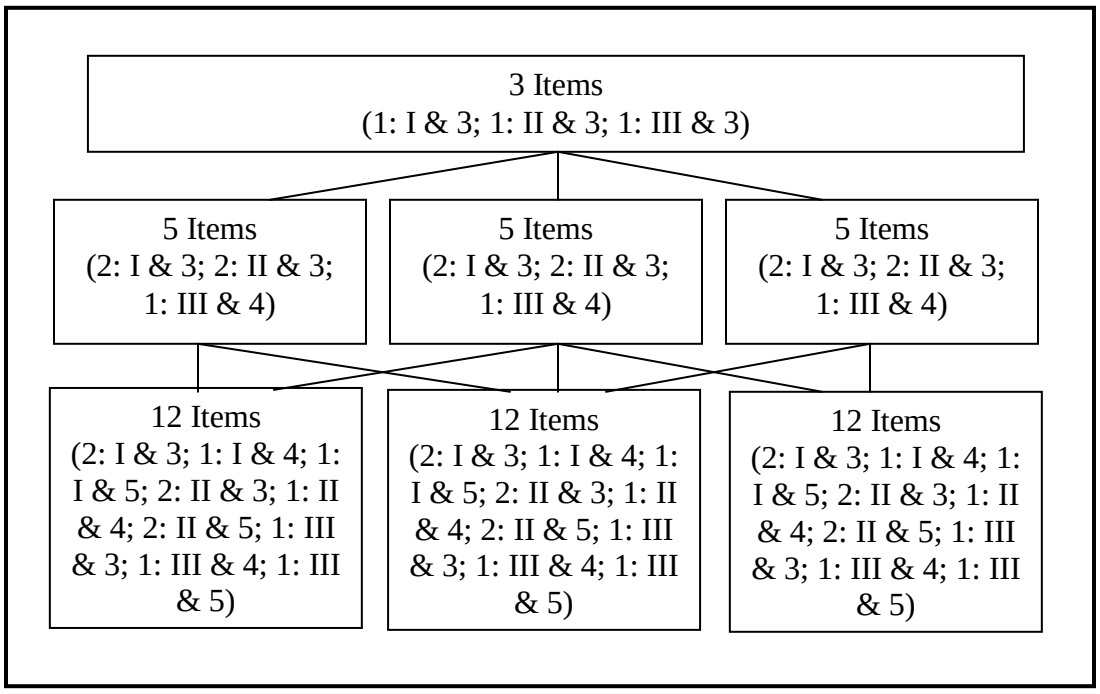

Figure 5. A 1-3-3 Panel Design with More Items at the Third Stage

Figure 6 shows a 1-3-3 panel design with equal numbers of items at each stage. At the first stage there were three 3-category items from content area I and three 3-category items from content area II. At the second stage there was two 3-category and one 4-category items from content area I, one 4-category item from content area II, and two 3-category items from content area III. At the fourth stage there was one 3-category item from content area II, one 3-category item and one 4-category item from content area III, and one 5-category item from each of the three content areas.

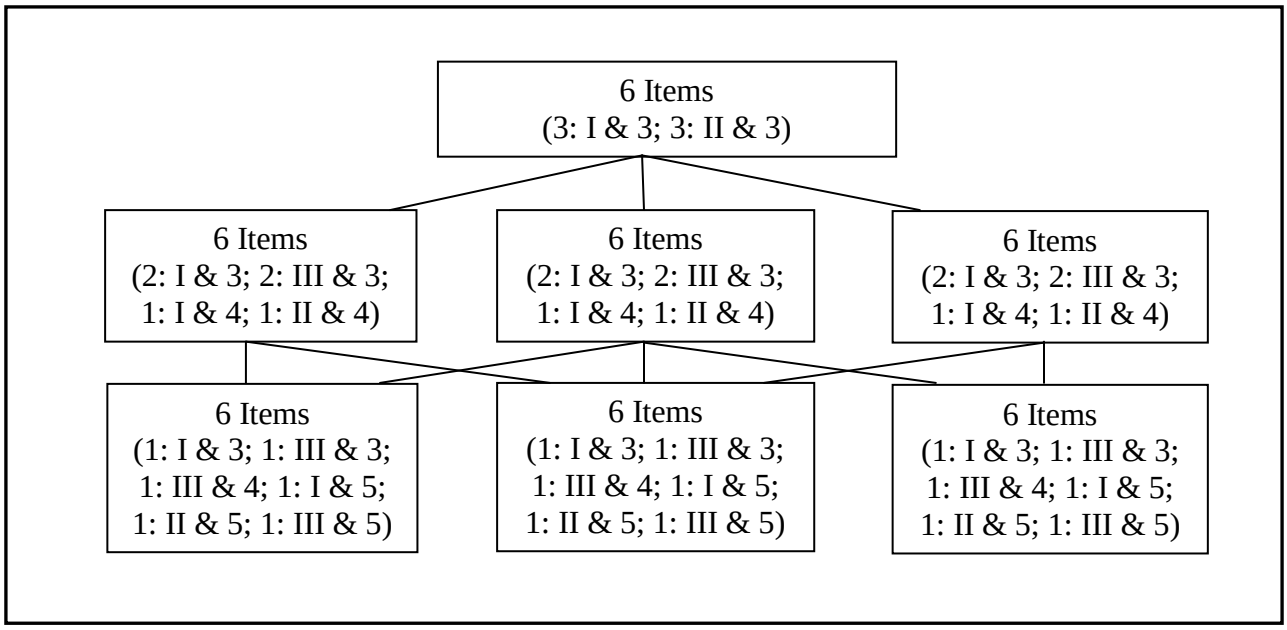

Figure 6. A 1-3-3 Panel Design with Equal Numbers of Items per Stage

Three routing rules were employed. In the maximum information routing rule condition, simulees were routed to the next stage module that provided the most information at the current θ estimate. In the fixed θ condition, simulees were routed to the easy modules if the current θ estimate was less than -1.0, to the moderate modules if the estimate was between -1.0 and 1.0, and to the difficult modules if the current estimate was greater than 1.0. Simulees were routed by number-right score under the third routing-rule. If the current score was in the lowest third of the score distribution, they were routed to the easy modules. If it was in the middle third of the score distribution, they were routed to the moderate modules, and they were routed to the difficult modules if the score was in the upper third.

Three panels were designed, one for each of the item by stage conditions. At the outset of the test, all simulees were administered a module of moderate difficulty. After administering the module, proficiency was estimated using MLE. In the maximum information routing condition, simulees were routed to an easy, moderate, or difficult

module by selecting the module that provides the most information at the current proficiency estimate. Depending on which module provided the most information cross-module routing was possible at the second and third stages. Cross-module routing was only possible at the next greater (or lesser) level of module difficulty, so that a simulee may have been routed to a moderate module after having completed an easy module or may have been routed to a moderate module after having completed a difficult module. Routing was not possible between easy and difficult modules at subsequent stages. In the fixed θ condition, simulees were routed based on whether the current θ estimate was above or below a θ-level threshold. In the number-right routing condition simulees were routed to each of the modules based on a split of the raw score distribution into thirds.

Data Analysis. Recovery of known trait estimates was evaluated using Pearson product moment correlations. Other methods to evaluate recovery of known traits included bias, root mean square error (RMSE), and average absolute differences (AAD). These statistics were calculated over each of the ten replications and averaged. The equations to compute these statistics for each replication are as follows:

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

$$RMSE = \left(\frac{\sum_{k=1}^{n} (\hat{\theta}_k - \theta_k)^2}{n} \right)^{½} \quad , \qquad (2)$$

$$AAD = \frac{\sum_{i=1}^{n} |(\hat{\theta}_k - \theta_k)|}{n} \quad , \qquad (3)$$

where $\hat{\theta}_k$ the estimated trait level for simulee k, θk is is the known trait level, $\bar{\hat{\theta}}_k$ is the mean of the estimated trait, and $\bar{\theta}_k$ is the mean for the known trait.

In addition to the statistics described above, tables showing the item-and module-level parameter estimates, the means and standard deviations for the estimated θ-values, and exposure rates of items under each of the designs are presented. These statistics were averaged over the ten replications, as well. Graphs of test and module-level information functions are presented for each of the MST test panels. Plots of the difference between known and estimated θ-values conditional on known θ for each replication and plots of the standard errors conditional on known θ for all replications under each of the designs are also presented.

CHAPTER 4: RESULTS

Item Pool Construction. Construction of a suitable item pool was more troublesome for CAT than for the MST designs. The imposition of the PR30 exposure control method forced the original 157 item pool to be increased to 208 to satisfy the item exposure limit of 30%. In particular, the number of content area II, 4 and 5-category items, and all content area III items were required to be increased. The same 208 item pool was used for all test designs. Though not necessary to meet specifications for the three MST panels, the MST designs benefited from this increase, as well. The appendix includes the list of item parameters for the 208 items. Figure 7 shows the information function for the 208 item pool.

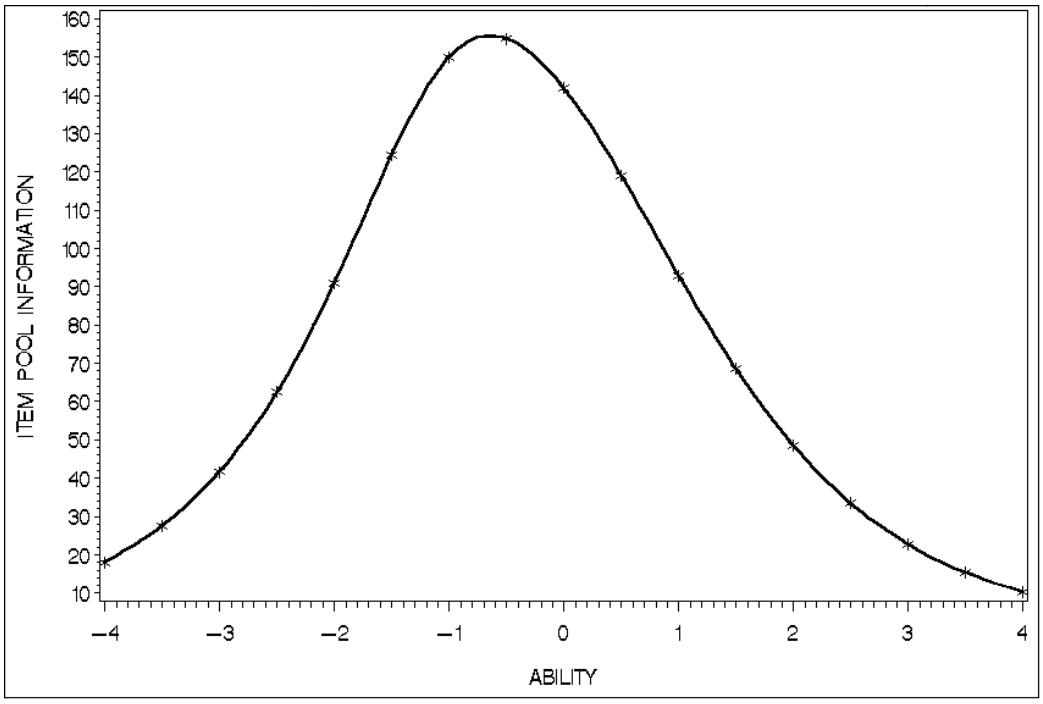

Figure 7. Item Pool Information Function

MST Construction. As stated previously, the MST design tests were assembled by hand. This was accomplished by selecting items that met design specifications by module and stage and then generating information functions to ensure that each module within a stage contained similar amounts of information. Figures 8a-10c show the information functions at each stage for the MST designs. At stage 1, the information function is centered on 0, reflecting a module designed to reflect a moderate level of difficulty. The stage 2 and stage 3 information functions show that similar amounts of information are provided at the easy, moderate, and difficult modules shown from left to right. The functions for the easy and difficult modules at stage 2 and stage 3 also provide more information at the extreme that corresponds to the targeted difficulty level. The function for the easy module provides the most information at the lower end of the θ-scale, while the function for the difficult module provides the most information at the upper end of the θ-scale.

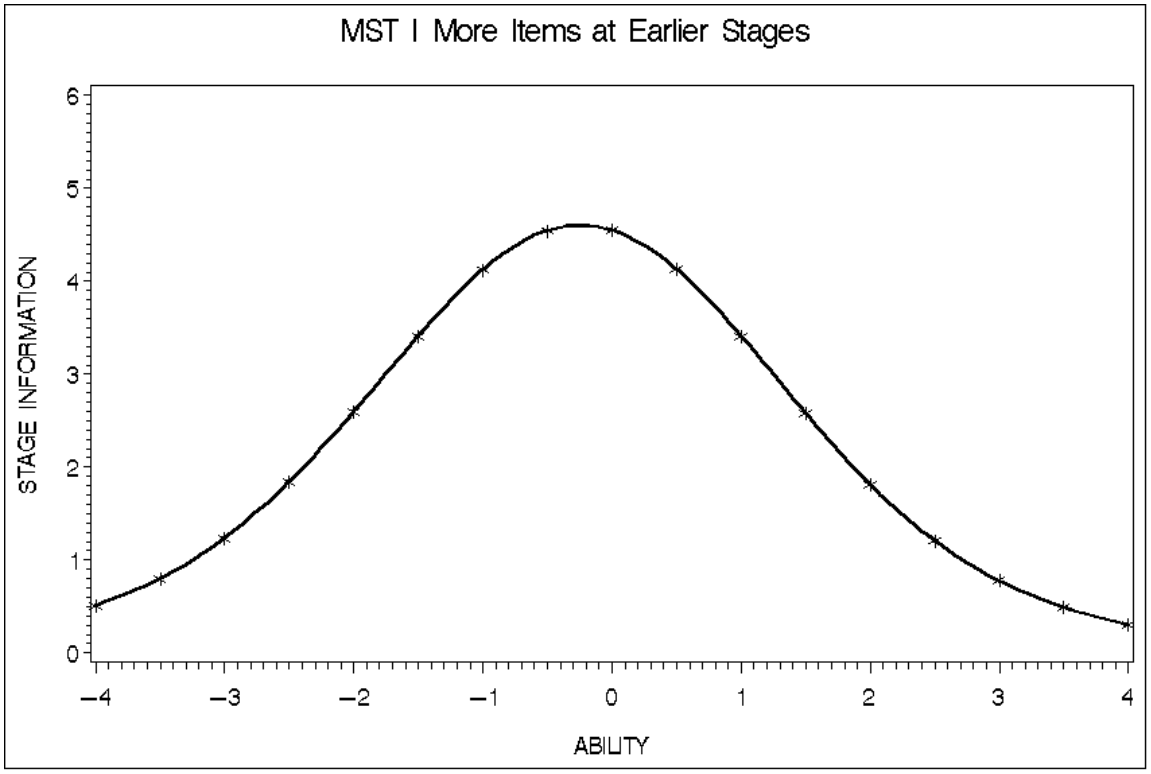

Figure 8a. MST I, Stage 1 Information Function

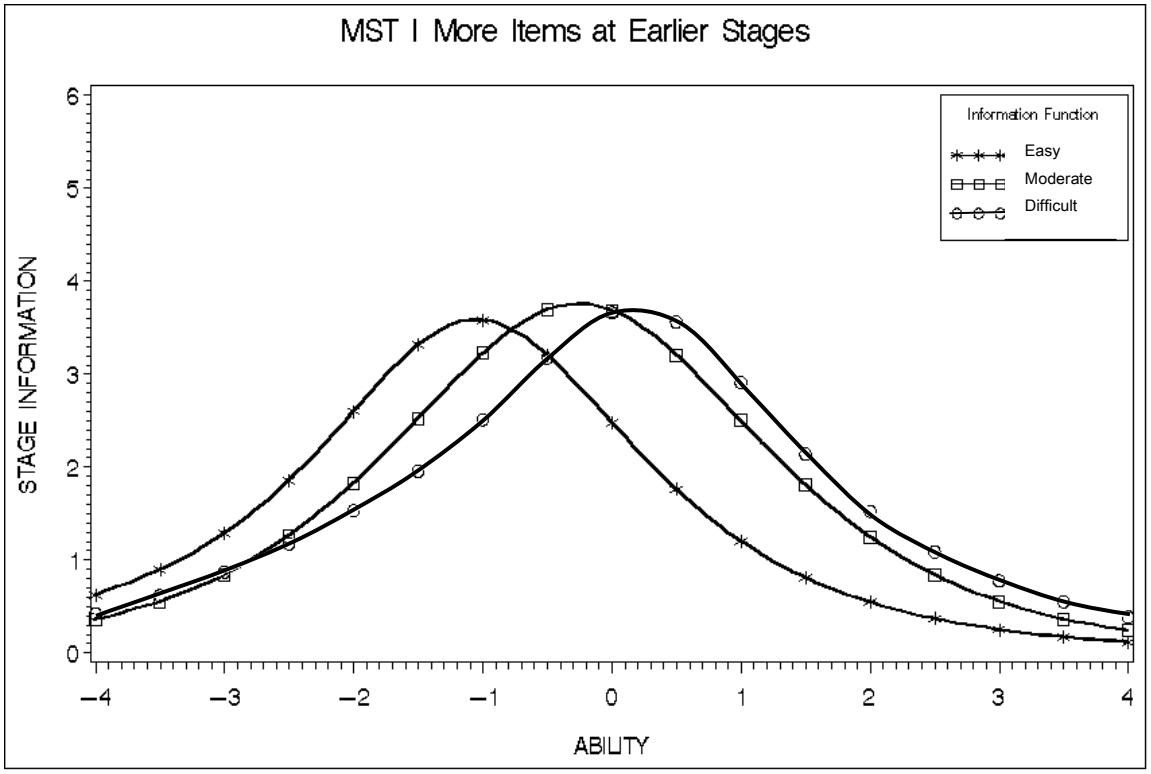

Figure 8b. MST I, Stage 2 Information Functions

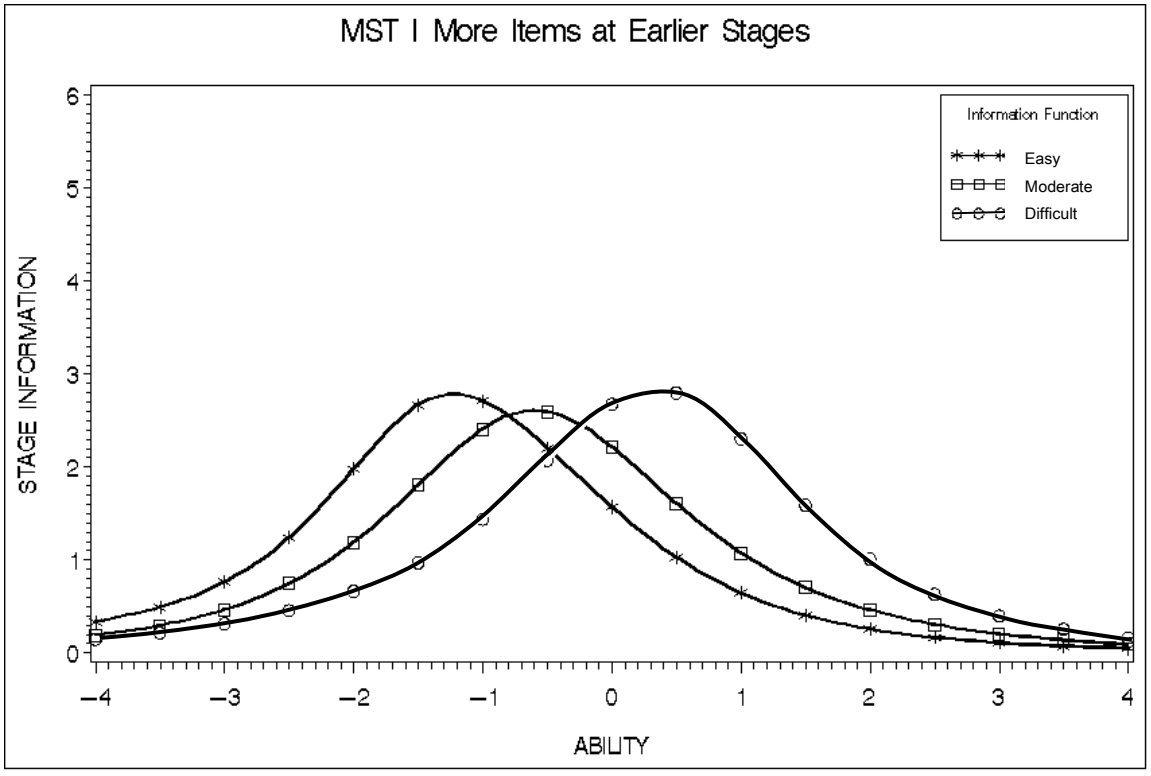

Figure 8c. MST I, Stage 3 Information Functions

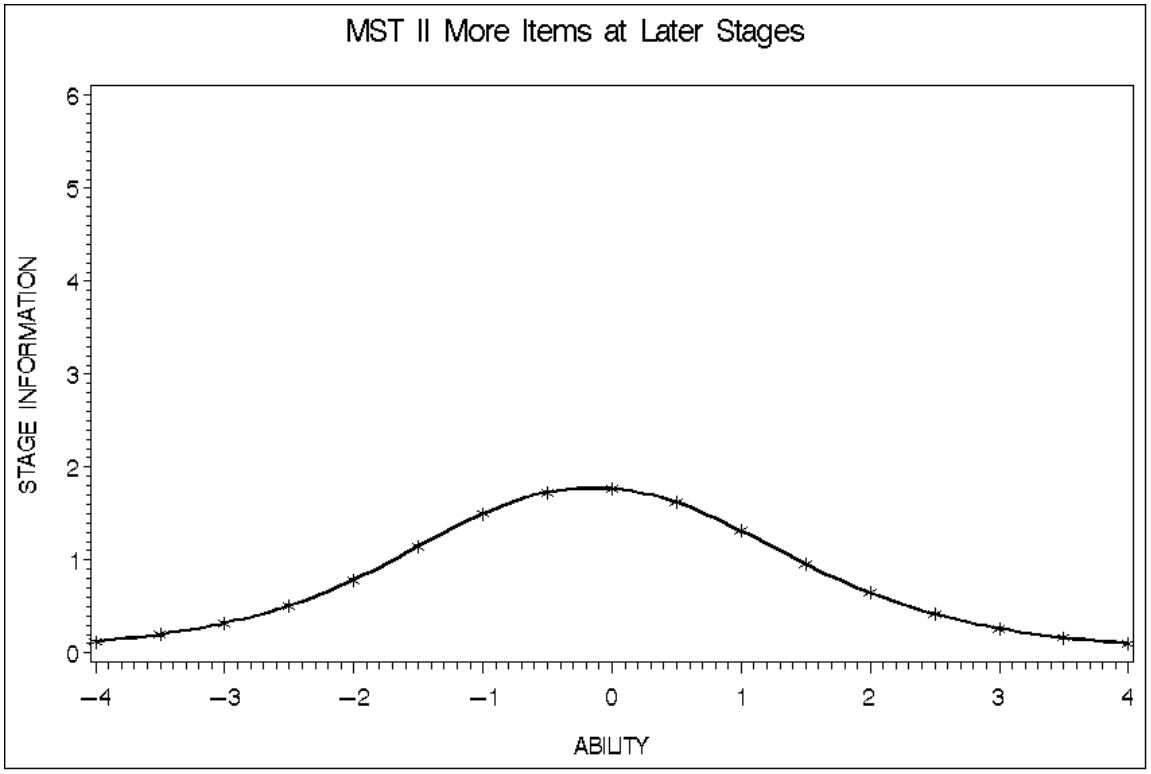

Figure 9a. MST II, Stage 1 Information Function

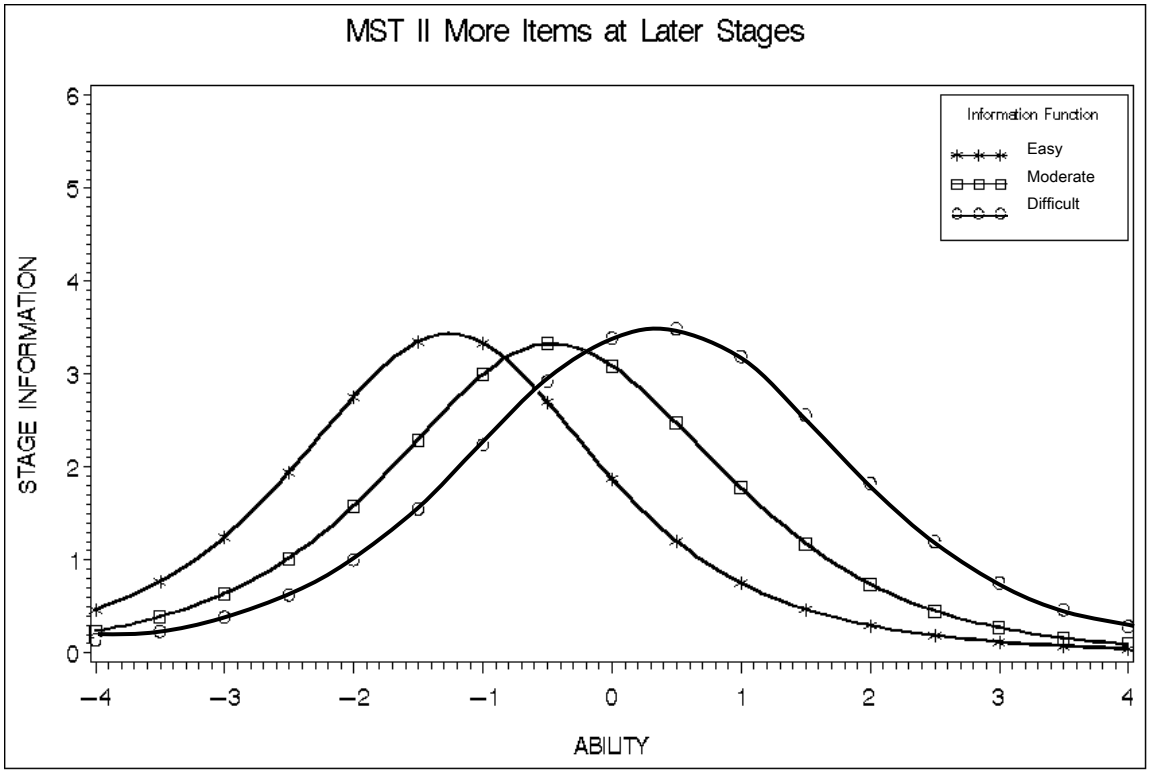

Figure 9b. MST II, Stage 2 Information Function

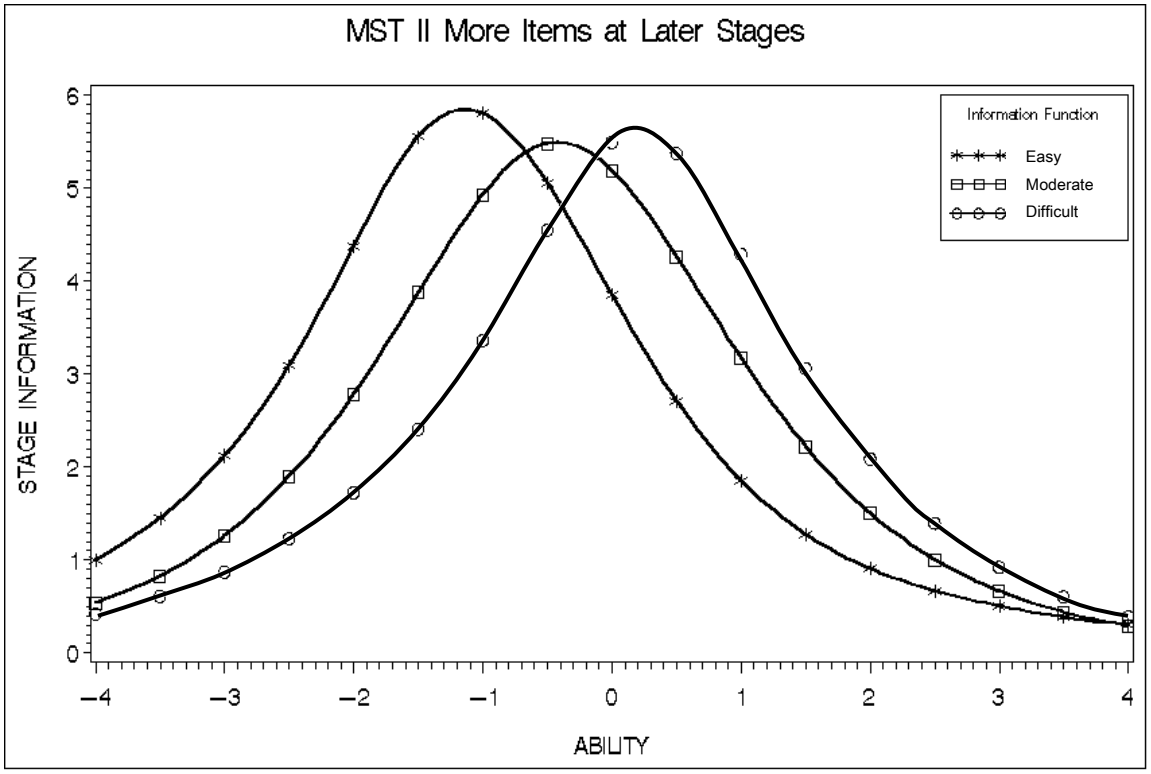

Figure 9c. MST II, Stage 3 Information Functions

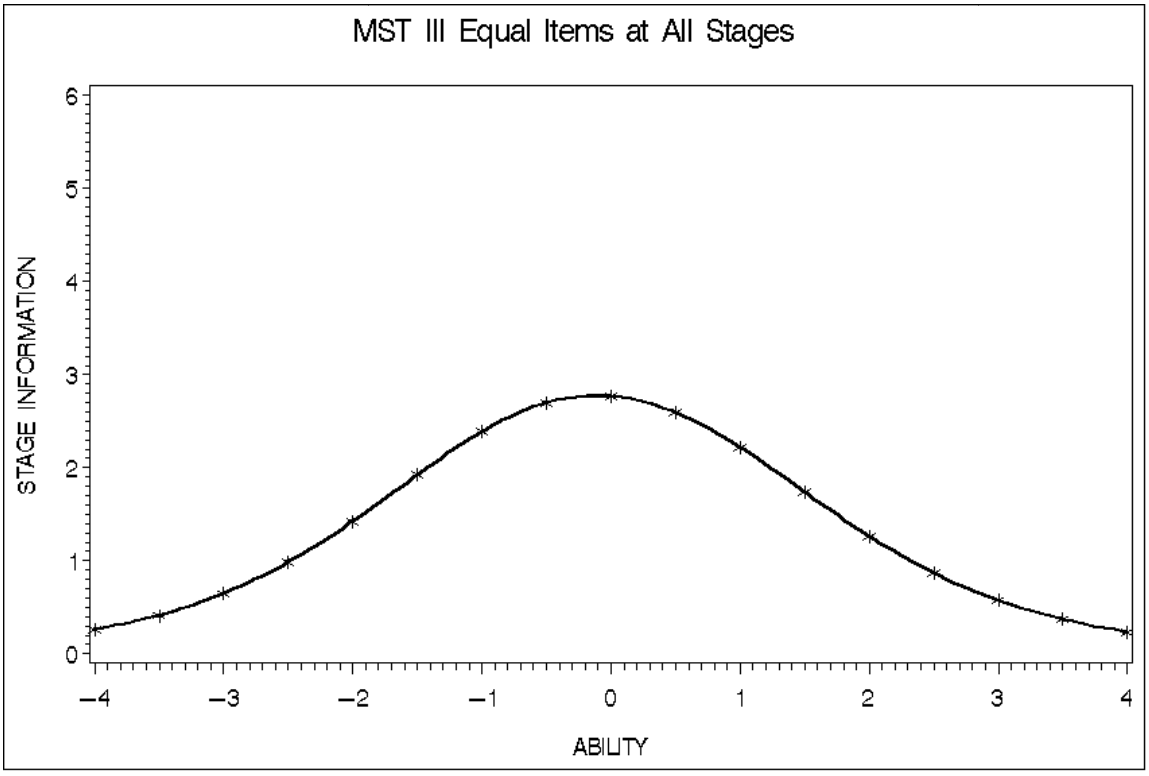

Figure 10a. MST III, Stage 1 Information Function

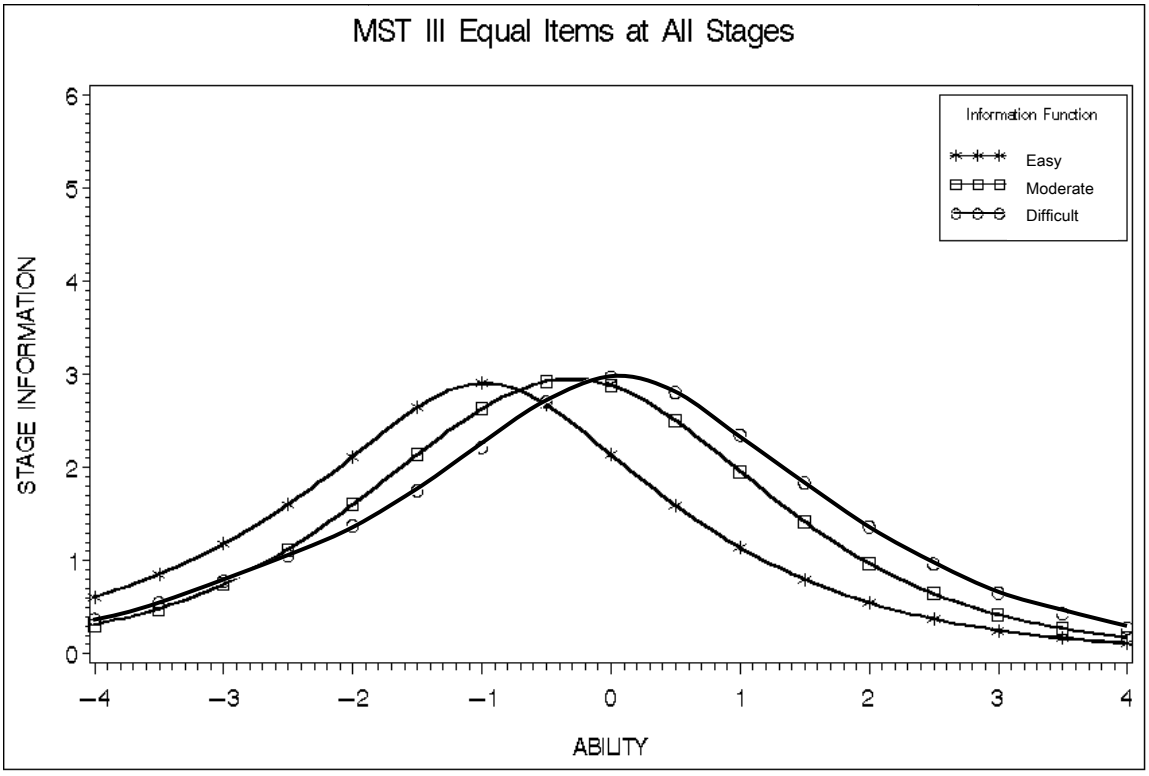

Figure 10b. MST III, Stage 2 Information Function

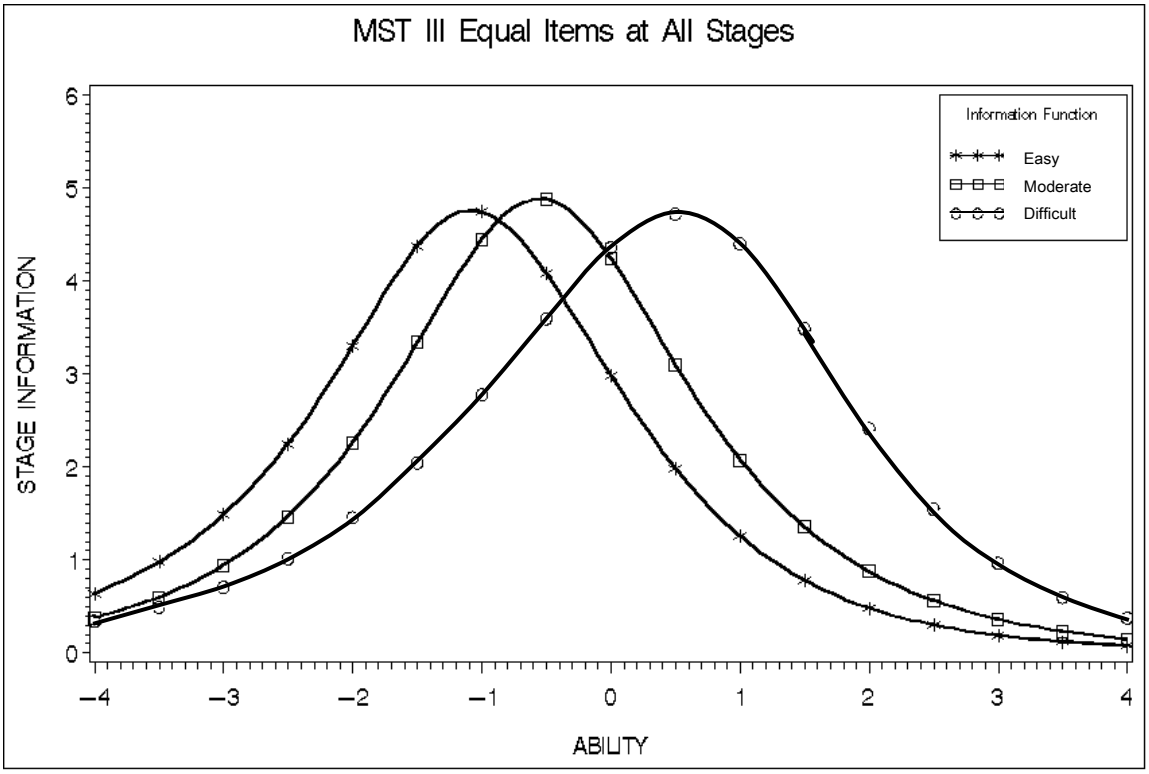

Figure 10c. MST III, Stage 3 Information Functions

Figures 11 through 13 show the total test information functions for all MST designs. All three peak around the midpoint of the theta distribution. MST III peaks somewhat more to the right of the theta distribution than the other two tests. MST II provides slightly more information than the other two designs. This was due in part to the characteristics of the available items and to the desire to limit the exposure of certain items across all three tests.

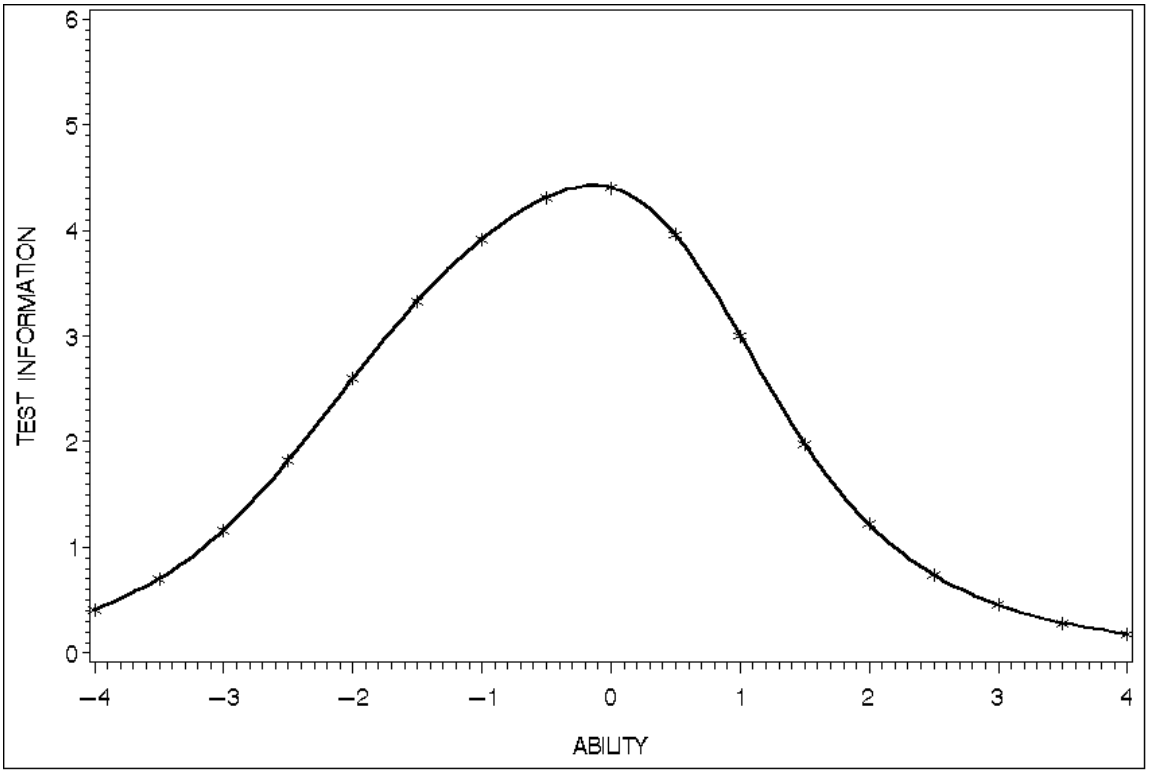

Figure 11. Test Information Function for MST 1

1

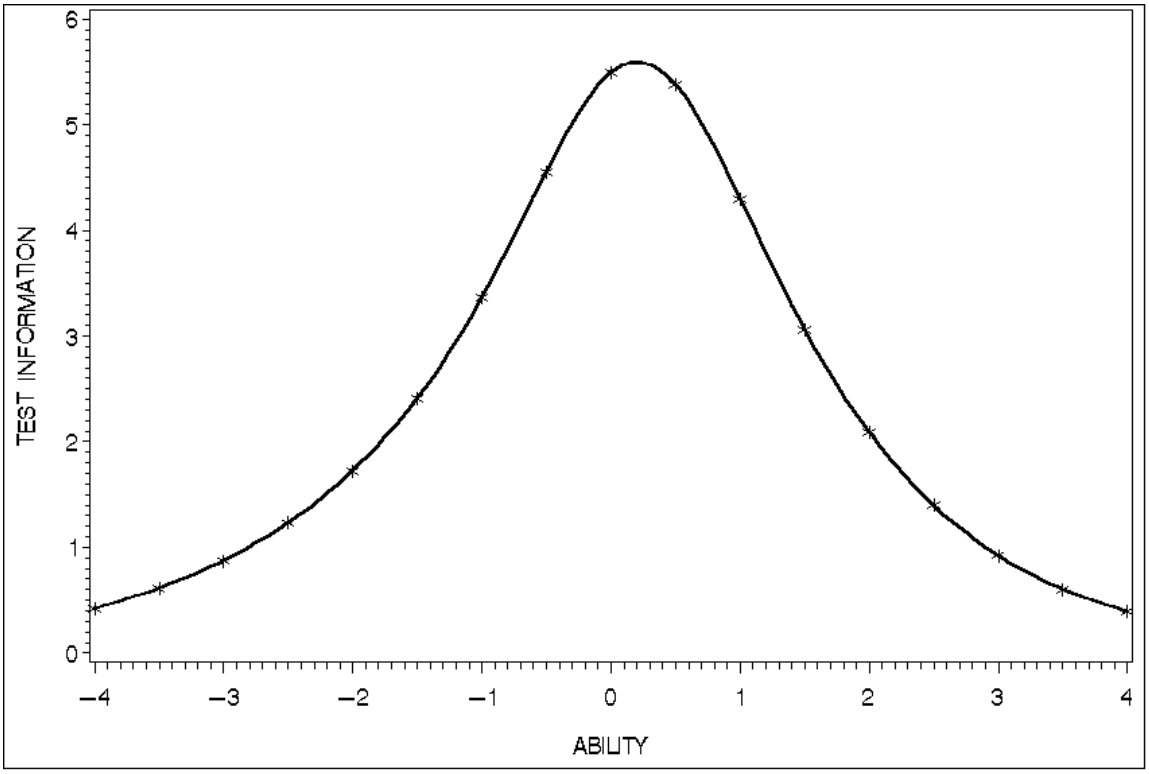

Figure 12. Test Information Function for MST II

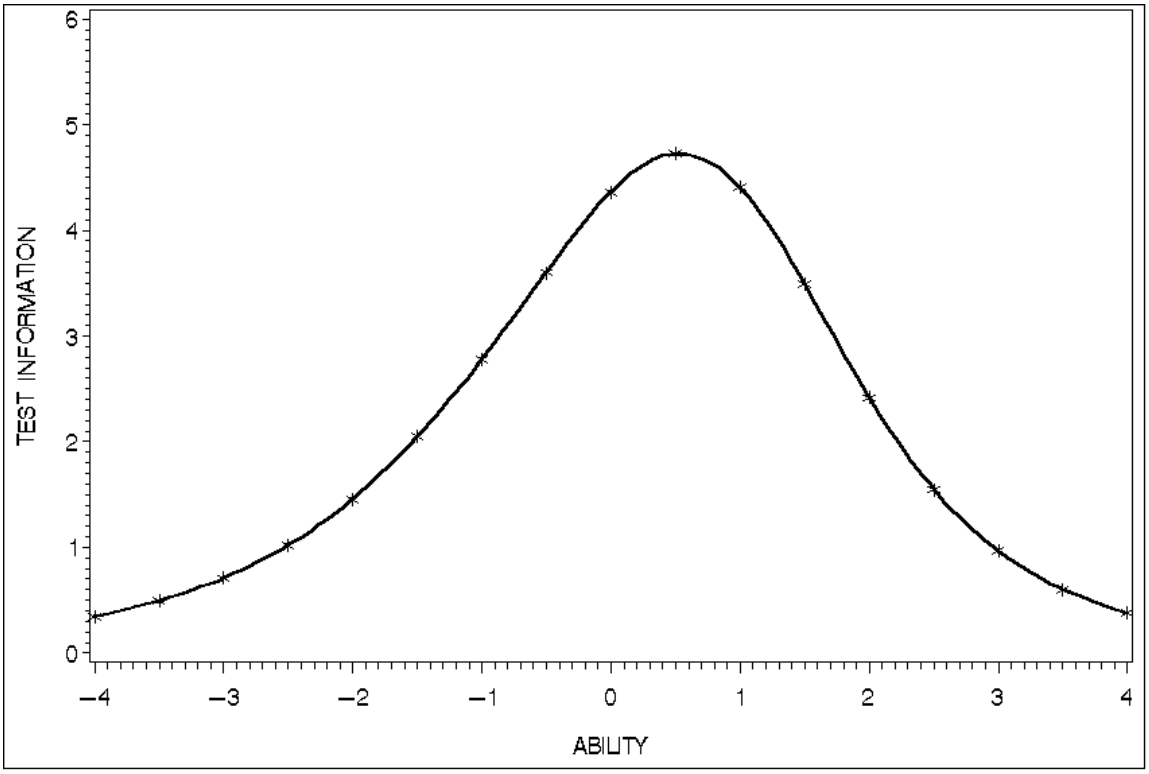

Figure 13. Test Information Function for MST III

All of the MST designs started with a module of moderate difficulty at stage 1. Three routing decision rules were implemented at stages 2 and 3. The first was routing according to maximum information in which the module at subsequent stages was selected according to which of the three yielded the most information at the current θ estimate. The second rule routed was based on a fixed level of θ: Ability estimates that were less than -1 were routed to the easy module at subsequent stages and those greater than 1 were routed to the difficult module at subsequent stages. The third rule routed simulees based the number-right in the previous module. Maximum stage-level scores were calculated at each stage, and simulees were routed such that those with a score at the lowest third of the score range were routed to the easy modules, those at the middle

third were routed to the moderate modules, and those with a score at the highest third of the score range were routed to the difficult modules.

Nonconvergence. Cases of nonconvergence were tabulated to determine whether any of the test designs yielded better results with respect to estimating simulee ability. Table 5 shows the average, minimum, and maximum number of cases of nonconvergence. MST designs are referred to as MST I, more items at earlier stages, MST II, more items at later stages, and MST III, equal number of items per stage. CAT yielded an average of 2.6 cases of nonconvergence with a minimum of 1 case and a maximum of 4 cases of nonconvergence. Two of the four MST designs, MST I and MST III, had better average results under all routing rules with a minimum of 0 cases and maximums ranging from 3 to 5. MST II performed similarly to CAT with an average of 2.6 cases under the maximum information and θ routing rules with a minimum of 0 cases and maximums of 4 to 5. MST II using the number-right routing rule performed the worst in terms of nonconvergence with an average of 7.1 cases and a minimum and a maximum of 3 and 11, respectively. In light of the larger number of nonconvergent cases under the number-right routing rule, each nonconvergent case was reviewed to determine whether $\hat{\theta}$ was unable to be estimated due to a perfect score or a 0 score. No case of nonconvergence was due to a perfect or 0 score. Among the MST designs, the maximum information and θ routing rules had fewer nonconvergent cases than the number-right routing rule for designs I and II, but the number-right routing rule produced the fewest number of nonconvergent cases under the MST III design.

Table 5. Nonconvergent Cases by Test Type

	CAT	MST I			MST II			MST III		
		MI	θ	NR	MI	θ	NR	MI	θ	NR
Average Number of Cases	2.6	1.1	1.3	2.4	2.6	2.6	7.1	2.1	2.1	1.8
Minimum	1	0	0	0	0	0	3	0	0	0
Maximum	4	4	4	5	4	5	11	4	4	3

Known and Estimated θ. Mean estimated thetas and correlations between known and estimated thetas are provided in Table 6. With respect to all designs, mean ability was estimated close to the center point of 0.0. Estimated θ ranged from 0.028 to 0.044 with the MST II design using the maximum information routing rule having the lowest $\hat{\theta}$ and the MST III design using the number-right rule having the highest $\hat{\theta}$. The mean known θ was 0.028.

Pearson correlations between the known and estimated θ-values were calculated as a measure of how well each of the tests recovered the known ability estimate. CAT yielded the highest correlation between the known and estimated θ at 0.957. Across all routing rules, MST II produced the next highest correlations, which ranged from 0.935 to 0.949, followed by MST I (0.934 to 0.941) and MST III (0.930 to 0.938). Among the MST designs the maximum information routing rule performed best in terms of recovering θ, but the fixed θ rule performed nearly as well with correlations that were only slightly lower by 0.001 to 0.002. The number-right routing rule yielded the lowest correlations.

Table 6. Average Estimated $\boldsymbol{\theta}$ and Correlation Between Known and Estimated $\boldsymbol{\theta}$ by Test Design Across Ten Replications

Test Design & Scoring Rule	$\hat{\theta}$ Mean (Min. Max.)	Correlation Mean (Min. Max)
CAT	0.029 (0.007 0.048)	0.957 (0.936 0.961)
MST I: Maximum information	0.030 (0.012 0.044)	0.941 (0.936 0.948)
MST I: Fixed θ	0.032 (0.017 0.039)	0.940 (0.935 0.946)
MST I: Number-right	0.036 (0.028 0.047)	0.934 (0.922 0.942)
MST II: Maximum information	0.028 (0.015 0.047)	0.949 (0.943 0.957)
MST II: Fixed θ	0.024 (0.010 0.038)	0.947 (0.940 0.953)
MST II: Number-right	0.030 (0.026 0.046)	0.935 (0.920 0.944)
MST III: Maximum information	0.032 (0.021 0.047)	0.938 (0.931 0.948)
MST III: Fixed θ	0.029 (0.008 0.045)	0.936 (0.927 0.945)
MST III: Number-right	0.044 (0.028 0.059)	0.930 (0.921 0.938)

Table 7 provides the average standard deviations of estimated theta. CAT, MST II (maximum information), MST II (fixed θ), and MST III (maximum information) performed similarly with regard to the standard deviation. The mean standard deviation

for MST II using the maximum information routing rule was the lowest at 1.055. The standard deviation was 1.056b for the remaining three. Within MST test type, the number-right routing rule produced the largest mean standard deviations and the widest ranges between minimum and maximum mean standard deviations, indicating more variability the estimation of θ compared to CAT and the MSTs under the maximum information and fixed the other routing rules.

Table 7. Standard Deviation for Estimated Theta Across Ten Replications

Test Design & Scoring Rule	Standard Deviation		
	Grand Mean	Minimum Mean	Maximum Mean
CAT	1.056	1.037	1.071
MST I: Maximum information	1.068	1.049	1.094
MST I: Fixed θ	1.067	1.046	1.095
MST I: Number-right	1.077	1.056	1.112
MST II: Maximum information	1.055	1.040	1.078
MST II: Fixed θ	1.056	1.027	1.086
MST II: Number-right	1.068	1.048	1.103
MST III: Maximum information	1.059	1.031	1.078
MST III: Fixed θ	1.056	1.027	1.086
MST III: Number-right	1.085	1.062	1.102

Average Error Statistics. Average error statistics are provided in Table 8. Three error statistics were calculated, bias, the root mean square error (RMSE), and the average absolute difference between known and estimated theta.

Bias is an index of error in item selection. Smaller bias indicates better item selection when items are near the true θ-level. If θ is higher than the average difficulty of the item, bias will be positive, and will be negative otherwise. All of the designs produced a negative average bias with the exception of the MST II when the maximum information and fixed θ routing rules were used. In the latter instances, average bias was positive. The bias estimate are not solely due to item selection, however, since the MST II number-right routing rule using the same items produced an average negative bias, implying that in the case of the MSTs the routing rules also had an impact on θ estimation. Bias was lowest for the MST II maximum information routing, 0.000, and highest for the MST I and MST III number-right routing, -0.010 and -0.016, respectively.

RMSE is another measure of error in item selection and is an indicator of how well the estimate of ability can consistently approximate its true value. CAT produced the lowest average RMSE, 0.308, followed by the MST II, MST I, and MST III designs. The maximum information and θ routing rules caused the average RMSEs to be the lowest among the MST designs ranging from 0.332 (MST II maximum information rule) to 0.365 (MST I θ rule); the number-right routing rule produced the highest average RMSE in all designs.

The absolute error provides an indicator of the magnitude of the difference between true ability and its approximation. Once again, CAT yielded the lowest average absolute difference between known and estimated θ, 0.241 followed by the MST II, MST I, and MST III designs. Among the MST designs, the lowest average absolute difference was found using the maximum information rule (0.274, 0.251, 0.279 for MST I, MST II,

and MST III, respectively), but the θ rule yielded only slightly higher values (0.278, 0.258, and 0.283, respectively). The number-right routing rule yielded the highest average absolute differences among the MST designs.

Table 8. Average Error Statistics by Test Design Across Ten Replications

Test Design & Scoring Rule	Bias Mean (Min. Max.)	RMSE Mean (Min. Max.)	Average Absolute Difference Mean (Min. Max.)
CAT	-0.002 (-0.020 0.017)	0.308 (0.286 0.316)	0.241 (0.226 0.249)
MST I: Maximum information	-0.002 (-0.023 0.019)	0.361 (0.339 0.380)	0.274 (0.265 0.287)
MST I: Fixed θ	-0.003 (-0.026 0.011)	0.365 (0.346 0.381)	0.278 (0.267 0.292)
MST I: Number-right	-0.010 (-0.019 -0.000)	0.384 (0.367 0.411)	0.287 (0.278 0.298)
MST II: Maximum information	0.000 (-0.018 0.009)	0.332 (0.302 0.360)	0.251 (0.238 0.263)
MST II: Fixed θ	0.004 (-0.010 0.018)	0.338 (0.315 0.363)	0.258 (0.249 0.269)
MST II: Number-right	-0.005 (-0.018 0.002)	0.379 (0.347 0.419)	0.279 (0.026 0.046)
MST III: Maximum information	-0.004 (-0.019 0.021)	0.367 (0.343 0.387)	0.279 (0.265 0.288)
MST III: Fixed θ	-0.001 (-0.019 0.020)	0.372 (0.349 0.389)	0.283 (0.271 0.288)
MST III: Number-right	-0.016 (-0.030 -0.000)	0.400 (0.389 0.413)	0.302 (0.294 0.308)

Average standard errors are provided in Table 9. The standard error provides an indication of the precision of measurement and consequently, how well the test estimates the true θ. CAT was the most precise in estimation with an average standard error of 0.301. CAT was followed by the MST II maximum information design with an average standard error of 0.311. MST II performed the best in terms of the standard error among all the MST designs, and MST III performed the worst. By routing rule, maximum information performed the best, and the number-right routing rule performed the worst.

Table 9. Average Standard Errors by Test Design Across Ten Replications

Test Design & Scoring Rule	Standard Error		
	Grand Mean	Minimum Mean	Maximum Mean
CAT	0.301	0.276	0.304
MST I: Maximum information	0.335	0.333	0.338
MST I: Fixed θ	0.342	0.339	0.344
MST I: Number-right	0.355	0.353	0.359
MST II: Maximum information	0.311	0.309	0.313
MST II: Fixed θ	0.320	0.318	0.322
MST II: Number-right	0.347	0.342	0.354
MST III: Maximum information	0.342	0.341	0.345
MST III: Fixed θ	0.351	0.348	0.353
MST III: Number-right	0.373	0.370	0.375

Exposure Rates. Average exposure rates are shown in Table 10. Then mean exposure rates was the same for CAT, MST I, and MST II because each was a fixed length test of 20 items. The mean exposure rate for MST III was lower because it was a fixed length test containing 18 items. CAT with the PR30 exposure control procedure was only

slightly over the specified exposure rate level at 0.031. Since only one panel was constructed for each of the MST designs, the maximum exposure rate was 1.00. Due to the fact that only one panel was constructed per design the items that were administered in the test were limited to those in each module at each test stage. In contrast, the adaptive nature of the CAT allowed for the best use of the item pool.

Table 10. Average Exposure Rates by Test Design Across Ten Replications

Test Design & Scoring Rule	Item Exposure Rate		
	Grand Mean	Minimum Mean	Maximum Mean
CAT	0.072	0.001	0.301
MST I: Maximum information	0.072	0.000*	1.00*
MST I: Fixed θ	0.072	0.000*	1.00*
MST I: Number-right	0.072	0.000*	1.00*
MST II: Maximum information	0.072	0.000*	1.00*
MST II: Fixed θ	0.072	0.000*	1.00*
MST II: Number-right	0.072	0.000*	1.00*
MST III: Maximum information	0.065	0.000*	1.00*
MST III: Fixed θ	0.065	0.000*	1.00*
MST III: Number-right	0.065	0.000*	1.00*

* Only one panel was constructed per design

Average standard deviations for exposure rates are shown in Table 11. The average standard deviation for CAT at 0.076 was lower than it was for any of the MST designs, largely due to the PR30 exposure control procedure that bounded the

maximum exposure rate. Within the three MST designs, the maximum information routing rule always provided the most control over the standard deviation of the exposure rate, and the number-right routing rule provided the least. Comparing the three designs, MST II tended to provide the most control, followed by MST I, and MST III.

Table 11. Average Standard Deviations of Exposure Rates Across Ten Replications

	Standard Deviation of Exposure Rates		
	Mean	Minimum	Maximum
CAT	0.076	0.058	0.088
MST I: Maximum information	0.212*	0.211*	0.212*
MST I: Fixed θ	0.217*	0.216*	0.218*
MST I: Number-right	0.226*	0.226*	0.227*
MST II: Maximum information	0.170*	0.168*	0.171*
MST II: Fixed θ	0.181*	0.177*	0.184*
MST II: Number-right	0.218*	0.217*	0.221*
MST III: Maximum information	0.186*	0.185*	0.187*
MST III: Fixed θ	0.191*	0.190*	0.193*
MST III: Number-right	0.205*	0.204*	0.205*

* Only one panel was constructed per design

The average frequency of exposure rates is provided in Table 12. It is clear that CAT provided the best control over exposure rates, with most item exposure rates falling in the .01-.05 percentage range and none occurring over .35 percentage interval. None of the MST designs achieved such control. Note, however, that among the MST maximum

information designs, MST II maximum information, in particular, that even though the exposure rates are larger than those of CAT, that the exposure rates are controlled over a continuous range. For example, in MST II maximum information the frequency of exposure ranges from .11 to .60 percent; there are no skipped intervals. This is largely true in MST I and MST III maximum information, as well. In the fixed θ and number-right routing conditions there is no pattern to the exposure rates intervals, suggesting that even though it does not provide the same control as PR30, the maximum information routing rule made better use of the available information than either of the other two routing rules.

Table 12. Frequency of Exposure Rates Across Ten Replications

Exposure Rate	CAT	MST I MI*	MST I θ*	MST I NR*	MST II MI*	MST II θ*	MST II NR*	MST III MI*	MST III θ*	MST III NR*
1.0	0	2	2	1	0	0	0	1	1	0
.91-.99	0	0	0	0	0	0	0	0	0	0
.81-.90	0	0	0	0	0	0	12	0	0	0
.71-.80	0	0	0	3	0	0	0	0	0	6
.61-.70	0	0	5	7	0	10	0	0	6	0
.51-.60	0	3	5	0	2	0	5	6	6	6
.41-.50	0	7	0	0	16	3	5	6	0	6
.36-.40	0	4	0	7	4	9	0	1	0	0
.31-.35	3	3	0	0	8	0	0	11	0	0
.26-.30	6	0	0	2	5	0	0	0	4	5
.21-.25	9	12	8	1	12	5	0	3	8	1
.16-.20	7	1	12	0	5	24	8	4	11	0
.11-.15	8	0	0	0	1	0	4	5	0	0
.06-.10	71	0	0	0	0	0	0	0	0	0
.01-.05	104	0	0	3	0	0	0	0	0	6
0.0	0	176	176	184	155	157	174	171	172	178
% Not Admin	0%	85%	85%	88%	75%	75%	84%	82%	83%	86%

* Only one panel was constructed per design

Item Overlap. The amount of items shared by simulees was calculated by examining the each of the audit trails, and item overlap indices are provided in Table 13. The overall item overlap provides an index of overlap across all simulees regardless of estimated proficiency level. If estimated proficiency was within 1 logit, simulees were considered to have the same ability. If estimated proficiency was greater than 1 logit, they were considered to have different abilities.

Table 13. Item Overlap Across Ten Replications

Test Design & Scoring Rule	Item Overlap		
	Overall Overlap Grand Mean (Min, Max)	Similar Abilities Grand Mean (Min, Max)	Different Abilities Grand Mean (Min, Max)
CAT	2.490 (2.335, 3.656)	1.629 (1.364, 3.634)	2.636 (2.514, 3.660)
MST I: Maximum information	13.753* (13.684, 13.825)	10.732* (10.676, 10.770)	14.317* (14.241, 14.389)
MST I: Fixed θ	14.490* (14.297, 14.619)	10.663* (10.487, 10.782)	15.200* (14.993, 15.335)
MST I: Number-right	15.577* (15.489, 15.674)	11.805* (11.714, 11.835)	16.280* (16.192, 16.383)
MST II: Maximum information	7.582* (7.306, 9.308)	4.943* (4.746, 5.051)	10.177* (10.070, 10.327)
MST II: Fixed θ	9.740* (6.840, 10.708)	4.628* (4.153, 4.943)	11.543* (11.202, 11.818)
MST II: Number-right	14.455* (14.288, 14.700)	9.099* (8.817, 9.524)	15.452* (15.282, 15.670)
MST III: Maximum information	10.695* (10.608, 10.782)	7.262* (7.208, 7.309)	11.334* (11.241, 11.445)
MST III: Fixed θ	11.232* (11.142, 11.412)	6.997* (8.314, 8.509)	12.022* (11.905, 12.193)
MST III: Number-right	12.649* (12.586, 12.721)	8.452* (8.314, 8.509)	13.431* (13.355, 13.491)

* Only one panel was constructed per design

CAT had the lowest overlap rates among all test designs. This was the result of two factors, the PR30 exposure control procedure and the efficient use of the item pool, in which all of the items were administered. The mean overall overlap was 2.49. For

simulees with similar abilities the mean overlap was 1.629; for those with different abilities it was 2.636.

Overlap indices were greater for all of the MST designs, largely due to the fact that only one panel was constructed per design. MST II, maximum information, had the least mean overall overlap at 7.582 items. This was followed by MST II, fixed θ, with a mean overall overlap of 9.74 items. MST III, maximum information, followed with 10.695 items. All other designs followed; the range of overall mean overlap was 11.232 (MST II, fixed θ) to 15.577 (MST I, number right). Similar to the patterns seen earlier, maximum information had the best performance within test design, and number-right routing had the worst performance. MST II had the best performance across all test designs, followed by MST III, and MST I. It should be remembered, however, that MST III had fewer items (18 versus 20), and this may have contributed to better performance with respect to average overlap.

Fewer items were shared among simulees with similar abilities than among simulees with different abilities, which in itself provided some measure of exposure control since fewer items were exposed to persons of similar proficiency. MST II, fixed θ, had the smallest mean overlap for simulees with similar abilities of 4.628 items, followed by MST II, maximum information at 4.943 items. MST I, number-right, had the largest mean overlap for simulees with similar abilities, 11.805. Similar patterns held true for simulees with different abilities with maximum information having the best performance within test design, and number-right routing having the worst performance. Across designs, MST II had the best performance followed by MST III, and MST I.

Conditional Standard Errors. The mean standard errors for each replication were plotted against known θ for all conditions as shown in Figures 14-23. For all conditions, standard error was lowest in the middle of the range, inversely corresponding to the peak of the information function, and increased as θ became more extreme. Compared to CAT, there was greater degradation in standard error estimation relative to known θ at the upper end of the distribution in all MST designs. Degradation was most pronounced in the number-right routing rule condition.

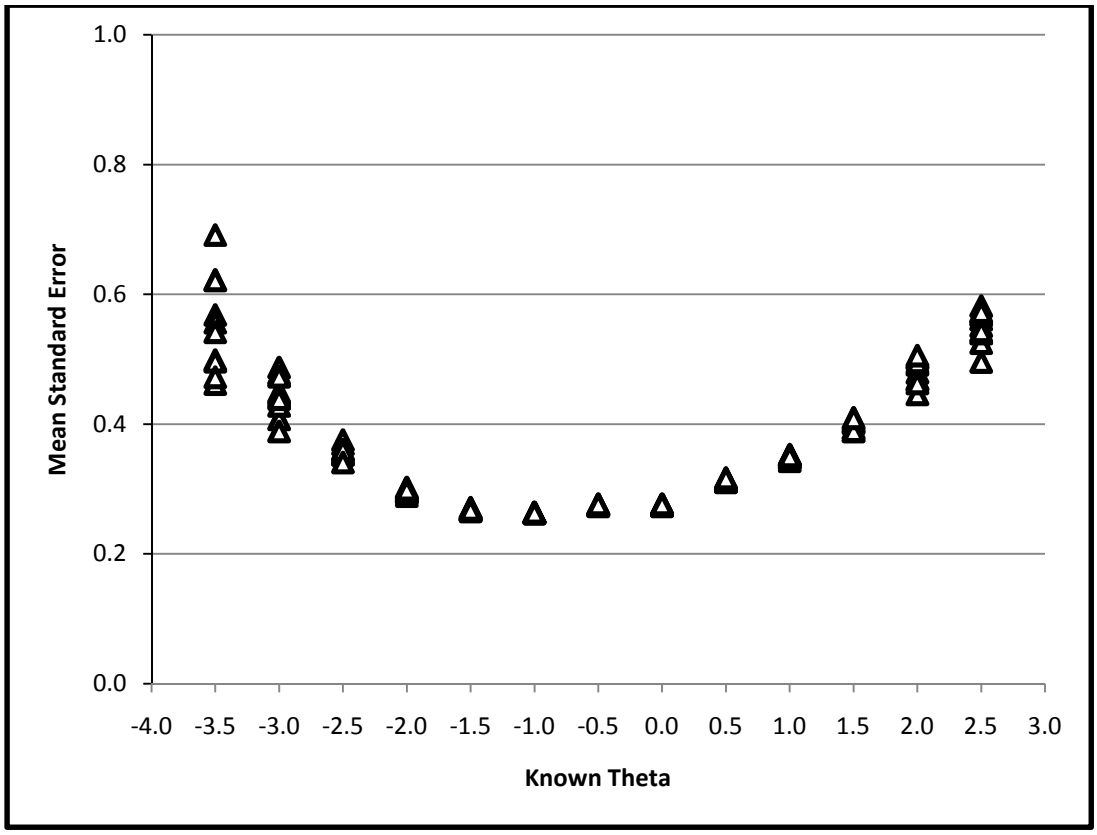

Figure 14. Mean Conditional Standard Error Plots for CAT

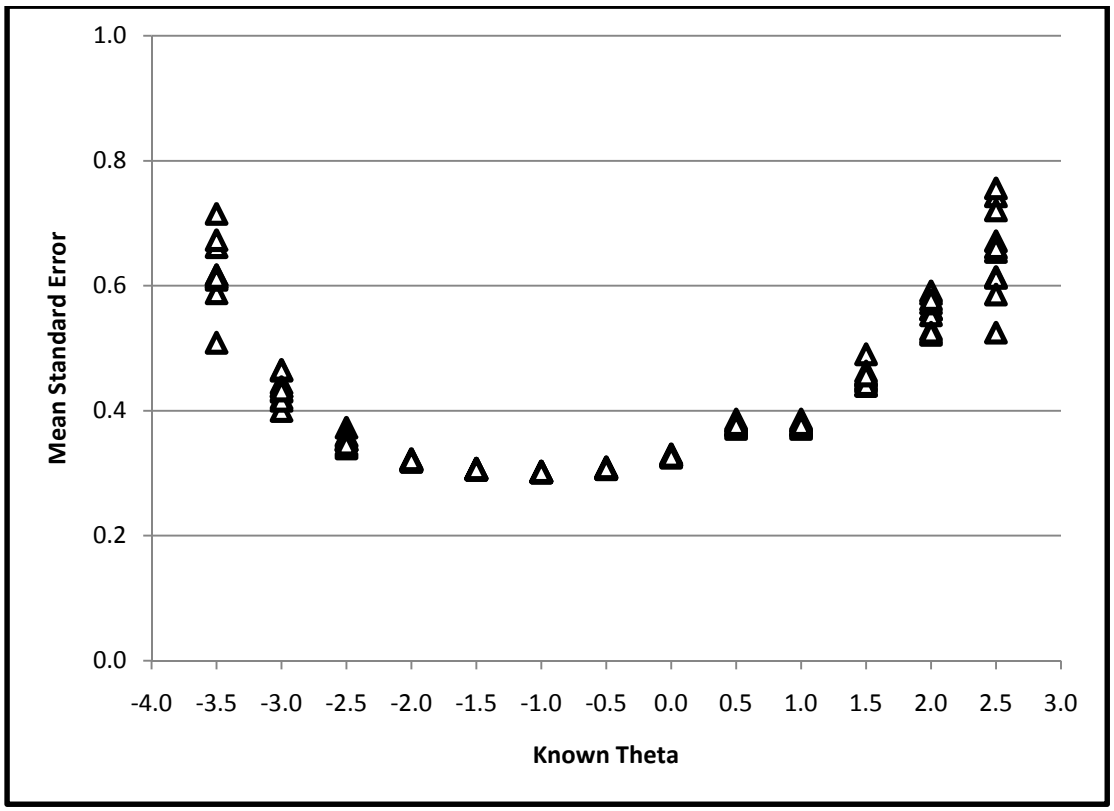

Figure 15. Mean Conditional Standard Error Plots for MST I Maximum Information

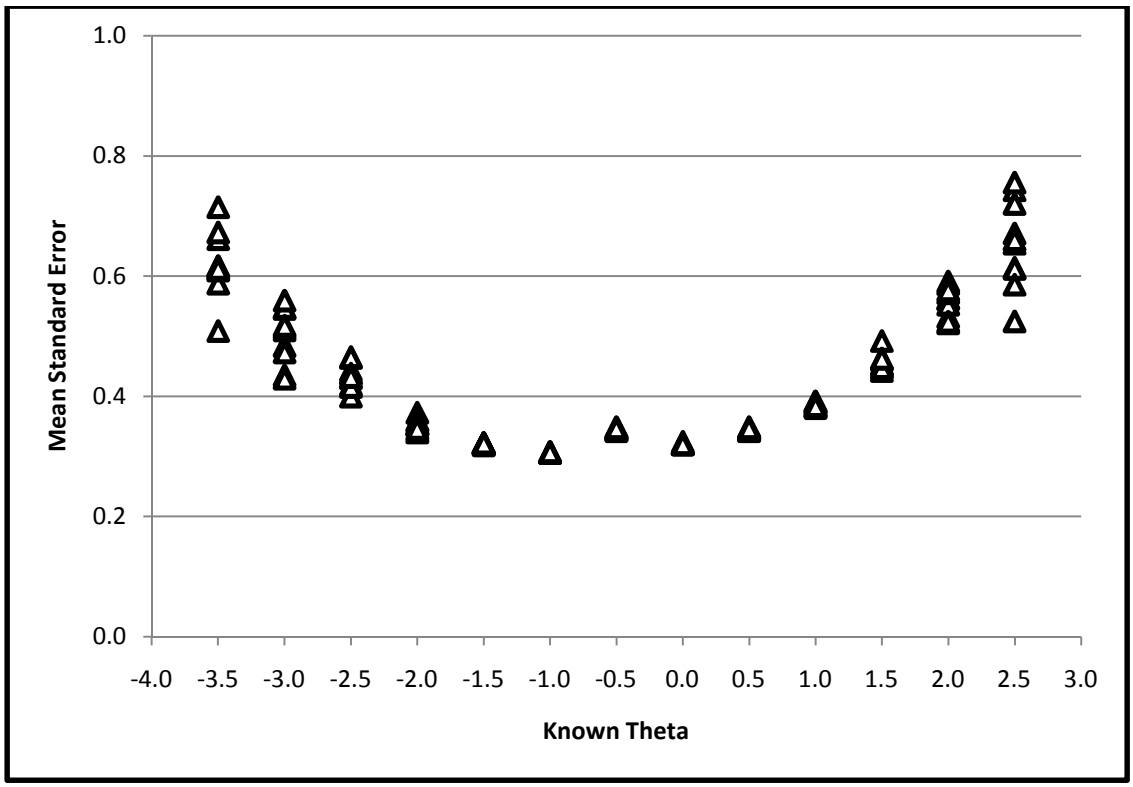

Figure 16. Mean Conditional Standard Error Plots for MST I Fixed θ

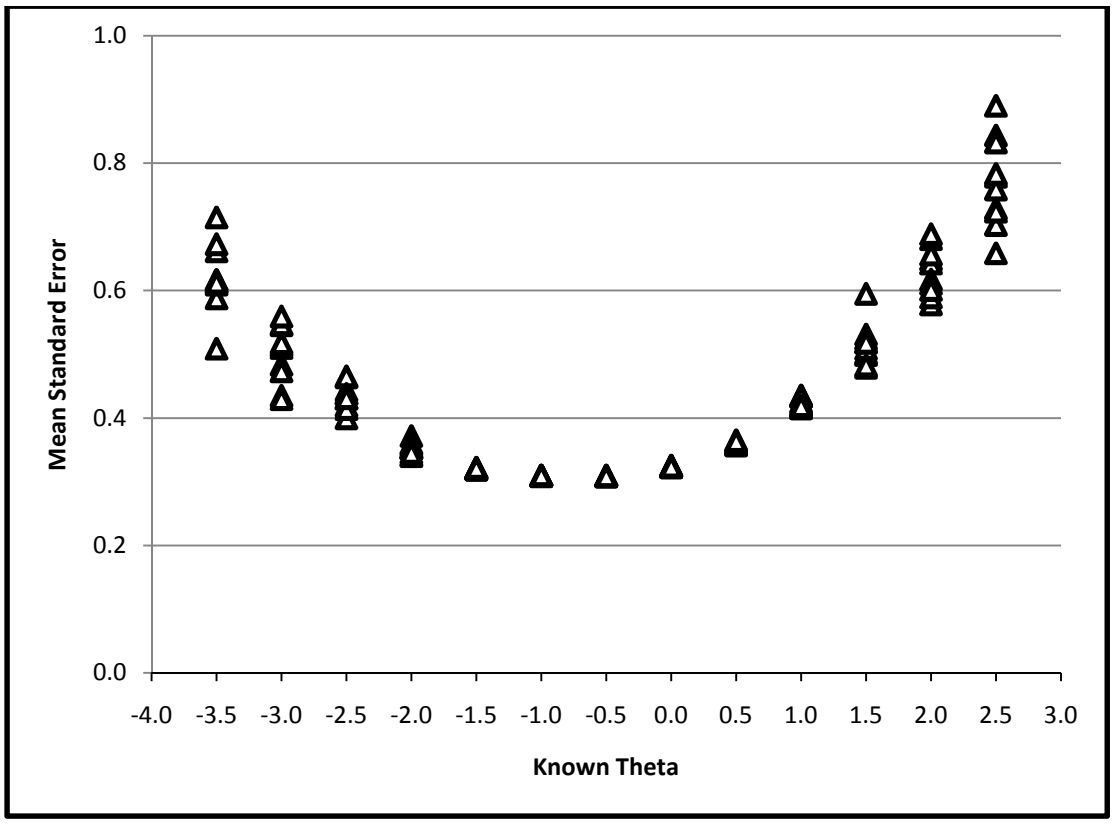

Figure 17. Mean Conditional Standard Error Plots for MST I Number-Right

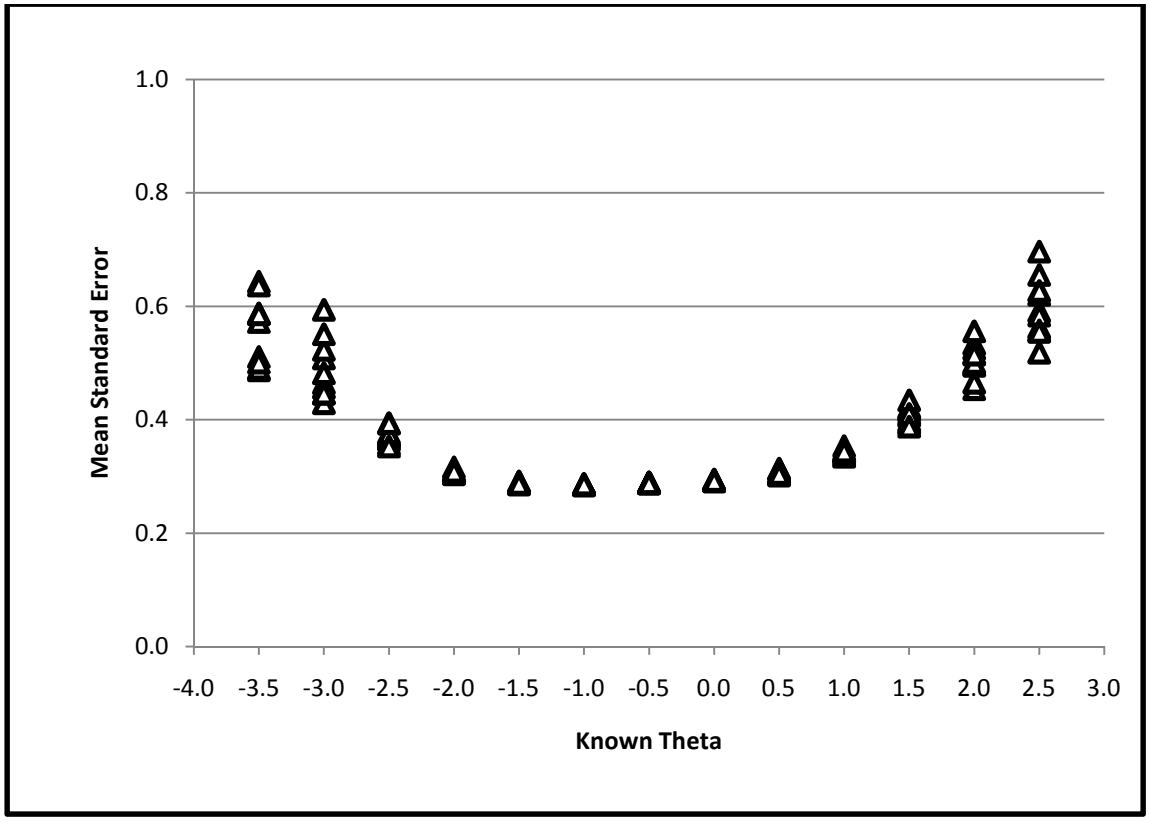

Figure 18. Mean Conditional Standard Error Plots for MST II Maximum Information

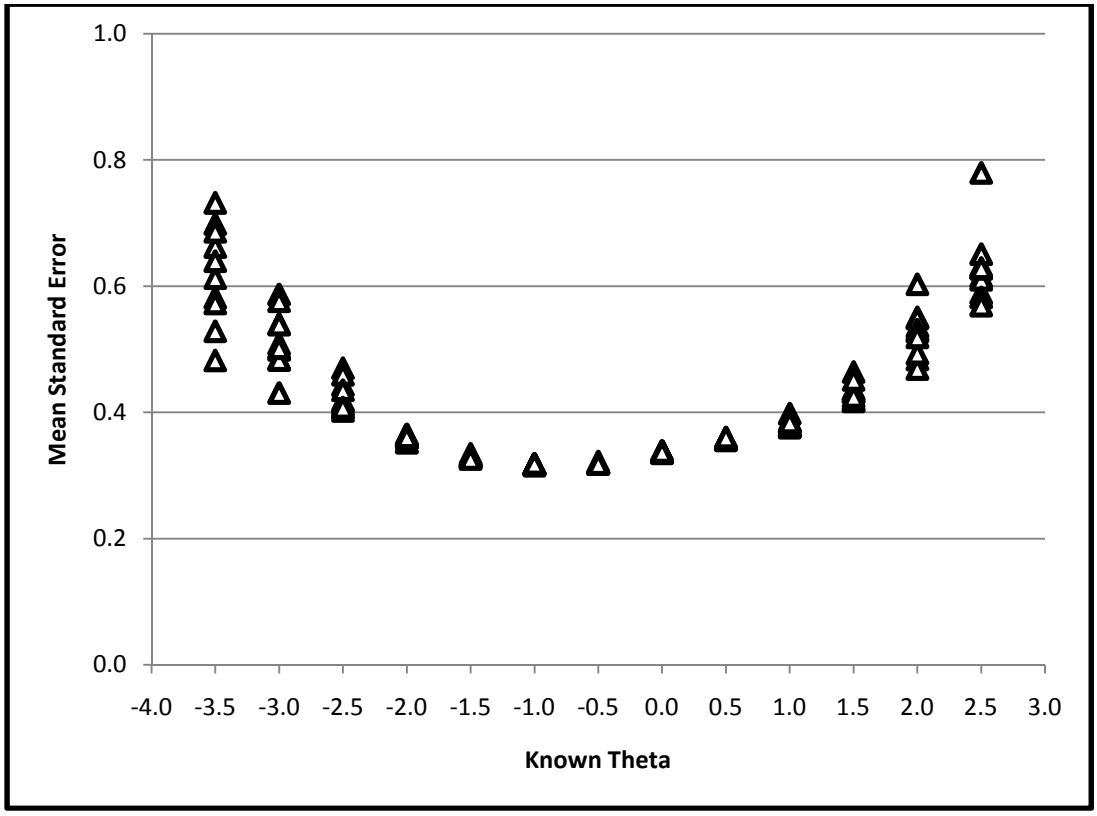

Figure 19. Mean Conditional Standard Error Plots for MST II Fixed θ

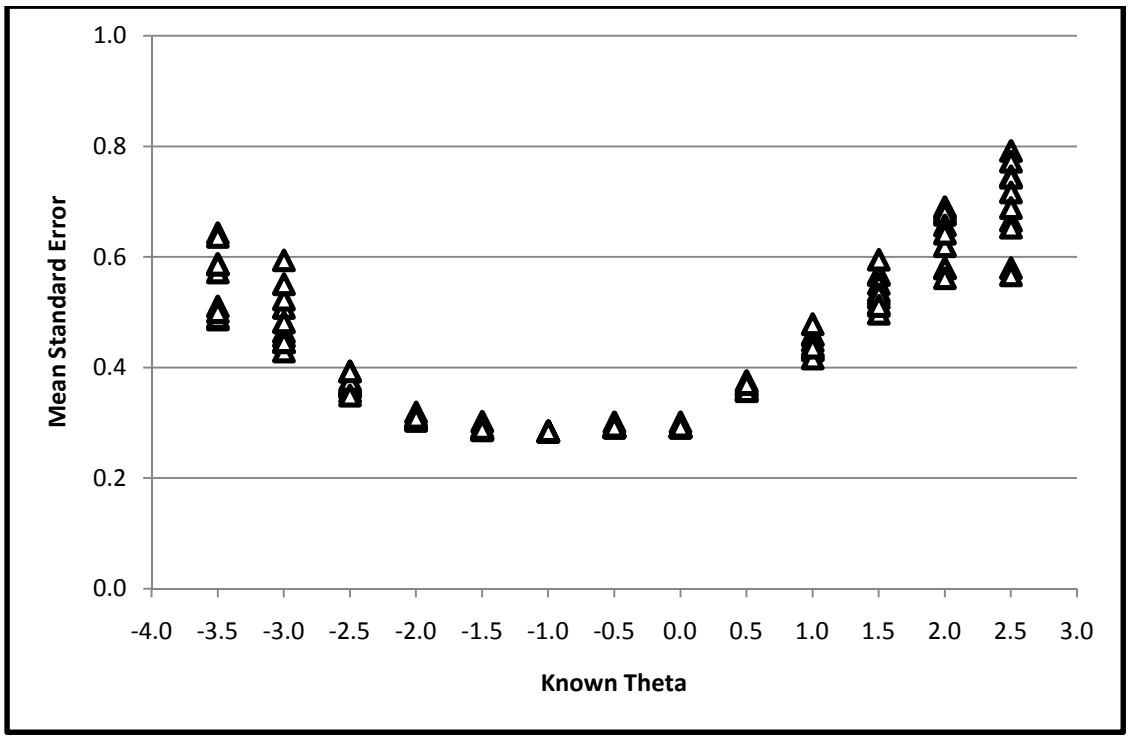

Figure 20. Mean Conditional Standard Error Plots for MST II Number-Right

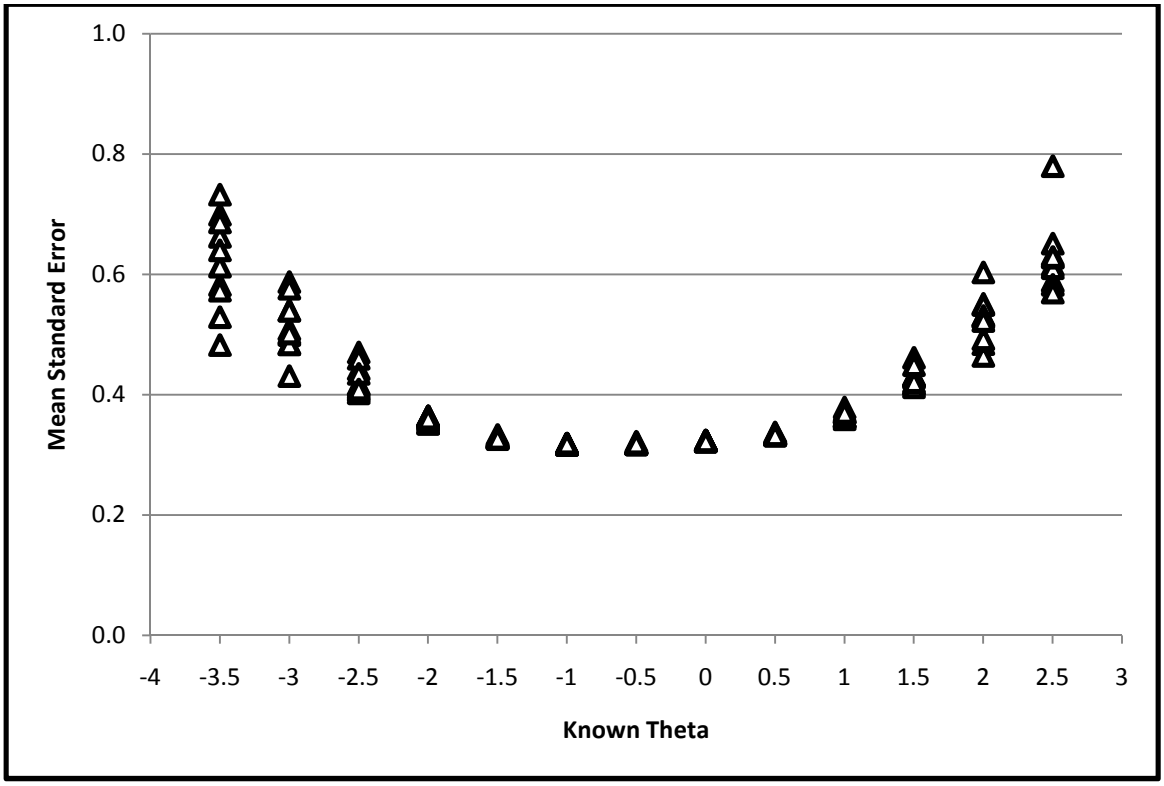

Figure 21. Mean Conditional Standard Error Plots for MST III Maximum Information

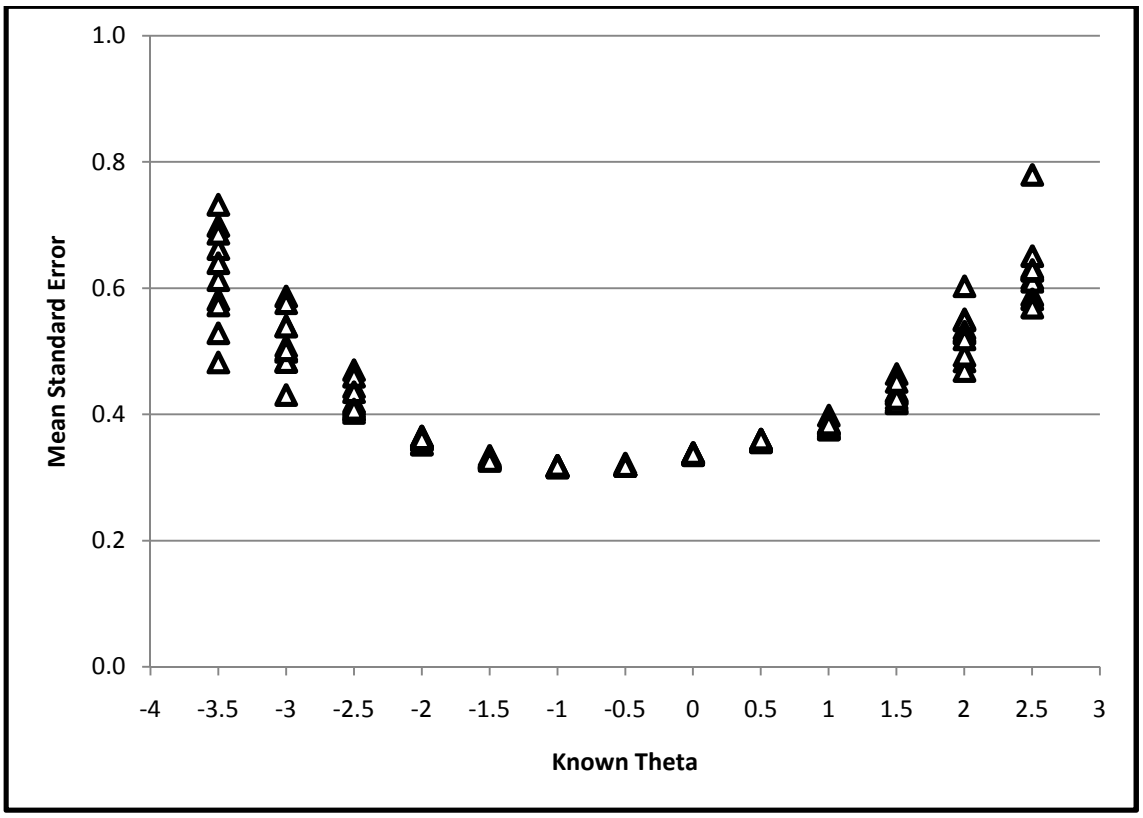

Figure 22. Mean Conditional Standard Error Plots for MST III Fixed θ

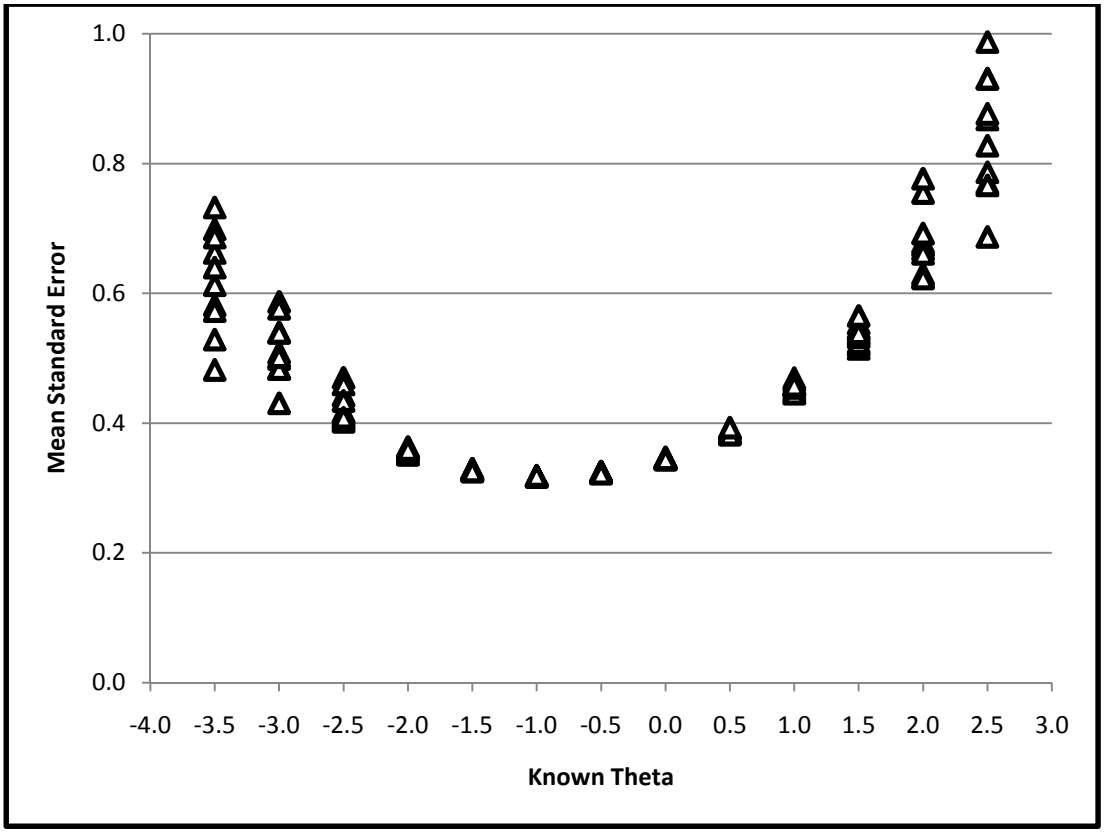

Figure 23. Mean Conditional Standard Error Plots for MST III Number-Right

Conditional Bias. The bias for each replication is provided in Figures 24-33. Note that in all MST designs bias tended to occupy a narrower range than did CAT across most of the θ distribution. From these graphs, it appears likely that the slightly higher correlation between estimated and known θ in CAT relative to some of the other designs, namely those employing maximum information, was due to the presence of fewer outliers rather than a tighter range of estimation through the majority of the distribution.

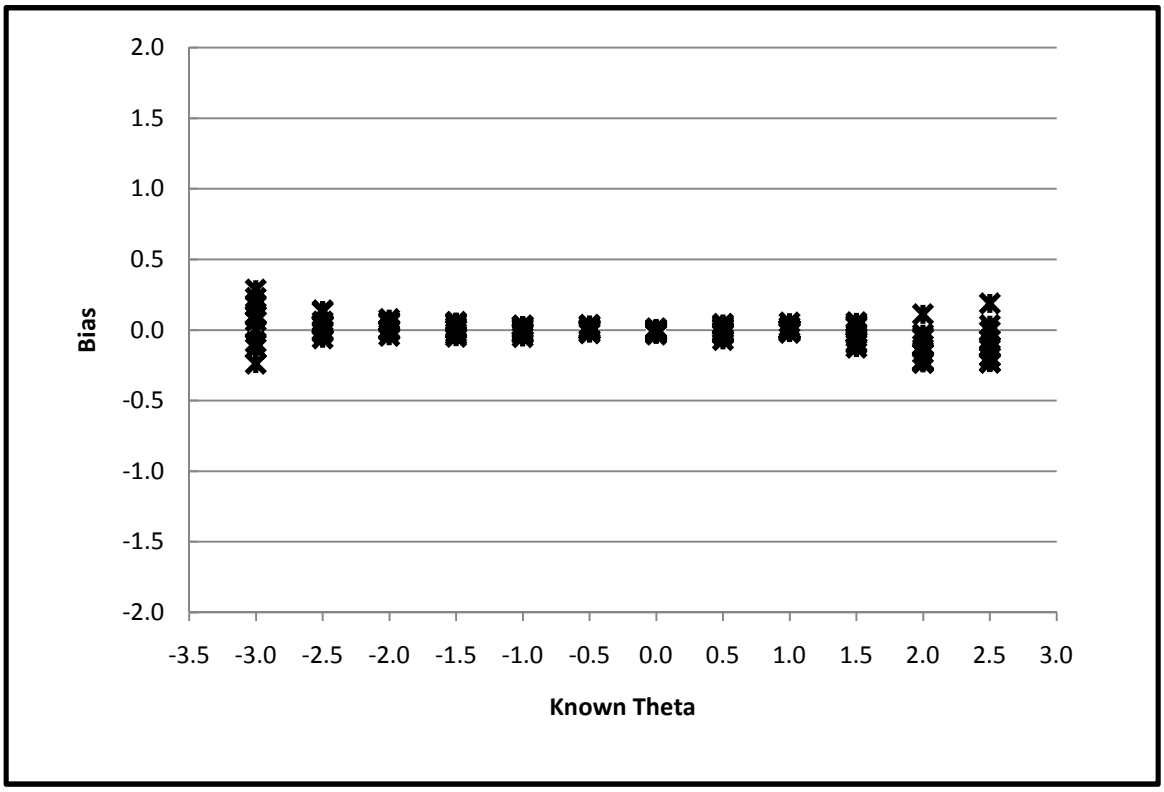

Figure 24. Conditional Bias Plot for CAT

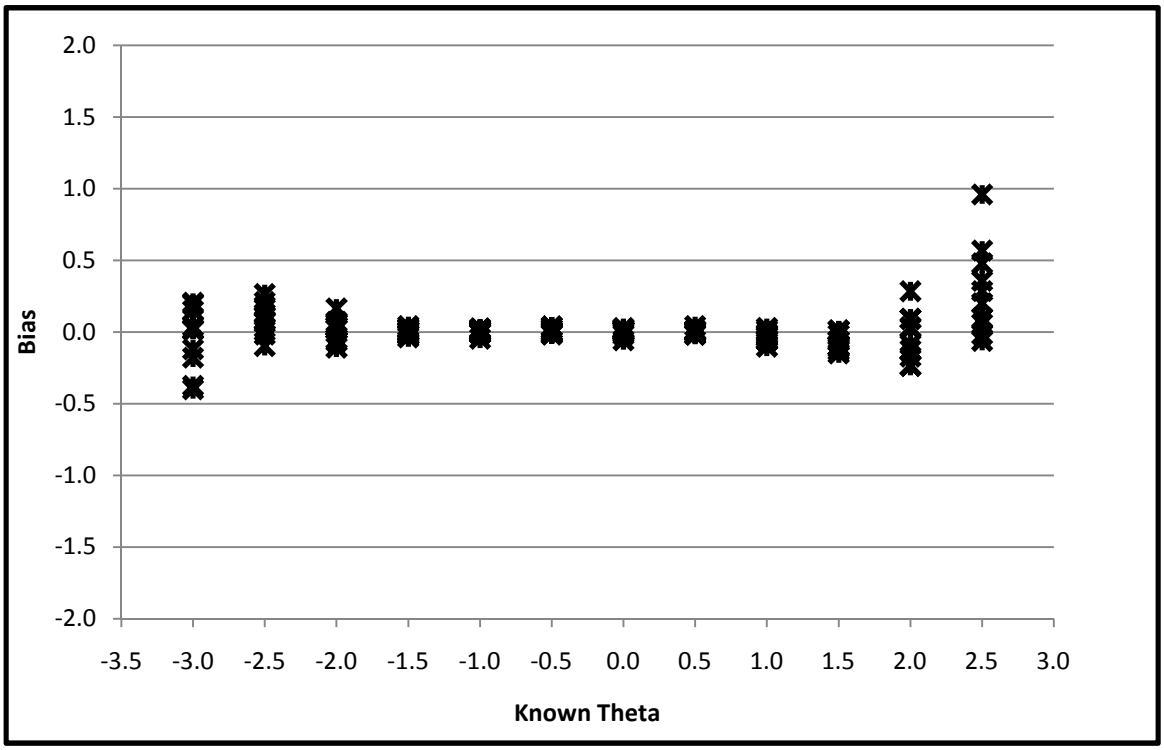

Figure 25. Conditional Bias Plot for MST I Maximum Information

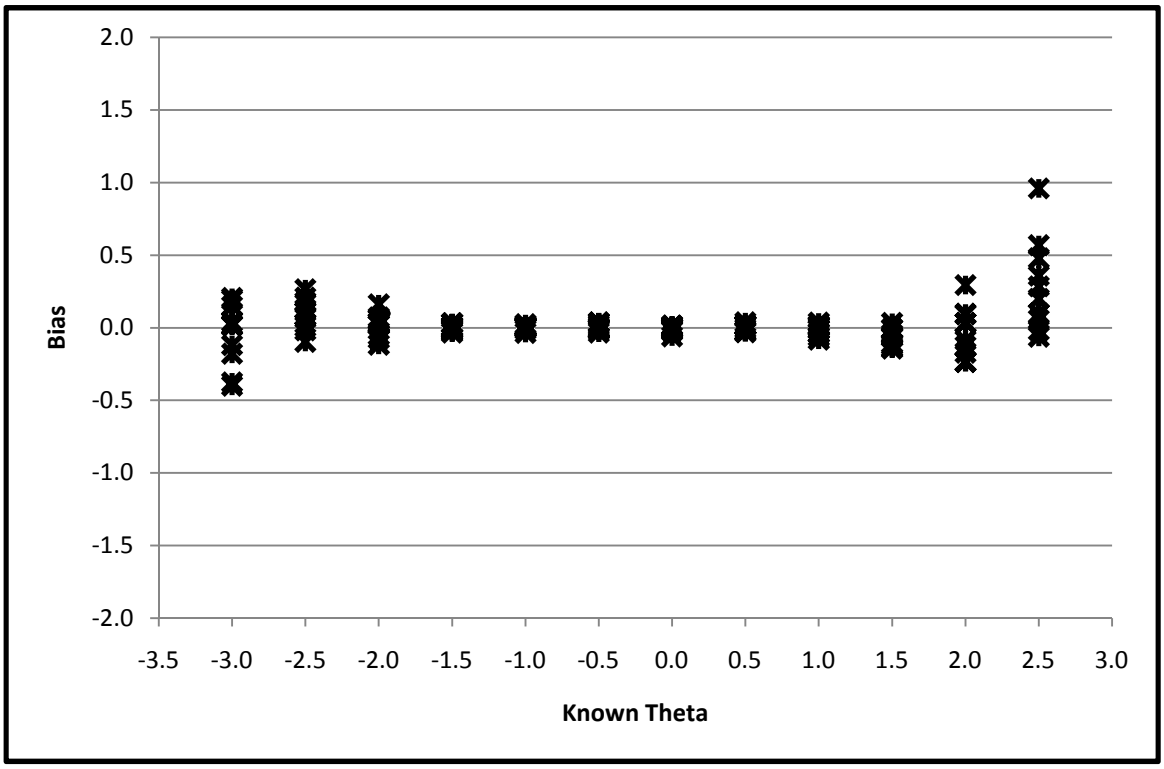

Figure 26. Conditional Bias Plot for MST I Fixed θ

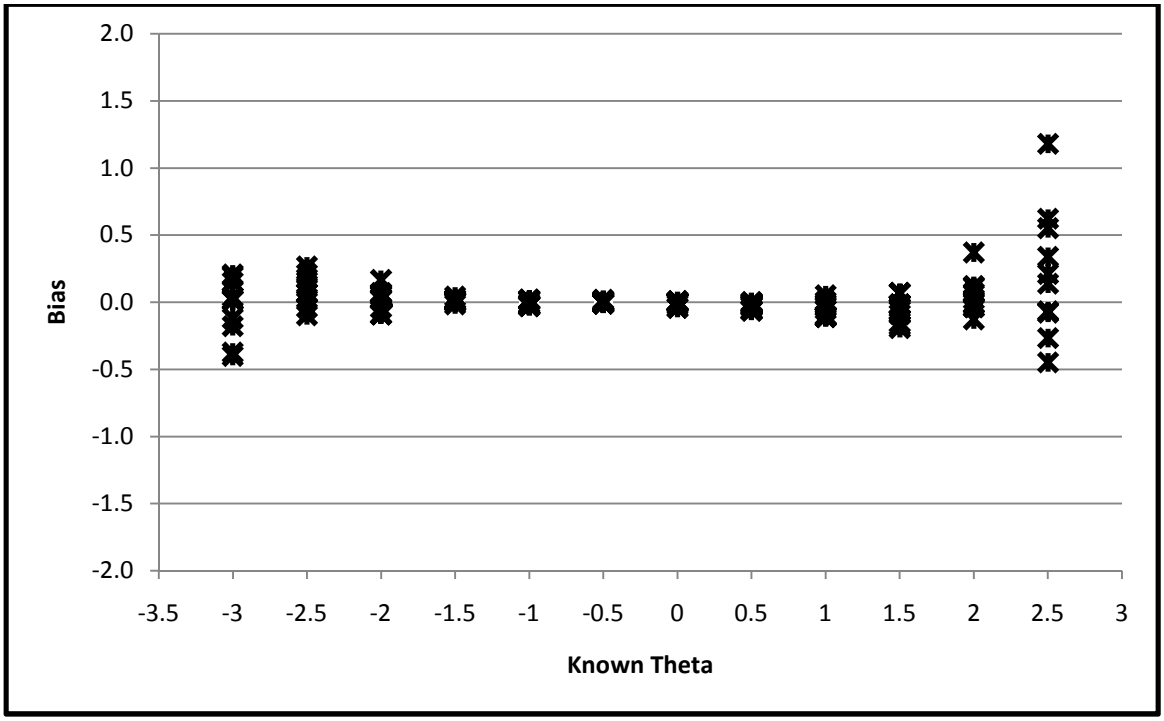

Figure 27. Conditional Bias Plot for MST I Number-Right

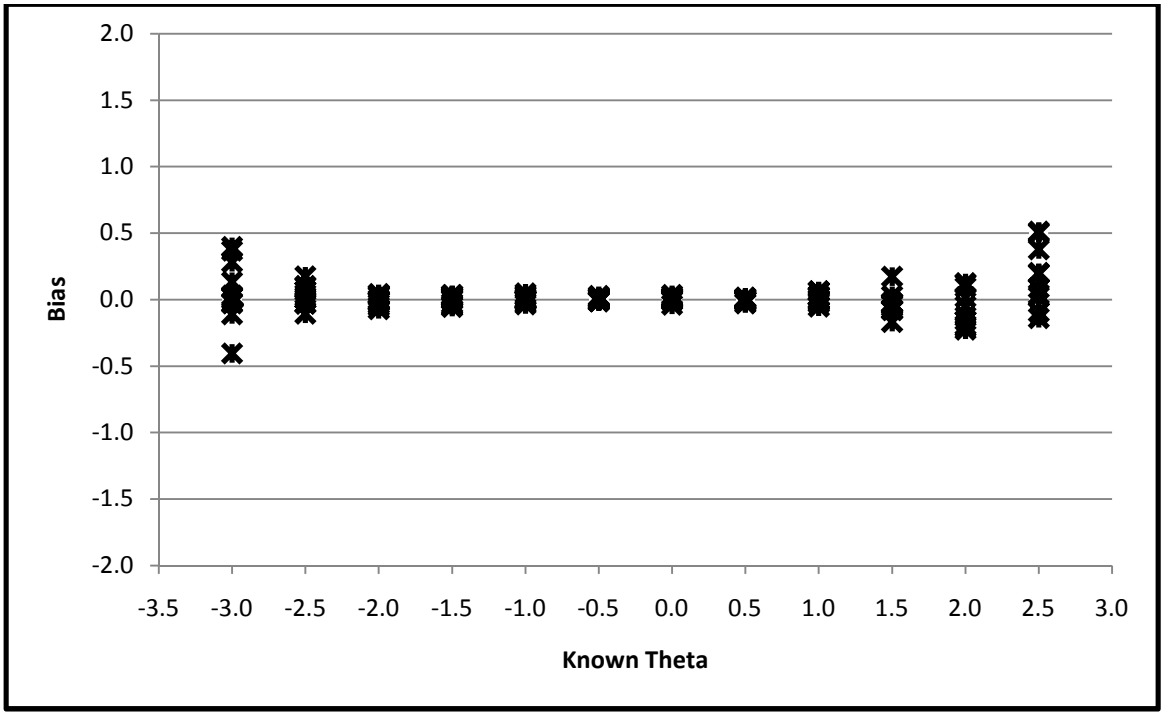

Figure 28. Conditional Bias Plot for MST II Maximum Information

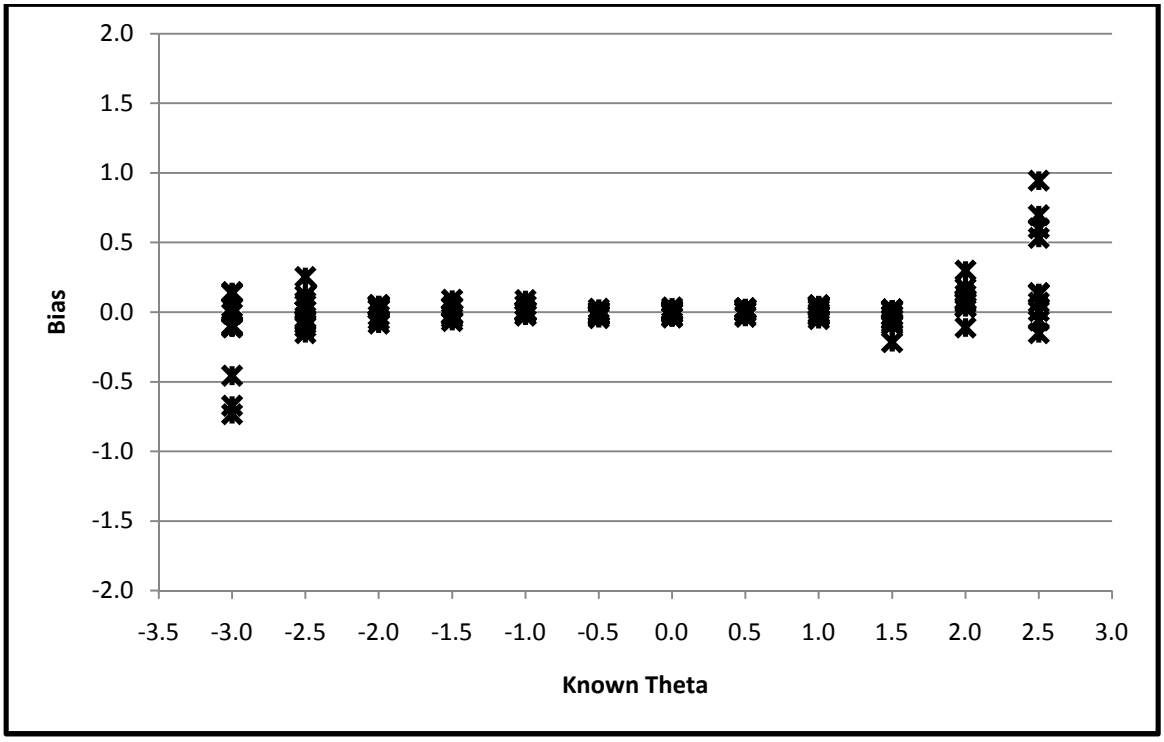

Figure 29. Conditional Bias Plot for MST II Fixed θ

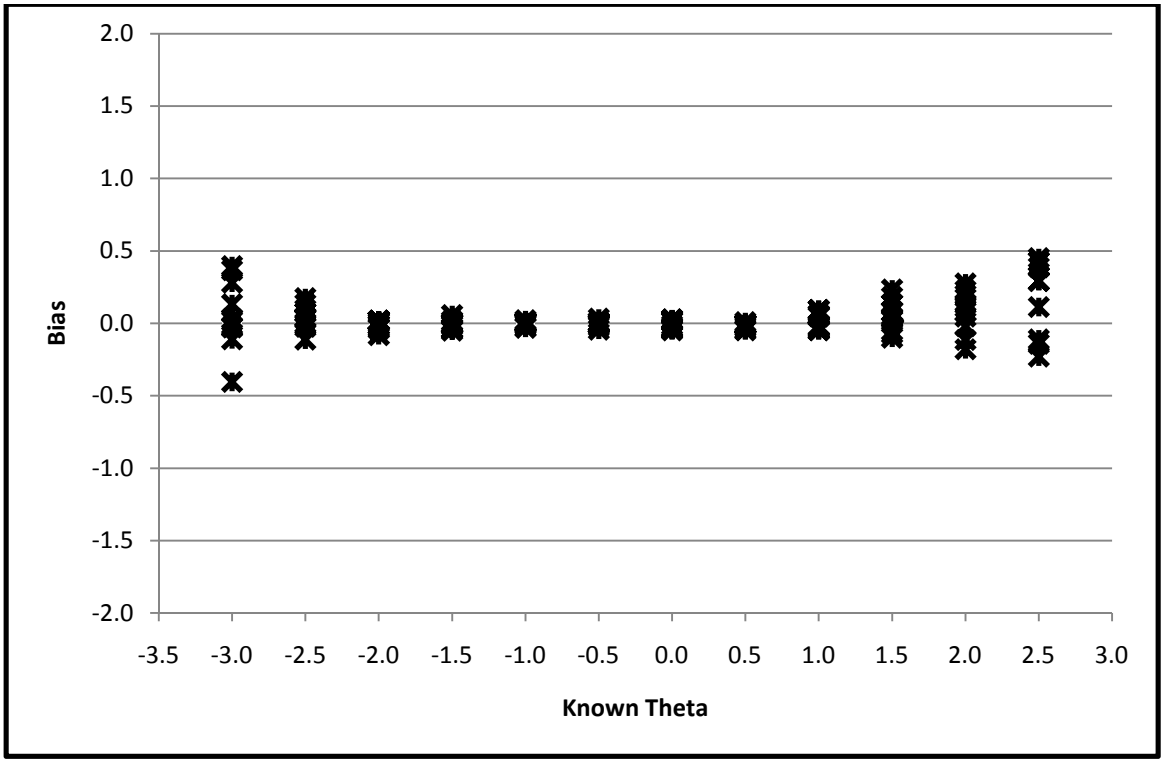

Figure 30. Conditional Bias Plot for MST II Number-Right

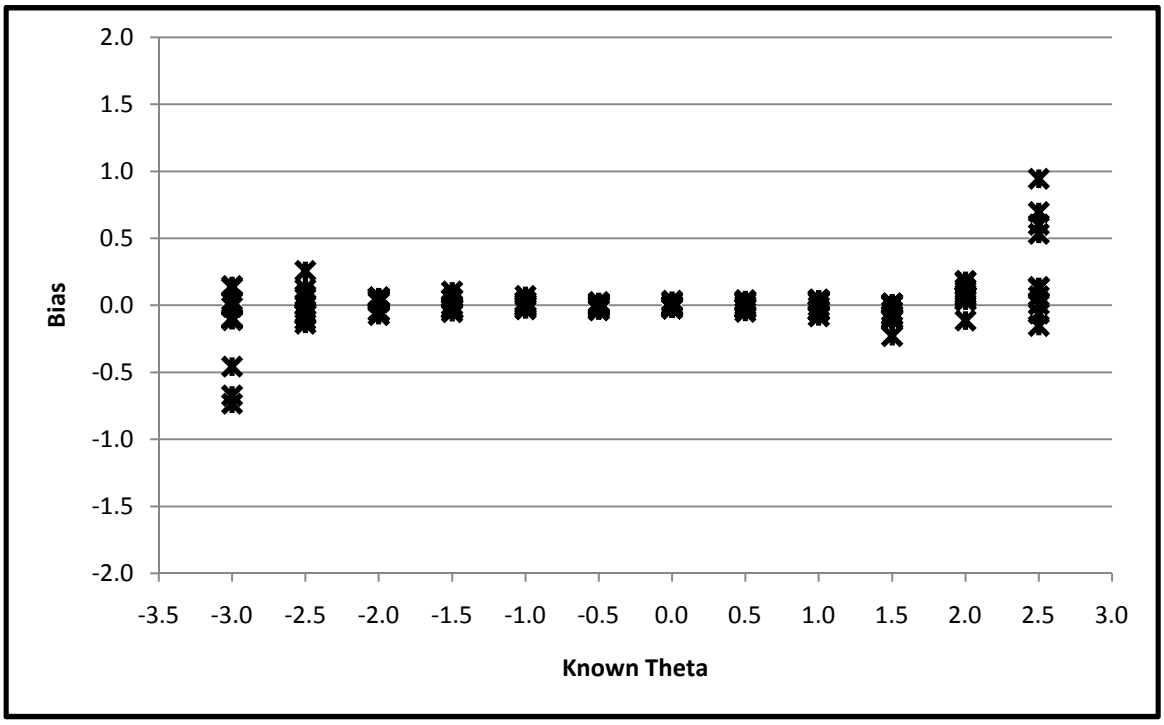

Figure 31. Conditional Bias Plot for MST III Maximum Information

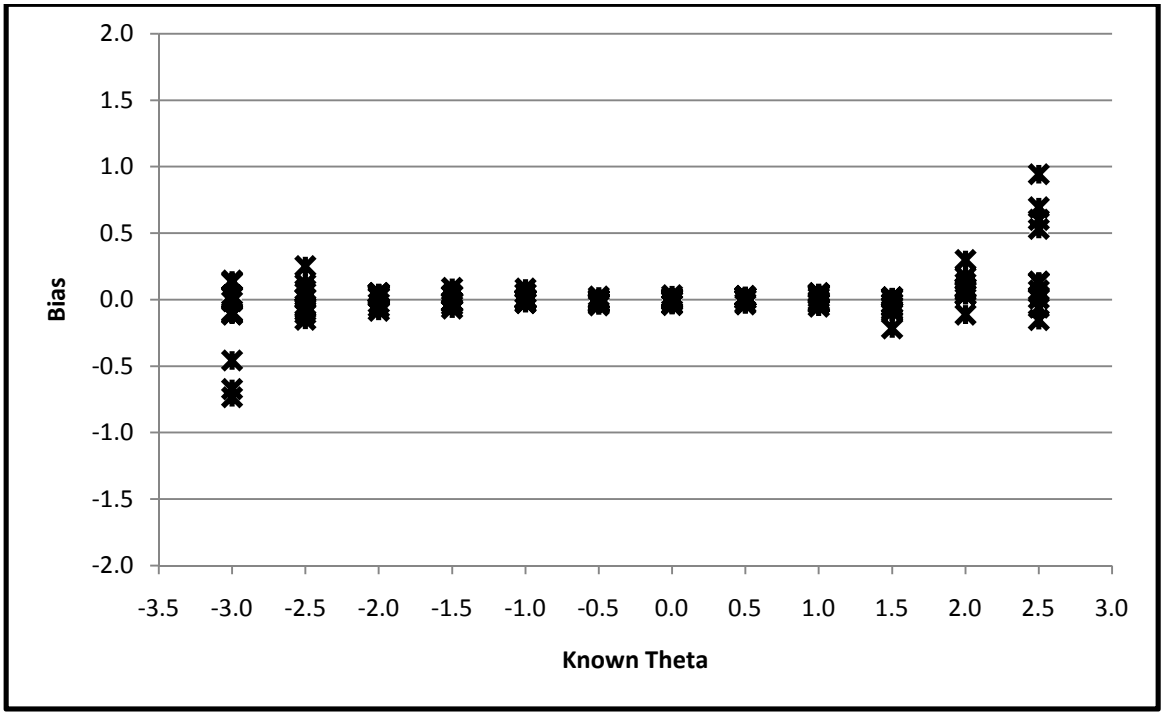

Figure 32. Conditional Bias Plot for MST III Fixed θ

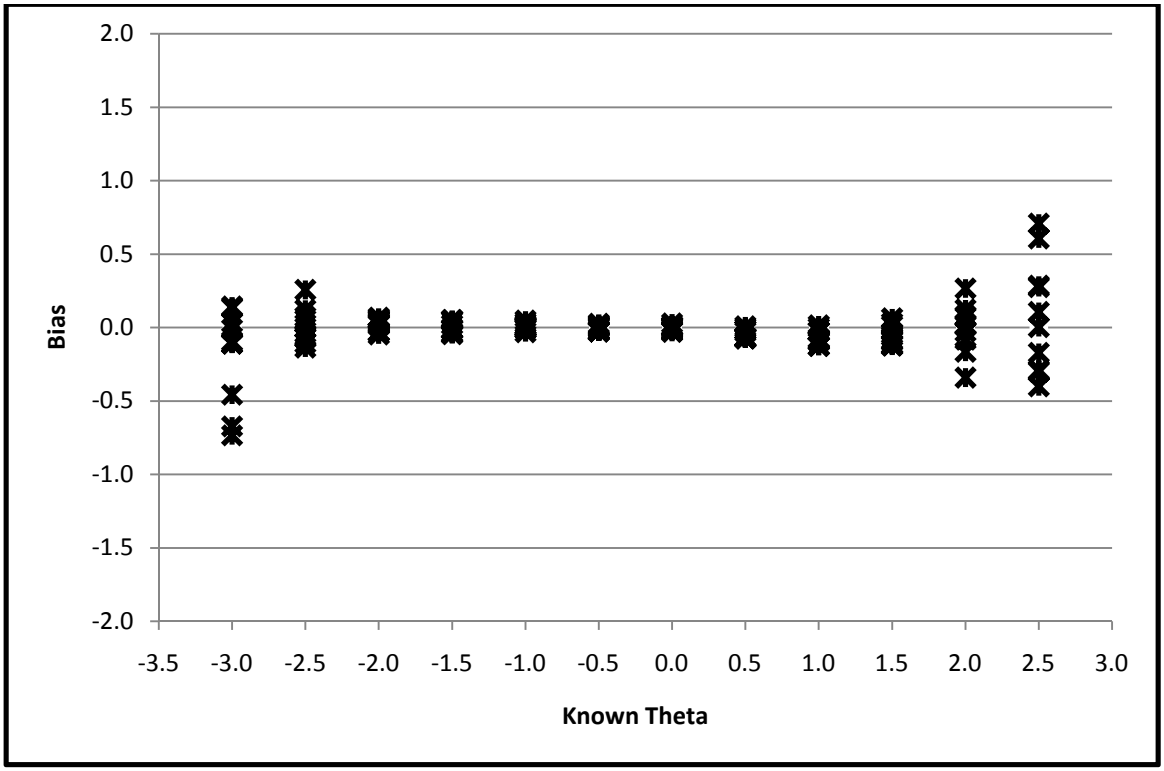

Figure 33. Conditional Bias Plot for MST III Number-Right

CHAPTER 5: DISCUSSION

This chapter comprises three sections. The first of which broadly recounts what has been learned from this research and provides perspective on how it relates to both CAT and the MST designs. The second discusses each of the research hypotheses and provides conclusions that are drawn from the results. The third covers study limitations and provides suggestions for future research.

Item Pool. An item pool of sufficient size is necessary to address exposure control considerations, as well as other nonstatistical properties such as content and item type. This research has shown the difficulties encountered when there are not enough items of certain types to meet content specifications. To meet the requirements of the PR30 exposure control procedure, the item pool was inflated to provide more of the item types that were in short supply for the purposes of CAT. The MST designs did not suffer to the same extent as CAT because items were selected by hand with intent to minimize the exposure of items while still meeting the content specifications. An obvious conclusion that should be drawn from this is that the makeup of the item pool with respect to nonstatistical item properties should be taken into account when writing test specifications. It must be noted, however, that the design of the CAT resulted in potentially 1,000 different tests, one for each simulee, from the same item pool whereas only three different MST tests were created in this research and administered to 1,000 simulees under each MST condition. Had multiple panels been created under the MST designs, similar problems may have been encountered. Furthermore, given that the MST

panels were constructed by hand the item pool might not have been as efficiently used as it was in the CAT paradigm.

Nonconvergence and Test Statistics. The average number of cases of nonconvergence was similar across conditions with the exception of the MST II design under the number-right score routing condition (7.1 average cases). For all other conditions, the average number of convergent cases ranged from 1.1 to 2.6, with the MST I design under the maximum information and θ routing rules and the MST III design under the number-right scoring rule having with the fewest number of nonconvergent cases. Given that the average number of nonconvergent cases was low overall, however, the differences appeared to have little impact on the resulting statistics.

CAT yielded the highest correlation between known and estimated θ followed by the MST II and MST I designs, respectively, under the maximum information and θ routing rules. Recall that the MST II design had the largest number of items at the later stages of the test, and the MST I design had the largest number of items at the earlier test stages. The MST III design, with equal number of items at each stage, produced the next highest correlations under the maximum information and θ routing rules. In all of the MST designs the number-right routing rule yielded the lowest correlations and supports the assertion that the number-right score is not a sufficient statistic for ability estimation when item discrimination is not equal across an item pool (Lord, 1980).

The MST II design using maximum information routing produced the least bias, and the MST III design under the number-right routing rule produced the most. Within all of the MST designs, the number-right scoring rule yielded the highest amount of bias.

With regard to the other two error statistics, RMSE and absolute bias, a similar pattern was observed: CAT produced the lowest error followed by the MST II, MST I, and MST III designs, respectively. Maximum information and θ routing, in that order, produced the next highest error in each of the MST designs. The number-right routing rule had the highest error within each of the MST designs.

Item Exposure and Item Bank Utilization. Since they were fixed length tests, average item exposure was similar for CAT, MST I, and MST II. The average item exposure was somewhat less for MST III because it was shorter than the other tests by two items. While the CAT PR30 exposure control method posed initial problems in that the original item pool did not contain the necessary number of item types to satisfy the requirements of the method, the use of PR30 ensured efficient and controlled item bank utilization. Item exposure was limited to a maximum of 30% for the most part, and the lowest item exposure ranged from 1 to 5%. Although the original item pool was sufficient to construct the MSTs manually, item bank utilization was not as efficient. Due in part to the fact that only one panel was constructed under each of the designs and the same panel was used for each simulee, maximum exposure rates were as high as 100% for a minority of items and a large percentage of the item pool was not used at all.

Research Questions. At the outset of this research, three questions were posed. In this section each question will be presented and answered.

How do MST designs compare to CAT in the recovery of proficiency estimates, item pool utilization, and item exposure assuming a normal proficiency distribution?

Although they did not perform as well as CAT, the MSTs performed nearly as well with respect to the recovery of proficiency estimates. The MST that used the maximum information routing method to determine which module to administer in later test stages had the best performance because it took advantage of the dynamic assessment of information to select the next module for administration. The MST that used the fixed θ routing rule had the next best performance. Although it did not have the advantage of dynamic assessment, it utilized the current θ to determine subsequent module selection. The MST that used the number-right routing rule had the worst performance, though it must be noted that the correlation between known and estimated θ was only about 0.02 to 0.03 less than the correlation found in CAT.

Item pool utilization was clearly more efficient in CAT. All of the items in the item pool were used in CAT while a majority of the items were not used in the MST. Rather than being a shortcoming of the MST design, however, it is more likely that this was due to the study design. By design, a CAT is tailored to the examinee with each item selected to maximize proficiency estimation. In theory, a unique test could be administered to each examinee. In this research, only one panel per design type was constructed. Rather than taking advantage of the entire item pool, a predesigned subset of items was administered. It is typical to use more panels in an operational testing program.

Exposure rates were a direct result of item pool utilization. In addition to having the entire item pool available in CAT, the use of the PR30 exposure control method assured that overutilization was kept to a minimum. The MSTs had no such mechanism in place even though there was no repetition of items within a test, and the use of items between tests was kept to a minimum.

To what extent are proficiency estimates affected by a number-right routing rule versus maximum information routing rule in an MST?

Average proficiency estimates tended to be overestimated when number-right routing was used compared to maximum information, and average proficiency estimates were not as well recovered. Average correlations between estimated and known θ were lower by 0.01 in the number-right routing condition.

This research expanded on the original proposal by adding a fixed θ routing rule to the above two routing rules. The fixed θ rule served as a middle ground between maximum information and number-right routing. Recovery of average proficiency estimates was better for the fixed θ rule as compared to the number-right rule, but recovery was not as high as it was under the maximum information routing rule.

Which is the more optimal design for an MST, the same number of items per stage or varying numbers of items per stage?

Every effort was made to maintain equivalence in the amount of test information among the three designs, but there was more information in the MST II design than in the other two. More information may have contributed to more precise estimates in

this design. Even though the amount of test information was virtually the same for MST I and MST III, MST III had fewer items.

Subject to the above considerations, this research found that the optimal design was MST II with more items at later stages. No matter the routing rule, the MST II design had higher correlations between estimated and known θ and lower average RMSE than MST I, with more items at earlier stages, or MST III, with the same number of items at each stage. MST III had the worst performance of the three designs.

Conclusions, Limitations, and Directions for Future Research. This research compared three MST designs and CAT based on the Generalized Partial Credit Model. The CAT used the PR30 exposure control method (Revuelta & Ponsoda, 1998) and content balancing (Kingsbury & Zara, 1989). The MSTs were constructed using the top-down approach whereby items were selected to fit a test information function rather than separate module-level information functions. The MST designs used one of three routing rules and differing numbers of items per stage. Through simulations that utilized maximum likelihood estimation, the three MST designs and the CAT were evaluated based on precision of ability estimation.

As anticipated from previous research (Hambleton & Xing, 2006; Schnipke & Reese, 1997), CAT yielded the best precision of ability estimation and had the most efficient item pool utilization. Among the MST designs, the maximum information routing rule yielded the most precise estimates followed by the fixed θ routing rule and the number-right routing rule. In this research, increasing the number of items per stage

as the test progressed yielded the most precise estimates followed by decreasing the number of items per stage and an equal number of items per stage. This finding is congruent with previous research, which found that increasing the number of points of adaptation yields more precise estimates of proficiency (Lord, 1980, 1971; Luecht, Brumfield & Breithaupt, 2006; Luecht & Nungester, 1998). However, the differences among the CAT and the MST designs are small, leading to the observation that the advantage of one of the designs over the others is of little practical importance. The strongest contraindicator of use of any of these designs with a 3PL model is the utilization of a number-right routing rule, which decreased the efficiency of proficiency estimation in all three MSTs.

It appeared that the advantage of CAT rested more on the ability to more precisely estimate proficiency at the extremes of the distribution rather than better estimation across the distribution as compared to the MST designs, however. An examination of plots of mean bias conditional on known θ showed that the differences were smaller for all MST designs than they were for CAT especially in the central portion of the distribution, but the MST designs showed more bias than CAT at the extremes. Similarly, the plots of mean standard errors conditional on known θ were narrower in the middle of the distribution in the MST designs than in CAT, but the standard errors were more widely dispersed at the upper end of the distribution in the MST designs. This was similar to the finding by Schnipke & Reese (1997) that CAT yielded the least bias and estimation error at the ends of the distribution when compared to a paper-and-pencil test and two MST designs.

A possible limitation to this research was that the MST designs were not designed with the same number of items. MSTs I and II used 20 items, and MST III used 18 items. A second limitation was the creation of only one panel per design. This, in effect, resulted in comparing the results from a 208 item pool used in the CAT simulations to three 20 item pools in the MST simulations.

An obvious suggestion for future research would be to increase the number of panels per stage so that comparisons of item pool use between CAT and MST could be made. This research showed that among the MST designs, the maximum information routing rule used the item pool most efficiently whereas the other MST designs were more erratic. Only limited comparisons between MST and CAT could be made, however, since only one panel per MST design was constructed. Another direction may be investigating the ways to increase the precision of proficiency estimation at the extremes of the proficiency distribution in MST. The MST designs seemed to be more precise than CAT in the middle of the distribution but were outperformed by CAT at the extremes, which on average yielded better estimation for CAT. Among the ways to achieve this may be to use a bottom-up approach to test design that would concentrate on maximizing module-level information functions, strategically place the more informative items either earlier or later in the test, or perhaps use more modules per stage to increase the number of adaptation points without increasing the number of stages. Lastly, increasing the number of adaptation points by increasing the number of stages should be compared to increasing the number of adaptation points by increasing the number of modules per stage.

APPENDIX: ITEM POOL PARAMETERS FOR 208 ITEMS

a	b1	b2	b3	b4
1.26896	-1.88782	-0.26746		
1.00294	-1.69539	0.39325		
1.20666	-1.44836	0.2268		
1.26896	-1.88782	-0.26746		
1.14728	-1.5944	-0.1044		
0.99677	-1.53461	0.17483		
0.99012	-1.91167	0.51627		
1.30003	-1.07668	0.57728		
0.9292	-1.50797	-0.46655		
1.17071	-2.03572	-0.68564		
1.13622	-0.87285	0.36067		
1.46912	-1.17593	0.21479		
1.11538	-1.51209	0.36047		
0.83919	-1.85554	0.29492		
1.1651	-0.19985	1.39689		
0.76211	-2.51139	-0.05825		
0.92638	-2.13108	0.87678		
1.25695	-1.05348	0.29198		
0.87917	-1.26528	1.49852		
0.79378	-1.22877	-0.35169		
1.51894	-1.66146	-0.51892		
1.02104	-1.27753	0.69195		
1.09664	-1.1194	0.21004		
1.35429	-0.55027	0.63721		
0.95751	-2.10248	-0.8671		
1.1809	-1.81049	-0.40065		
0.8574	0.65249	2.97987		
0.74797	-1.00514	-0.1827		
0.81937	-0.87158	0.32232		
1.37115	-1.90615	-0.30797		
0.81056	-2.35579	-1.20095		
1.05167	-0.72453	1.70519		
1.08328	-0.21329	1.56579		
0.77614	-0.67093	0.03511		
0.95029	-1.74314	-0.9321		
1.09571	-2.35511	-0.57033		
0.85696	-1.62779	0.70587		
0.86188	-0.08083	2.20849		
0.93463	-0.70733	0.29255		

a	b1	b2	b3	b4
1.17333	-1.79634	-0.92122		
1.07837	-1.92016	-0.13712		
1.0523	0.09202	1.60966		
0.94781	-0.17044	1.50496		
0.80935	-2.65267	-0.34577		
1.07825	-1.98076	0.10488		
1.17263	-0.91388	0.64182		
0.84797	-0.42875	2.21011		
0.94235	-2.26398	-0.76542		
1.07888	-0.94033	0.78497		
0.91086	-0.48723	1.86509		
0.76673	-0.92262	1.62756		
0.89797	-0.40157	1.23629		
1.04327	-1.82437	0.15109		
1.08708	-1.4905	031298		
1.01724	-1.28366	049914		
0.84476	-1.19056	0.73216		
1.00209	-0.58964	1.24712		
1.04461	-1.41834	0.5584		
1.07346	-0.9227	0.36164		
0.99503	-0.66249	0.95177		
0.91509	-0.22166	1.4722		
1.04822	-2.21589	-1.45229		
1.03027	-0.67721	0.77223		
1.26983	-0.27093	1.43433		
0.80362	-0.75802	1.3112		
1.02784	-1.74105	-1.81443		
0.80362	-0.75802	1.3112		
1.02784	-1.74105	-1.81443		
0.89174	-1.40101	0.37479		
1.51743	-0.87027	0.48821		
0.91089	-0.56611	1.50581		
1.0082	-1.39639	0.38909		
1.01723	-1.47845	0.47243		
0.9887	-1.20875	1.08747		
1.1288	-0.89004	0.9478		
0.90264	-0.39981	1.29631		
0.78567	-0.60109	1.11717		
0.97345	-0.59543	1.02483		
1.0027	-0.69066	1.08034		
0.90553	-0.43772	1.7506		
1.04761	-2.16862	-0.00678		
1.14522	-1.59367	-0.26557		

a	b1	b2	b3	b4
0.78567	-0.60109	1.11717		
0.97345	-0.59543	1.02483		
1.0027	-0.69066	1.08034		
0.90553	-0.43772	1.7506		
1.04761	-2.16862	-0.00678		
1.01593	-0.41502	1.29992		
0.98048	-0.51462	0.17988		
1.0092	-1.3496	0.4766		
1.05128	-1.07301	0.56475		
0.77545	-0.6313	1.86048		
0.6885	0.98096	3.56874		
0.68571	-2.82657	-0.28175		
0.91929	-1.73743	0.19849		
0.8481	-1.41659	1.08579		
0.8463	-0.42614	1.52972		
0.79736	0.73146	2.70044		
0.99081	-2.38197	-0.63769		
1.10192	-1.99727	-0.12137		
0.92631	-0.53242	0.889		
1.16893	0.20841	1.41967		
0.8685	-2.67267	-0.46793		
1.02727	-1.0172	0.9559		
0.8436	-1.6005	0.43024		
0.89142	-0.06885	1.88171		
1.0553	-0.59147	0.72161		
0.85171	-0.80643	-1.09665	-1.47687	
1.0088	-0.08751	-1.27945	-1.36244	
0.67832	-0.3044	-0.97649	-1.1921	
0.66028	-0.91032	-0.73444	-0.94466	
0.88521	-0.80066	-1.10219	-1.01131	
0.85171	-0.80643	-1.09665	-1.47687	
1.0088	-0.08751	-1.27945	-1.36244	
0.67832	-0.3044	-0.97649	-1.1921	
0.66028	-0.91032	-0.73444	-0.94466	
0.88521	-0.80066	-1.10219	-1.01131	
0.79867	-0.87511	-1.2461	-0.8757	
1.019	-0.70827	-1.07186	-0.78709	
0.66098	-1.06984	-1.00168	-0.76817	
0.60047	1.50207	-0.42778	-0.64246	
0.95365	-0.70642	-0.80831	-0.52452	
0.89003	-0.34928	-0.63366	-0.32431	
0.89949	-0.63956	-0.68847	-0.31627	
0.8112	-0.55984	-0.74122	-0.23095	

a	b1	b2	b3	b4
0.63035	0.57499	-0.67578	-0.24547	
0.67672	1.03838	-0.11362	-0.05239	
0.83089	0.48207	-0.24135	-0.0931	
0.7783	-0.0256	-0.07628	-0.01166	
0.68614	-0.35945	-0.5313	0.25397	
0.73453	-0.44914	-0.22937	0.35001	
0.75409	-0.08912	0.07972	0.35881	
0.63824	1.04903	-0.16117	0.44633	
0.69757	-0.05145	-0.09017	0 .41965	
0.68853	-0.93943	-0.49938	0.64865	
1.019	-0.70827	-1.07186	-0.78709	
0.66098	-1.06984	-1.00168	-0.76817	
0.60047	1.50207	-0.42778	-0.64246	
0.95365	-0.70642	-0.80831	-0.52452	
0.89003	-0.34928	-0.63366	-0.32431	
0.89949	-0.63956	-0.68847	-0.31627	
0.8112	-0.55984	-0.74122	-0.23095	
0.63035	0.57499	-0.67578	-0.24547	
0.67672	1.03838	-0.11362	-0.05239	
0.83089	0.48207	-0.24135	-0.0931	
0.7783	-0.0256	-0.07628	-0.01166	
0.68614	-0.35945	-0.5313	0.25397	
0.73453	-0.44914	-0.22937	0.35001	
0.75409	-0.08912	0.07972	0.35881	
0.63824	1.04903	-0.16117	0.44633	
0.69757	-0.05145	-0.09017	0 .41965	
0.68853	-0.93943	-0.49938	0.64865	
0.68952	0.90671	0.22468	0.58599	
0.71672	0.26124	0.32902	0.95279	
0.62891	1.05523	0.63408	1.18772	
0.66522	0.03439	0.13297	1.25887	
0.57558	-0.08613	0.61087	1.24899	
0.86478	1.09147	0.82349	1.51098	
0.60692	-0.26124	-1.4483	0.92072	-2.36198
0.86611	-1.11308	-1.73714	-1.25179	-1.28363
0.83685	-1.25601	-1.48188	-0.92566	-1.09533
0.98115	-0.97162	-1.39341	-1.30434	-0.87711
0.96171	-2.14813	-0.96711	-0.72315	-0.68086
0.98184	-2.16192	-0.70021	-0.78053	-0.58966
0.66992	-1.99762	-0.39888	-0.77744	-0.63538
0.74693	-2.43057	-0.72329	-0.81042	-0.6308
0.67272	-2.5996	-0.71435	-0.72227	-0.61158
0.91176	-1.17969	-1.26124	-0.88388	-0.53287

a	b1	b2	b3	b4
0.75118	-2.49894	-0.69841	-0.69815	-0.41206
0.65723	-1.79487	-0.61074	-0.29875	-0.45284
0.83883	-1.73412	-0.52159	-0.74524	-0.40754
0.77727	-2.32006	-1.00844	-0.89133	-0.38601
0.82402	-2.04604	-0.55377	-0.3546	-0.22815
0.65924	-3.13156	-0.75446	-0.46752	-0.20423
0.80385	-0.71676	-0.87966	-0.42454	-0.27872
0.75118	-2.49894	-0.69841	-0.69815	-0.41206
0.65723	-1.79487	-0.61074	-0.29875	-0.45284
0.83883	-1.73412	-0.52159	-0.74524	-0.40754
0.77727	-2.32006	-1.00844	-0.89133	-0.38601
0.82402	-2.04604	-0.55377	-0.3546	-0.22815
0.65924	-3.13156	-0.75446	-0.46752	-0.20423
0.80385	-0.71676	-0.87966	-0.42454	-0.27872
0.69796	-1.0913	-1.06383	-0.66754	0.08308
0.68187	-0.46739	-0.97975	0.34757	0.18502
0.72091	-1.32802	-1.11603	-0.10763	0.03936
0.8163	-1.52395	-0.29794	-0.49393	0.17389
0.64729	-0.61121	-0.67307	0.02281	0.32339
1.01198	-0.87659	0.12461	0.14859	0.48312
0.66578	0.7157	-0.42684	0.37317	0.41733
0.65847	-0.4061	-0.38419	0.30334	0.95127
0.92712	-0.97437	0.44477	0.39846	0.93442
0.75991	-1.55006	0.509	0.54239	1 09434
0.77218	-0.07815	0.86878	1.11343	1.25214
0.53765	-1.36519	1.45029	1.3348	2.34161
0.83883	-1.73412	-0.52159	-0.74524	-0.40754
0.77727	-2.32006	-1.00844	-0.89133	-0.38601
0.82402	-2.04604	-0.55377	-0.3546	-0.22815
0.65924	-3.13156	-0.75446	-0.46752	-0.20423
0.80385	-0.71676	-0.87966	-0.42454	-0.27872
0.69796	-1.0913	-1.06383	-0.66754	0.08308
0.68187	-0.46739	-0.97975	0.34757	0.18502
0.72091	-1.32802	-1.11603	-0.1076	0.03936
0.8163	-1.52395	-0.29794	-0.4939	0.17389
0.64729	-0.61121	-0.67307	0.02281	0.32339
1.01198	-0.87659	0.12461	0.14859	0.48312
0.66578	0.7157	-0.42684	0.37317	0.41733
0.65847	-0.4061	-0.38419	0.30334	0.95127
0.92712	-0.97437	0.44477	0.39846	0.93442
0.75991	-1.55006	0.509	0.54239	1 09434
0.77218	-0.07815	0.86878	1.11343	1.25214
0.53765	-1.36519	1.45029	1.3348	2.34161

REFERENCES

Adema, J.J. (1990). The construction of two-stage tests. *Journal of Educational Measurement, 27*(3), 241-253.

Ariel, A., Veldkamp, B.P., & Breithaupt, K. (2006). Optimal testlet pool assembly for multistage testing designs. *Applied Psychological Measurement, 30*(3), 204-215.

Armstrong, R.D., Jones, D.H., Koppel, N.B., & Pashley, P.J. (2004). Computerized adaptive testing with multiple-form structures. *Applied Psychological Measurement, 28*(3), 147-164.

Baker, F. (2001). *The Basics of Item Response Theory.* Retrieved October 26, 2006 from http://edres.org/irt/baker.

Bergstrom, B. A. & Lunz, M. E. (1999). *Innovations in Computerized Assessment.* (F. Drasgow & J. B. Olson-Buchanan, Eds.) Mahwah, New Jersey: Lawrence Erlbaum Associates.

Birnbaum, A. (1968). *Statistical Theories of Mental Test Scores.* (F.M. Lord & M.R. Novick, Eds.). Reading, Massachusetts: Addison-Wesley Publishing Company.

Boekkooi-Timminga E. (1987). Simultaneous test construction by zero-one programming. *Methodika, 1*(2), 101-112.

Boekkooi-Timminga E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics, 15*(2), 129-145.

Boyd, A.M. 2003. *Strategies for controlling testlet exposure rates in computerized adaptive testing systems.* Unpublished doctoral dissertation, University of Texas, Austin.

Breithaupt, K., Ariel, A., & Veldkamp, B.P. (2005). Automated simultaneous assembly for multistage testing. *International Journal of Testing, 5*(3), 319-330.

Chang, S.W., & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40(1),* 71-103.

Chen, S. (1997). A comparison of maximum likelihood estimation and expected a priori estimation in computerized adaptive testing using the generalized partial credit model. *Dissertation Abstracts International, 58,* 453.

Chen, S-Y., Ankenmann, R.D., & Chang, H-H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24*(3), 241-255.

Chuah, S.C., Drasgow, F., & Luecht, R. (2006). How big is big enough? Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education, 19*(3), 241-255.

Davey, T., & Parshall, C.G. (1995, April). *New Algorithm for Item Exposure and Exposure Control with Computerized Adaptive Testing.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California.

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

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.

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

Du, Y., Lewis, C., & Pashley, P.J. (1993). Computerized mastery testing using fuzzy set theory. *Applied Measurement in Education, 6*(3), 181-193.

Eignor, D. (2000). *Computerized Adaptive Testing: A Primer.* (2nd ed.). (H. Wainer, Ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Gershon, R. C. (2005). Computer Adaptive Testing. *Journal of Applied Measurement, 6(1)*, 109-127.

Glas, C.A.W. (1988). The Rasch model and multistage testing. *Journal of Educational Statistics, 13*(1), 45-52.

Gorin, J.S., Dodd, B.G., Fitzpatrick, S.J., & Shieh, Y.Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. *Applied Psychological Measurement, 29*(6), 433-456.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L., & Reckase, M. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347-360.

Hambleton, R.K., Swaminathan H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, California: Sage Publications.

Hambleton, R.K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education, 19*(3), 221-239.

Hetter, R.D. & Sympson, J. B. (1997). *Computerized Adaptive Testing From Inquiry to Operation.* (W.A. Sands, B.K. Waters, & J. R. McBride, Eds.) American Psychological Association: Washington, D.C.

Huitzing, H.H., Veldkamp, B.P., & Verschoor, A.J. (2005). Infeasibility in automated test assembly models: a comparison study of different methods. *Journal of Educational Measurement, 42*(3), 223-243.

Jodoin, M.G., Zenisky, A., & Hambleton, R.K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education, 19*(3), 203-220.

Kingsbury, G.G. & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-374.

Kingsbury, G.G. & Zara, A.R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4*(3), 241-261.

Koch, W.R. & Dodd, B.G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education, 2*(4), 335-357.

Lewis, C. & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*(4), 367-386.

Lord, F. M. (1971a). The self-scoring flexilevel test. *Journal of Educational Measurement, 8(3),* 147-151.

Lord, F. M. (1971b). A theoretical study of two-stage testing. *Psychometrika, 36*(3), 227-242.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Luecht, R.M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement,* 22(3), 224-236.

Luecht, R.M. (2000, April). *Implementing the Computer-Adaptive Sequential Testing (CAST) Framework to Mass Produce High Quality Computer-Adaptive and Mastery Tests.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, Louisiana.

Luecht, R.M. (2003, April). *Exposure Control Using Adaptive Multi-Stage Item Bundles.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois.

Luecht R., Brumfield, T., & Breithaupt K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education, 19*(3), 189-202.

Luecht, R.M. & Burgin, W. (2003, April). *Test Information Targeting Strategies for Adaptive Multistage Testing Designs.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois.

Luecht, R.M., de Champlain, A., & Nungester, R. (1998). Maintaining content validity in computerized adaptive testing. *Advances in Health Science Education, 3,*29-41.

Luecht, R.M. & Hirsch, T.M. (1992). Item selection using an average growth approximation of target information functions. *Applied Psychological Measurement, 16*(1), 41-51.

Luecht, R.M. & Nungester, R.J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*(3), 229-249.

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

Mead, A.D. (2006). An introduction to multistage testing. *Applied Measurement in Education, 19*(3), 185-187.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*(2), 177-195.

Muraki, E. (1990). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

Muraki, E., & Bock, R. D. (1993). PARSCALE (Version 3.0) [Computer program]. Chicago: Scientific Software International.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47*(4), 90-100.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*(350), 351-356.

Parshall, C. G., Davey, T., & Nering M.L. (1998, April). *Test Development Exposure Control for Adaptive Testing.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, California.

Parshall, C.G., Spray, J.A., Kalohn, J.C., & Davey, T. (2002). *Practical Considerations in Computer-Based Testing.* New York, New York: Springer-Verlag.

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

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

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

Reese, L.M., Schnipke, D.L., & Luebke, S.W. (1999). *Incorporating Content Constraints into a Multi-Stage Adaptive Testlet Design.* Newtown, Pennsylvania: Law School Admissions Council.

Revuelta, J. & Ponsoda,V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(4), 311-327.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2).

Schnipke, D.L., & Reese, L.M. (1997, March). *A Comparison of Testlet-Based Designs for Computerized Adaptive Testing.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois.

Stark, S. & Chernyshenko, O.S. (2006). Multistage testing: widely or narrowly applicable? *Applied Measurement in Education, 19*(3), 257-260.

Stocking, M. L., & Lewis, C. (1998). Controlling Item Exposure Conditional on Ability in Computerized Adaptive Testing. *Journal of Educational and Behavioral Statistics, 23(1),* 57-75.

Stocking, M. L. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17(3),*277-292.

Stocking, M.L., Swanson, L., & Pearlman, M. (1993). Application of an automated item selection method to real data. *Applied Psychological Measurement, 17*(2), 167-176.

Swanson, L. & Stocking, M. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*(2), 151-166.

Theunissen, T.J.J.M. (1985). Binary programming and test design. *Psychometrika, 50*(4), 411-420.

Theunissen, T.J.J.M. (1986). Some applications of optimization algorithms in test design and adaptive testing. *Applied Psychological Measurement, 10*(4), 381-389.

Thissen, D. & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567-577.

van der Linden, W.J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22*(3), 195-211.

van der Linden, W.J. (2000). Optimal assembly of tests with item sets. *Applied Psychological Measurement, 24*(3), 225-240.

van der Linden, W.J. (2002). *Some Alternatives to Sympson-Hetter Item-Exposure Control in Computerized Testing.* Enschede, The Netherlands: University of Twente.

van der Linden, W.J. & Adema, J.J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement, 35*(3), 185-198.

van der Linden, W.J. & Boekkooi-Timminga, E. (1988). A zero-one programming approach to Gulliksen's matched random subtests method. *Applied Psychological Measurement, 12*(2), 201-209.

van der Linden, W.J. & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika,* 54(2), 237-247.

van der Linden, W.J. & Glas, C.A. W. (Eds.) (2000). *Computerized Adaptive Testing: Theory and Practice.* Boston, Massachusetts: Kluwer Academic Publishers

van der Linden, W.J. & Hambleton, R.K. (Eds.). (1997). *Handbook of Modern Response Theory.* New York: Springer-Verlag.

van der Linden, W.J. & Luecht, R.M. (1994). *An Optimization Model for Test Assembly to Match Observed-Score Distributions.* Enschede, The Netherlands: University of Twente.

van der Linden, W. J. & Pashley, P. J. (2000). *Computerized Adaptive Testing Theory and Practice.* (W. J. van der Linden and C. A. W. Glas, Eds.) Dordrecht, The Netherlands: Kluwer Academic Press.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). *Computerized Adaptive Testing Theory and Practice.* (W. J. van der Linden and C. A. W. Glas, Eds.) Dordrecht, The Netherlands: Kluwer Academic Press.

Wainer, H., Brown, L.M., Bradlow, E.T., Wang, W., Skorupski, W.P., & Mislevy, R.J. (2006). *Automated Scoring of Complex Tasks in Computer-Based Testing.* (D.M. Williamson, I.I. Bejar, and R.J. Mislevy, Eds.) Mahwah, New Jersey: Lawrenece Erlbaum Associates.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185-201.

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

Wang, S. & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized testing. *Applied Psychological Measurement, 25*(4), 317-331.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(4), 427-450.

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

Weiss, D.J. (1974). *Strategies of Adaptive Ability Measurement.* Minneapolis: University of Minnesota.

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 methods. *Applied Psychological Measurement, 27(4),* 299-301.

Wu, I-L. (2001). A new computer algorithm for simultaneous test construction of two-stage and multistage testing. *Journal of Educational and Behavioral Statistics, 26*(2), 180-198.

VITA

Candace L. Macken-Ruiz was born in Pittsburgh, Pennsylvania. She attended the Pennsylvania State University in University Park, Pennsylvania and obtained a Bachelor of Arts degree in Economics with a second major in French. She then obtained a Master of Science degree in Epidemiology and Preventive Medicine from the University of Maryland Medical School in Baltimore, Maryland. She was employed by the federal government in Maryland and worked for a number of federal agencies including the Bureau of the Census, the Center for Medicare and Medicaid Services, the Indian Health Service, and the Substance Abuse and Mental Health Services Administration as a statistician. In September 2002 she entered the Graduate School at the University of Texas at Austin. En route to the accomplishment of her terminal degree, she obtained a Master of Arts degree in Program Evaluation.

Permanent Address: 10211 Brantley Bend, Austin, Texas, 78748

This dissertation was typed by the author.