This Dissertation Committee for Marc Anthony Johnson certifies that this is the approved version of the following dissertation:

An Investigation of Stratification Exposure Control Procedures in CATs Using the Generalized Partial Credit Model

Committee:

____
Barbara Dodd, Supervisor

____
Gary Borich

____
Tasha Beretvas

____
Keenan Pituch

____
Daniel Powers

An Investigation of Stratification Exposure Control Procedures in CATs Using the Generalized Partial Credit Model

by

Marc Anthony Johnson, B.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

December 2006

An Investigation of Stratification Exposure Control Procedures in CATs Using the Generalized Partial Credit Model

Publication No. ___________

Marc Anthony Johnson, Ph.D.

The University of Texas at Austin, 2006

Supervisor: Barbara G. Dodd

The *a*-stratification procedure of item exposure control was designed to stratify items by item discrimination to ensure that an adaptive test would administer items from the entire range of items, not just the most-informative ones. An improvement to the *a*-stratification method, the *a*-stratification with *b*-blocking procedure added stratification according to item difficulty in order to take into account any correlation that might exist within the item pool between item discrimination and item difficulty. These procedures have been shown to work well using dichotomous items. This dissertation explored both stratification procedures using *polytomous* item pools to investigate whether or not an optimum number of strata could be implemented when administering polytomous computerized adaptive tests.

In addition to the stratification procedures, two other exposure control conditions were studied. The randomesque procedure was used in one condition while a no exposure control condition served as a baseline condition. Items calibrated according to the generalized partial credit model were used to construct two item pools. Since the items covered three areas of science, content balancing procedures were incorporated to ensure that each adaptive test provided the appropriate balance of content. Maximum likelihood estimation was used to estimate ability levels from simulated CATs. The number of strata used with both stratification procedures ranged from two to five, to ensure enough items per stratum.

Along with descriptive statistics and correlations, bias and root mean squared error helped portray the accuracy of the simulated tests. Item exposure and item pool usage rates were used to show how much of the item pools were being used across administrations of the tests. Finally, item overlap rates were calculated to show how many of the same items were being used among simulated examinees of similar and different abilities.

The results of this study did not reveal an optimum number of strata for the stratification procedures with either item pool. Furthermore, the randomesque procedure outperformed the stratification procedures in terms of item exposure and item overlap rates for both item pools. This surprising result was not affected by the number of strata used within the stratification procedures.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

Computerized adaptive testing has become widely acknowledged as a means of tailoring tests to individual examinees and providing efficient estimates of examinees' abilities from these tests. The theoretical and psychometric basis for tailored (adaptive) testing was described by Frederic M. Lord (1980), followed by his proposal of using computers to achieve adaptive testing. Introducing computers into adaptive testing meant that computers could essentially present a different test to each examinee, using items from a common pool, and compute an ability estimate for each examinee based on the responses to the items. In other words, the examinees would take a computerized adaptive test (CAT) that was designed for their own individual level of ability – not having to wade through items that are not necessarily appropriate for them. Along with this, scores (ability estimates) could be obtained with the same or better accuracy as traditional paper-and-pencil tests, but with fewer items.

Although computerized adaptive testing has become widely used, there are several practical issues that limit its expansion into larger arenas, namely high-stakes testing. Even with the Graduate Record Examination (GRE) Board and National Council of State Boards of Nursing introducing computer adaptive versions of their tests in the mid '90s, other testing agencies are slow to move into computerized adaptive testing due to its limitations. Unfortunately, this reluctance is not without reason.

In 1994, the Kaplan Educational Center demonstrated just how simple it was to "steal" some of the item pool of the GRE. Kaplan instructed some of its employees to take the GRE over several weeks and remember all of the items that they encountered. These items were written down and kept in a journal by Kaplan. As time progressed, several items appeared repeatedly in the journal. It became apparent to Kaplan that a high proportion of the items on the GRE had been seen by its employees. Kaplan shared this finding with Educational Testing (ETS), the company responsible for the GRE.

Despite the repercussions against Kaplan for suggesting and overseeing such a seemingly unethical task, the issue of test/item security became an issue of concern for computer adaptive testing. Although it was proposed that the GRE item pools were too small for proper operation, simply enlarging them would not have solved the problem. Therefore, the item selection procedure used in the CAT became the next suspect.

Wainer and Eignor (2000) discussed analyses of real data on GRE item usage – data obtained during the first few years of the GRE-CAT operation – that focused on the number of times an item was used and its resulting rank, for the 2,000 most commonly used items from the GRE-Verbal and GRE-Quantitative tests. In this case, the item used most frequently had the highest ranking, 1, while the least used item had lowest, 2000. It was found that the GRE-Quantitative used its items more evenly than the GRE-Verbal and that, for both item pools, there was an exponential decline in item usage as the items' ranks decreased. What this tells us is that the item selection algorithm did a poor job of selecting items evenly, enhancing the opportunity of a test/item security breach. This was a problem of item exposure control.

The main concern with item exposure control in computerized adaptive testing is that if examinees can answer CAT questions correctly from any previous knowledge of particular items then the ability estimate generated through the CAT is not accurate. The ability estimates are intended to reflect how much of a domain, or single dimension, an examinee knows, not how much of the item pool an examinee knows ahead of time. A second issue of item exposure control concerns item pool utilization. Item pools are difficult, not to mention expensive, to develop for large-scale testing. Knowing this, it is more cost effective to use the entire item pool, which is not always the case with CAT.

Also, the effects of item overexposure are aggregated when CATs are administered over long periods of time. With this, items can be overexposed with the frequent administrations of the tests and examinees could help other examinees by alerting them to particular items on the test. However, when testing schedules are restricted to “windows” of times, then the effects may be subdued, if not eliminated. Restricting the test scheduling, however, is largely dictated by practical and policy issues and is often not a viable option. Constraints such as these have led to the use of statistical algorithms for controlling item exposure.

Since the Kaplan-GRE scandal, new security measures have been implemented within CAT to control the exposure of items. For example, a new conditional-exposure control method was incorporated into an adaptive version of the Scholastic Aptitude Test (SAT). However, as shown by Wainer and Eignor (2000), the item pool was still not used evenly. A knowledge-performance curve analysis was used to portray the relationship between the number of items answered correctly and the “percent of the item pool known by the examinee”. In this particular analysis, the item pool represents the “domain of knowledge” that the test is to tap. From this analysis, it was found that knowing 17% of the SAT-Verbal domain would result in a 50% correct score while knowing 33% of the domain would achieve a 75% correct sore.

From these numbers, it is apparent that the ratio of knowledge-to-performance for the SAT-Verbal is remarkably different from an ideal situation. In an ideal situation, if

50% of a domain is known, then one could expect that the performance score would be close to 50%, causing a 1:1 knowledge-to-performance ratio. With the SAT-Verbal, though, the ratio is about 1:3 at one point along the knowledge continuum and about 1:2 at another, showing an inefficient use of the item pool.

Eignor, Stocking, Way and Steffen (1993) presented simulation studies done with item exposure control methods for the computer adaptive versions of the GRE and SAT. The GRE CAT used the Sympson-Hetter exposure control methodology, which controls item exposure in a probabilistic fashion. This approach distinguishes between the probability that an item is selected for an examinee and the probability that the item is administered. The goal of this method is to control the probability of the item being administered, since overexposure can result if the item is administered each time it is selected. As will be discussed in greater detail later, the Sympson-Hetter procedure is a multistage process that involves, first, generating exposure control parameters for items, then a random number generation to determine whether or not an item is actually administered. It should be noted that since the GRE item pools contain sets of items that are based on common stimulus material, exposure control parameters are generated for the stimulus material as well as for the items.

Simulation studies on the GRE CAT revealed that the highest exposure rates for items and item sets were in the .2 to .3 range for all three GRE item pools – GRE Verbal (GRE-V), GRE Quantitative (GRE-Q), and GRE Analytical (GRE-A). This means that an item or passage could appear on 20 to 30% of the CATs administered to the typical population. The desired maximum rate of exposure was set to .2. For the GRE-V and

GRE-Q item pools, the average exposure rate for all used items and sets was just over 10%, while it was just under 9% for the GRE-A item pool (See Eignor et al., 1993).

In contrast to the GRE, the SAT CAT – SAT Verbal and SAT Math – involved a randomization procedure for controlling item exposure. As outlined by McBride and Martin (1983), the first item was randomly chosen from a set of the eight best items, the second from the seven best items, the third from the best six items, and so forth. The idea was that the eighth and subsequent items were optimal for the examinee. In other words, after the initial items – those before the first optimal item – examinees would be presented items optimal for them and the items, in theory, would vary from examinee to examinee.

Results from SAT CAT simulation studies showed that the highest exposure rates for both the SAT-V and SAT-M were in the .5 to .6 range. However, the average exposure rate for all used items and passages in the SAT-V was just over 11%, while it was just over 10% for the SAT-M CAT (see Eignor et al., 1993).

Since these early simulation studies, and perhaps before, security in computer adaptive testing has generated great concern leading to several new developments in controlling item exposure (Chang & Ying, 1999; and Chang, Qian, & Ying, 2001). Researchers have looked at different ways of controlling item exposure by developing new techniques for handling the phenomenon and performing various studies comparing the techniques with one another. Comparing exposure control techniques with one another, researchers have taken aim at proposing which technique appears to work the best given a particular computer adaptive testing situation. In doing so, other issues have been taken into account, including item pool size and content balancing restraints, when prescribing the best procedure.

It is worth mentioning that only until recently most of the previous research on item exposure control methods has been with items that follow the multiple-choice format. This could be due to the fact that most tests that have been developed into CATs have also followed this item format. However, with the advent of using more performance-based items in CATs, more research is needed in controlling item exposure and item pool use with polytomously scored items, items that produce multiple-category scoring options. An example of this type of item is a Likert-type item in which the response could be one of five categories (i.e., strongly disagree, disagree, neutral, agree, and strongly agree). Other types of polytomous items include essay questions and constructed-response math items, which are scored on the basis of what a student *can do*, rather than the typical “right/wrong” criterion. As a result, these types of items use statistical models that facilitate complex scoring schemes, such as partial credit scoring. Statistical models that use such scoring schemes are the graded response model (Samejima, 1969), partial credit model (Masters, 1982), the generalized partial credit model (Muraki, 1992).

Performance assessment items, such as the types just mentioned, have achieved an increasing amount of attention, especially within the computerized adaptive testing framework, for several reasons. Proponents of these items argue that they represent a better means of finding what the students can and cannot do. For example, an item that involves five steps for arriving at the solution could score the examinees according to the number of steps that were performed correctly. In this case, each step might be worth one point leading to a possible range of scores of 1 to 5 for that particular item. The scores achieved by the examinees will reflect how far they were able to correctly work through the problem. Knowing which steps were incorrectly performed gives the examinees an indication of what particular knowledge they lack instead of the typical global notion that they do not understand a particular problem. In other words, this type of task makes it easier to pinpoint where an error in understanding occurs rather than in a typical multiple-choice item.

From this, it appears very crucial that efforts of expanding computerized adaptive testing to include more performance-based assessments continue. Specifically, it is important to continue examining the many innovations that have been designed for traditional multiple-choice CATs and see how well they apply to CATs for polytomously scored items.

This dissertation will focus on the integration of performance-based items and item exposure control methods – methods that have been designed more for traditional multiple-choice assessments – to further analyze the potential of these exposure control procedures in performance-based assessments. The exposure control procedures that will be investigated in this simulation study are the *a*-stratification procedure (Chang & Ying, 1999) and *a*-stratification with *b*-blocking procedure (Chang, Qian, & Ying, 2001). The goal of analyzing these methods is to determine if there is an *optimum number of strata* needed to stratify a polytomous item pool in order to achieve the perceived advantages of these item exposure control procedures. These procedures will also be compared to a frequently used randomization procedure and a no item exposure control condition. The latter condition serves a s a baseline condition.

CHAPTER TWO: LITERATURE REVIEW

The purpose of this literature review is to provide the theoretical framework for investigating item exposure control methods in computerized adaptive testing. The first section will discuss item response theory, its assumptions and models. The assumptions are crucial to the appropriate uses of item response theory and its models and, therefore, are discussed first. Following the assumptions, the models of item response theory are presented. The dichotomous item response theory models will be discussed followed by models that are used with items having ordered-response categories (i.e., Likert-scale items, or items scored based on partial credit).

Following the discussion of item response theory, the subject of adaptive testing will be discussed. A brief introduction to computerized adaptive testing will be presented followed by the well-known advantages it has over traditional paper-and-pencil testing. In addition to this, several components of computerized adaptive testing will be discussed beginning with the topic of item pools in adaptive testing. Next, the most common item selection techniques for progressing through an adaptive test will be reviewed as well as criticisms of using such procedures. Along with this, typical procedures for terminating a computerized adaptive test (CAT) will also be mentioned.

Ability estimation procedures in adaptive testing will also be discussed including their advantages and disadvantages in the scope of computerized adaptive testing. Content balancing and item exposure control will conclude the section on computerized adaptive testing. Within the item exposure control section, several procedures that have been developed for enhancing test security will be addressed along with empirical research outlining their advantages and disadvantages.

Lastly, previous research in dichotomous and polytomous item response theory on procedures of exposure control will be reviewed. This section will focus on the previous research investigating stratification methods of exposure control in an attempt to provide the foundation for the investigation conducted in this dissertation.

Item Response Theory

Computing an examinee's ability from responses obtained through a CAT is based on item response theory (IRT), which is essentially a collection of mathematical models that characterize items and examinees on a common scale. Within this framework, the scale that indicates the difficulty of an item is the same scale that is used to assign scores (ability estimates) to all examinees. With this, one of the benefits of IRT is that examinees can be compared; using the ability estimates, regardless of the items each examinee is administered.

IRT is a measurement theory that was developed to address areas that have been problematic for the *classical test theory* (CTT) that had dominated the measurement models and procedures for constructing tests and interpreting scores. It was revealed that CTT had some shortcomings that affected its utility in certain applications. First, item statistics determined through CTT depended on the particular group of examinees with which they were obtained. This prevented the generalization of the item statistics to other groups of examinees without careful sampling procedures. Also, the ability estimates depended upon the particular choice of items used on a test, preventing the generalization of the ability estimates to other items that could have been used without extensive equating procedures.

The property of invariance distinguishes IRT from CTT in that the item statistics or estimates do not depend on a particular group of examinees and the ability estimates do not depend on a particular set of items. In other words, assuming decent model fit, using an IRT model to fit response data should result in the same item characteristic curves regardless of the distribution of ability used to estimate the item parameters and the ability estimates should be the same regardless of the items administered from a given calibrated item pool. This invariance has been shown to hold within a linear transformation (Lord, 1980).

The statistical models that IRT plays host to can be classified into two categories: dichotomous and polytomous models. Dichotomous IRT models are used when the items are scored according to a "right/wrong" criterion. In this case, the examinee either gets the item right or wrong. Polytomous IRT models, however, are used when more complex scoring schemes are necessary for the items. Typical item types for polytomous IRT models include essay questions, constructed-response math items, and Likert-type items. These items are not scored on a strict "right/wrong" criterion, but on a partial-credit or graded-response criterion.

In general, IRT models give the probability of an examinee answering an item in a given way, incorporating both the estimation of ability and the item parameters of the models. An item characteristic curve (ICC) represents the mathematical expression that relates the probable success on an item to an examinee's ability. ICCs differ across IRT models due to the parameters used to estimate them.

IRT does have some assumptions, which are addressed by Hambleton and Cook (1977), that the data (item responses) must meet in order for its models to be properly used. One assumption of many IRT models is that the items used must measure the same thing – a single dimension of knowledge or underlying trait, such as verbal proficiency, statistical reasoning, and spatial memory, to name a few. This assumption is referred to as unidimensionality. It should be understood that this assumption cannot be strictly met as factors such as motivation and anxiety most often affect test performance. However, as Lord (1968) pointed out, showing a unidimensional structure to a set of items "may provide a tolerably good approximation".

Local independence, an assumption related to unidimensionality, is classified into two forms: strong and weak. The strong form of local independence refers to the item responses of an individual examinee being statistically independent. In other words, a response to one item is not influenced by the response on the other items within the test. This is related to unidimensionality, in a sense, because only the underlying ability measured through the items influences item response, not some other ability. The weak form of local independence is the idea that pairs of test items are uncorrelated for examinees of the same ability level. The difference between the two forms of local independence is that the strong form defines the condition of being "statistically uncorrelated" while the weak form defines the condition of just being "uncorrelated".

The third assumption of IRT is that the probability of responding in a given category is a mathematical function of the item parameter(s) and the person's trait level. A graphical representation of the mathematical function is called an item characteristic curve (ICC). The different IRT models specify this relationship between probable success and ability in different forms, leading to differing displays of the ICCs.

Dichotomous IRT Models

Rasch (One Parameter Logistic) Model

The Rasch model, proposed by Georg Rasch (1960), is the simplest IRT model since it characterizes items using just one parameter, an item's difficulty (*b*). Commonly known as the one-parameter logistic model (1PL), the Rasch model can be mathematically represented as

$$P(\theta) = \frac{1}{1 + e^{-(\theta - b)}} \qquad (1)$$

where θ denotes an examinee's ability and *b* represents an item's difficulty. The 1PL gives the probability of an examinee with ability θ responding correctly to an item of difficulty *b*.

Since the application of the 1PL model assumes that only item difficulty influences examinee performance, two other assumptions are implied. First, since there are no indices of item discrimination allowed in this model, it is assumed that all items are equally discriminating. The second assumption is that correct responses are unlikely due to guessing. Within the 1PL, there are no parameters indicating the probability of correct responses through guessing, which can happen using multiple-choice items.
As Figure 1 illustrates, ICCs resulting from the 1PL will only differ in location; slopes and lower asymptotes of the curves for all items will be the same. Also, since guessing is not modeled under this model, the lower asymptotes of the ICCs will be zero. From this graph, the items' difficulty, *b*, is represented by the trait level at the "point of inflection."

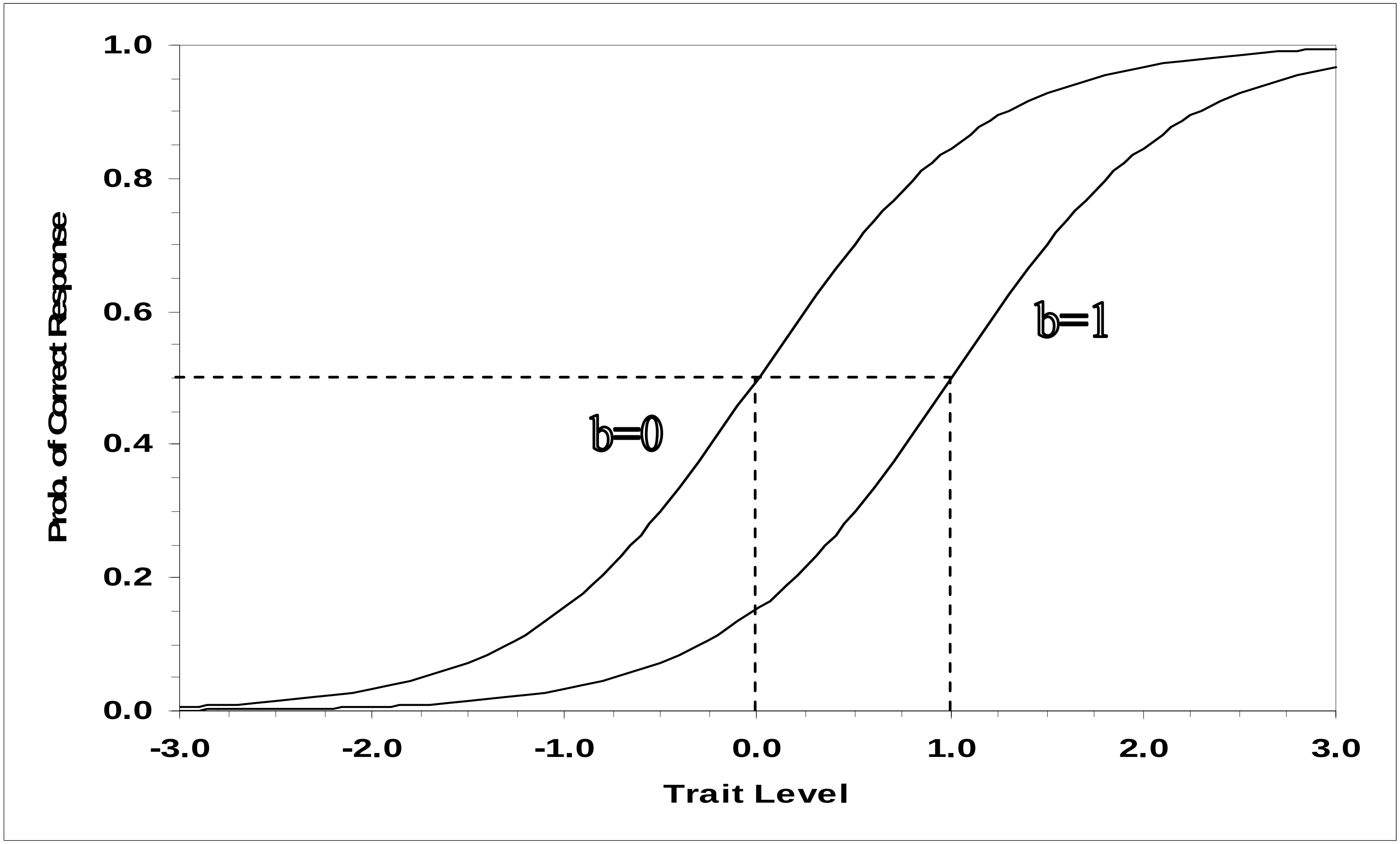


Figure 1. Item characteristic curves for the 1PL.

This point on the ICC – where the probability of a correct response is .50 - signifies where the rate of change shifts from increasingly accelerating to increasingly decelerating.

Two Parameter Logistic Model

Allowing the items to have different discrimination capabilities adds a parameter to the 1PL. The resulting model, the two-parameter logistic model (2PL) was proposed by Birnbaum (1968) and is given by

$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}} \qquad (2)$$

where the addition of the *a* parameter signifies that each item might discriminate among the examinees differently. From this, it can be said that some items will discriminate among the examinees better than others. ICCs from the 2PL will differ in location and slope, but not lower asymptotes (see Figure 2). The discrimination parameter is proportional to the slope of the ICC, so that higher discrimination parameters will yield steeper slopes than items with lower discrimination parameters.

The differences in item discrimination parameters reveal the differences in utility of separating examinees into different ability levels. For example, with a high discrimination parameter an item is more useful for separating examinees into different ability levels than an item with a low discrimination parameter. The usual range for item discrimination parameters is between 0 and 2.

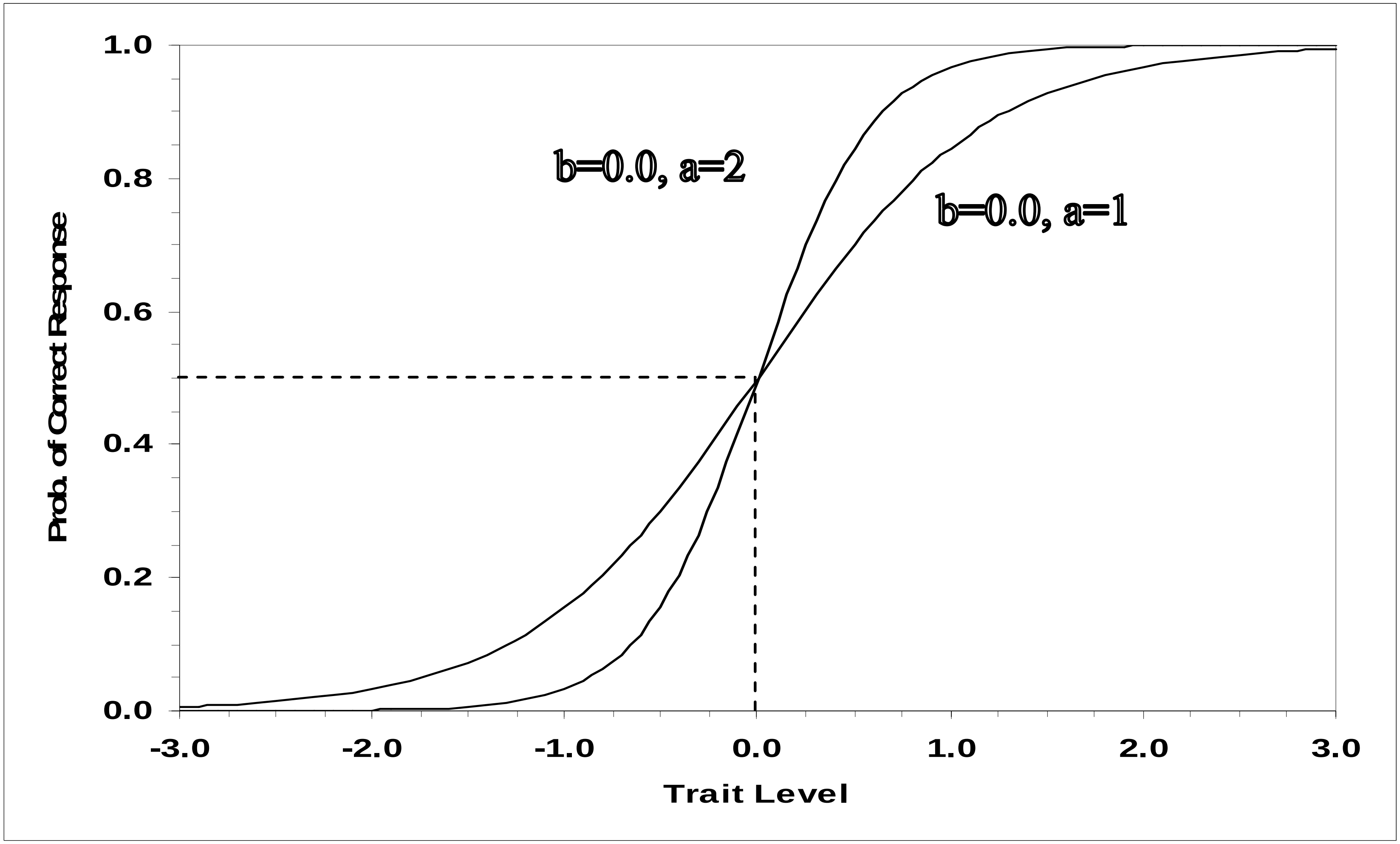


Figure 2. Item characteristic curves for the 2PL.

Three Parameter Logistic Model

Although the Rasch and 2PL models do not account for examinees guessing on test items, it cannot be ruled out in multiple-choice items. It can occur that an examinee gets an item correct that is outside the appropriate range of difficulty. Given this situation, the most logical solution is to account for the capacity for guessing that can take place. The three-parameter logistic model (3PL), also proposed by Birnbaum (1968) and defined as

$$P(\theta) = c + \frac{1-c}{1+e^{-a(\theta-b)}} \tag{3}$$

does just that. The *c* parameter represents the lowest probability of getting an item correct through guessing. For this model, *c* now represents the lower asymptote of the ICCs. Allowing for guessing raises the lower asymptote of the ICCs so that they do not have to be zero, as with the 1- and 2PL (see Figure 3). Of course, an examinee will likely only guess on an item that is seemingly difficult, in which case that examinee has low ability for answering that item.

Figure 3 shows two items, one with a guessing parameter and the other without a guessing parameter, therefore being modeled under the 2PL for comparison. Although both items have the same difficulty, the item modeled under the 3PL *appears* to have a different point of inflection. Even though the item difficulties are the same, the guessing parameter causes the shift of the inflection point with the 3PL item. The item difficulty of

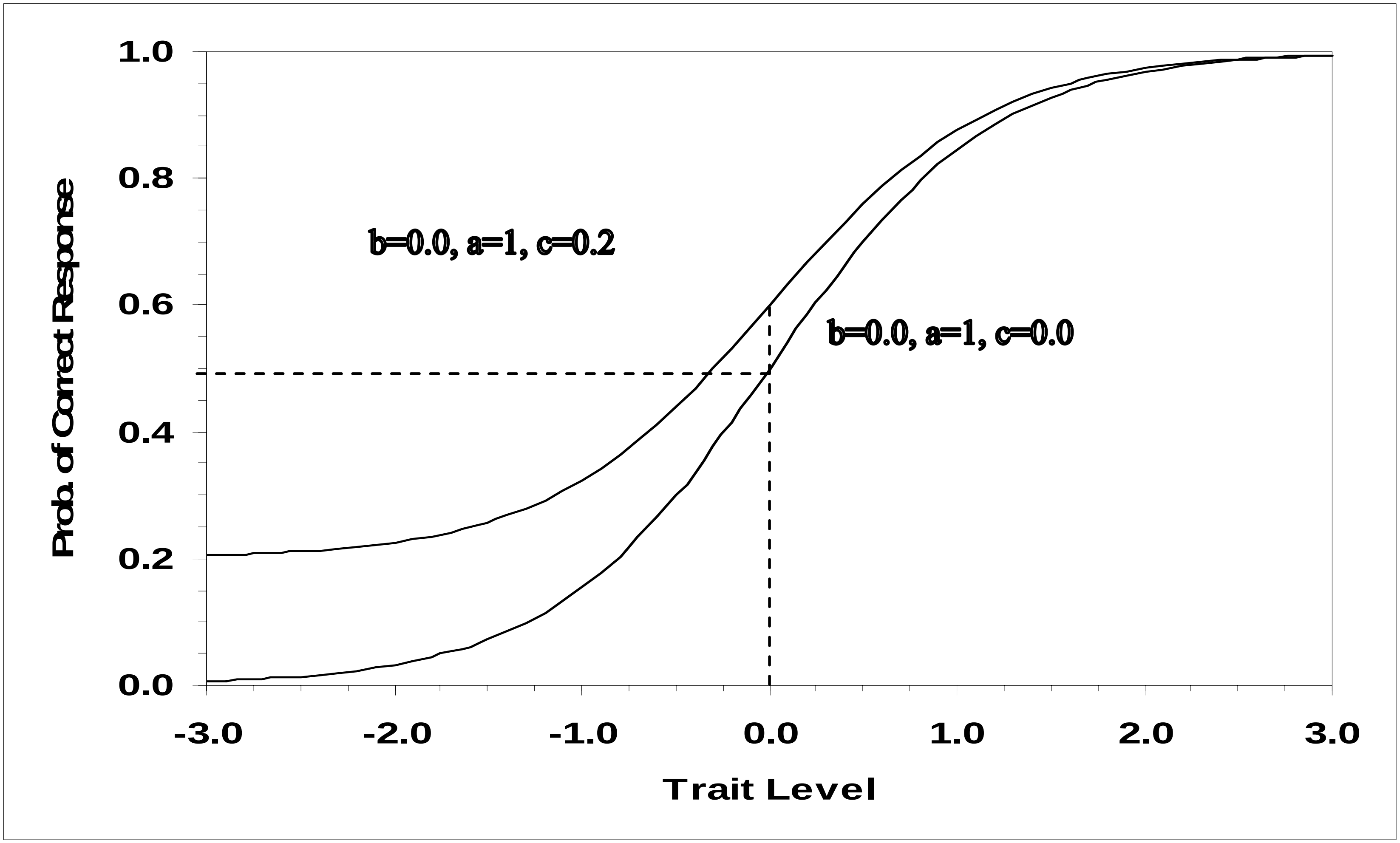


Figure 3. Item characteristic curves for the 3PL.

the 3PL item is associated with a higher probability of a correct response than the item with a guessing parameter of zero.

Urry (1977) stated that the 3PL is the most appropriate IRT model for multiple-choice items since those items usually vary in discriminatory power and can be answered correctly through guessing. Also, Urry outlined four conditions for choosing items for the item pools used for computerized adaptive testing, specifically for the dichotomous case. These conditions are: 1) the item's discrimination value must exceed 0.8; 2) the item difficulties must be evenly and widely distributed, for example from -2.0 to +2.0; 3) the "guessing" parameter must be less than 0.3 and 4) there must be at least 100 items in the item pool.

Item and Test Information for Dichotomous IRT Models

Related to item discrimination is the notion of item information – the amount of measurement precision an item provides at each point along the ability continuum. More specifically, an item with high discrimination at a particular ability level provides more information at that ability level than an item with low discrimination. In this sense, item information is related to the slope of the ICC. As will be discussed later, one goal of computerized adaptive testing is to select items that provide the most information at the examinee's current ability level. Birnbaum (1968) demonstrated the notion of item information, for dichotomous IRT models, using

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

where $I_i(\theta)$ represents the amount of information an item i provides at ability θ. $P_i^{'}(\theta)$ is the first derivative of $P_i(\theta)$ with respect to θ, $P_i(\theta)$ is the probability of a correct response given θ, and $Q_i(\theta) = 1 - P_i(\theta)$, the probability of an incorrect response.

Item information can be summed across a test to provide test information.

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \qquad (5)$$

The same principles hold here as well. The higher the test information the greater the measurement precision achieved. This should not be surprising since high test-information is the direct result of the high item information provided by the items that comprise test. Given this, test information can be made analogous to test reliability since it indicates how well the test measured the examinee's ability (Parshall, Spray, Kalohn, and Davey, 2002). This comparison is often made by assuming that the error of measurement has the same value at all ability levels. However, scores based on IRT may have different errors of measurement at different levels of ability preventing the precision of a test from being easily described by a single estimate of reliability. From this, Green, Bock, Humphreys, Linn, and Reckase (1984) showed that an average of the measurement error – across all ability levels – could be used to generate a "marginal reliability" estimate

$$\bar{\rho} = \frac{\sigma_\theta^2 - \bar{\sigma}_{e^*}^2}{\sigma_\theta^2} \qquad (6)$$

where σ_θ^2 is the variance of proficiency and $\bar{\sigma}_{e^*}^2$ is the average of the values of error variance.

Furthermore, as defined by Birnbaum (1968), the amount of information a test provides is inversely related to the precision with which ability is estimated:

$$SE\left(\hat{\theta}\right) = \frac{1}{\sqrt{I(\theta)}}, \qquad (7)$$

where $SE\left(\hat{\theta}\right)$ is the standard error of estimate of ability.

Polytomous IRT Models

This section provides an overview of some of the polytomous IRT models. Although several polytomous IRT models exist, this section will only provide an in-depth explanation of three of the models: the graded response model, the partial credit model, and the generalized partial credit model. Other polytomous models will be introduced, but any readers interested in those models should refer to the original sources.

Graded Response Model

Samejima (1969) proposed a graded response model that deals with ordered polytomous categories, such as letter grading, A, B, C, D, and F, used in the evaluation of students' performance; strongly disagree, disagree, agree, and strongly agree used in attitude surveys; or partial credit given in accordance with an examinee's degree of attainment in solving a problem. The graded response model requires two steps for determining response probabilities. The first step involves determining the characteristic curves for each score category of each item,

$$P^{*}{}_{ix}(\theta) = \frac{\exp[a_i(\theta - b_{ix})]}{1 + \exp[a_i(\theta - b_{ix})]}, \qquad (8)$$

representing the probability of an examinee responding to an item in a particular category or higher, using b_{ix} as the between category "threshold" parameter for each of the *x* categories of item *i* and a_i as the slope parameter for item i. The category threshold parameters represent the ability level needed to respond above the threshold *x* with .50 probability $P_{ix}^{*}(\theta)$ represents an "operating characteristic curve" for the threshold between adjacent categories of an item. Under this model, an item may have $(m_i + 1)$ scoring categories, with $x_i = 0,1,\ldots,m_i$ representing the possible scores for the item. Figure 4 shows the operating characteristic curves for a five-category item. The *actual* category response probabilities, called category response curves, are computed using the operating characteristic curves

$$P_{ix}(\theta) = P_{ix}^{*}(\theta) - P_{i(x+1)}^{*}(\theta) \qquad (9)$$

This equation is simply a subtraction of the probabilities of an examinee responding in adjacent categories of an item. In order to use this formula, the probability of responding in or above the lowest category is defined as $P_{i0}^{*}(\theta) = 1.0$ and the probability of responding above the highest category is $P_{x_i+1}^{*}(\theta) = 0.0$. Figure 5 shows the category response curves for the same five-category item represented in figure 4. The graded response model simplifies to the 2PL when there are only two categories.

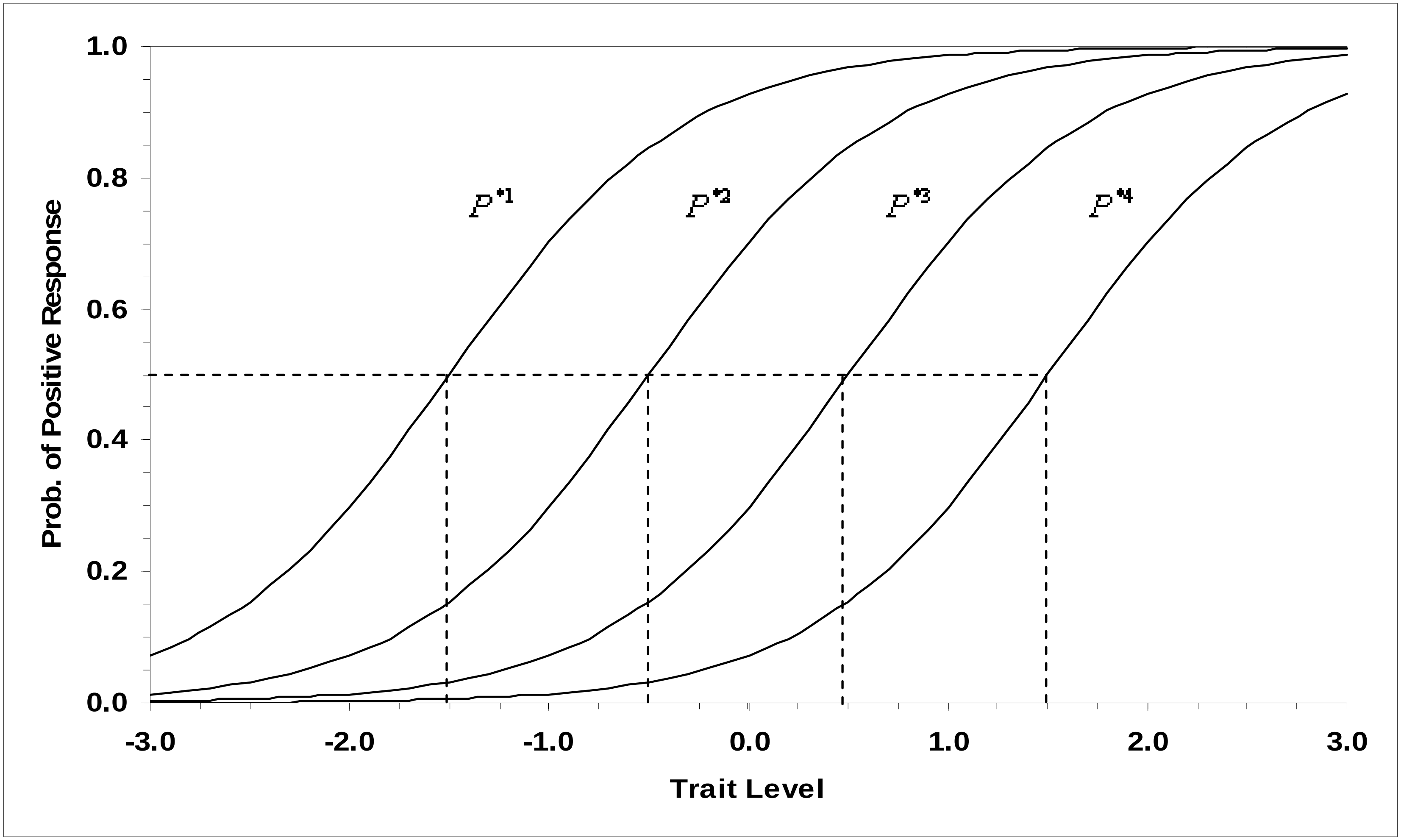


Figure 4. Operating characteristic curves for a five-category item under the graded response model.

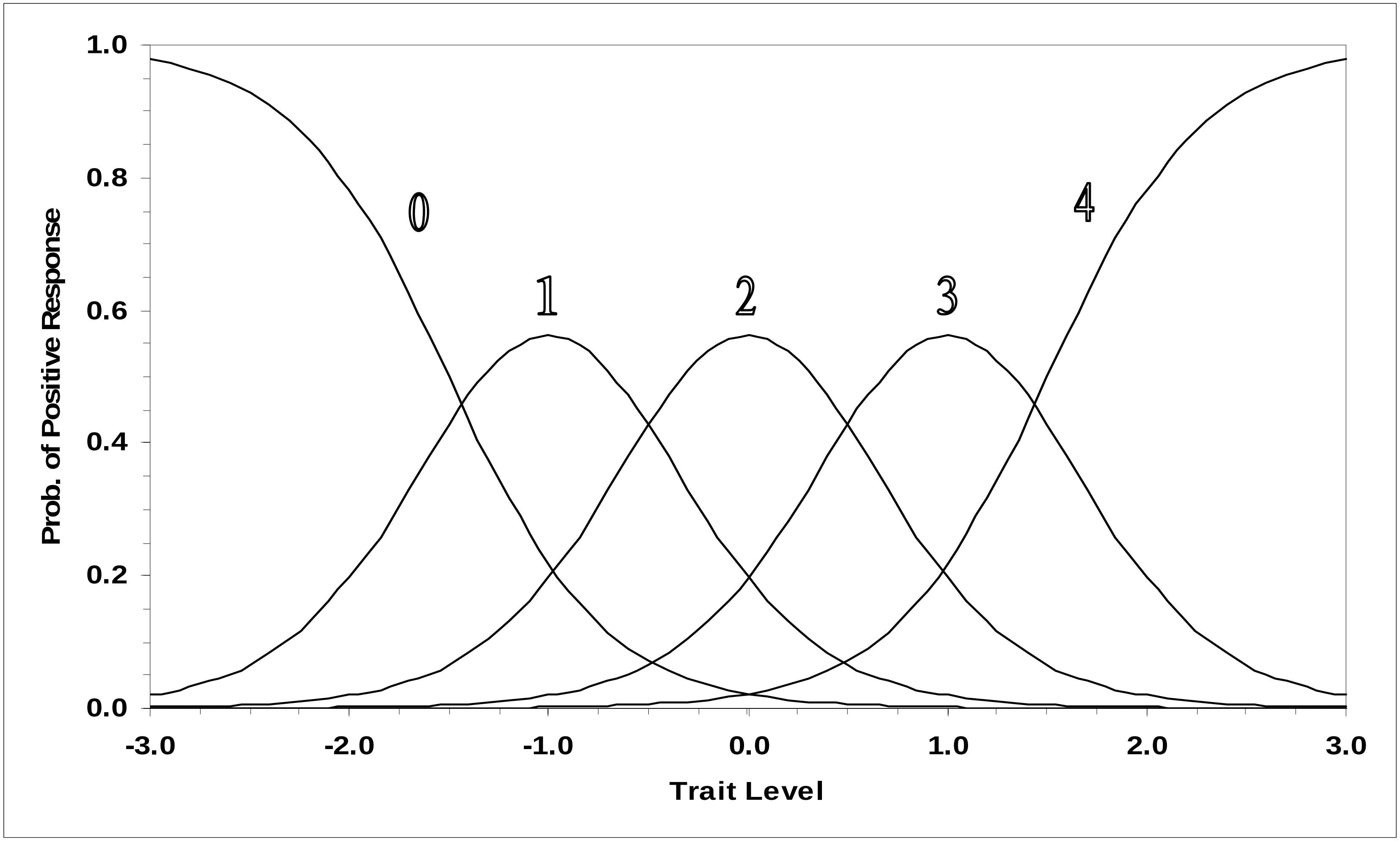


Figure 5. Category response curves for the five-category item under the graded response model.

Partial Credit Model

The partial credit model (Masters, 1982), like the Rasch model for dichotomies, only uses difficulty parameters to characterize the items, or more specifically the response categories of the items. The probability of receiving a score x on item i can be obtained through

$$P_{ix}(\theta) = \frac{\exp \sum_{k=0}^{x} (\theta - b_{ik})}{\sum_{h=0}^{m} \exp \sum_{k=0}^{h} (\theta - b_{ik})}, x = 0,1,\ldots,m_i, \tag{10}$$

where m_i is the number of categories, minus one, for item i and b_{ik} is the step difficulty for the x category. The step difficulty is the difficulty associated with the transition from one category to the next. Also, the step difficulty is the point on the ability scale where two consecutive category response curves intersect, arriving at its name of category intersection parameter.

Explicitly, Equation 10 states that the probability of an examinee responding in category *x* on an m_i step item is a function of the difference between an examinee’s trait level and a category intersection parameter (Embretson & Reise, 2000). This equation simplifies to the Rasch model, for the dichotomous case, when there are only two categories. Figure 6 is an illustration of a three-category item under the partial credit model. The intersection points of adjacent category curves are represented by the step difficulties – the “difficulty” associated with the transition from one category to the next.

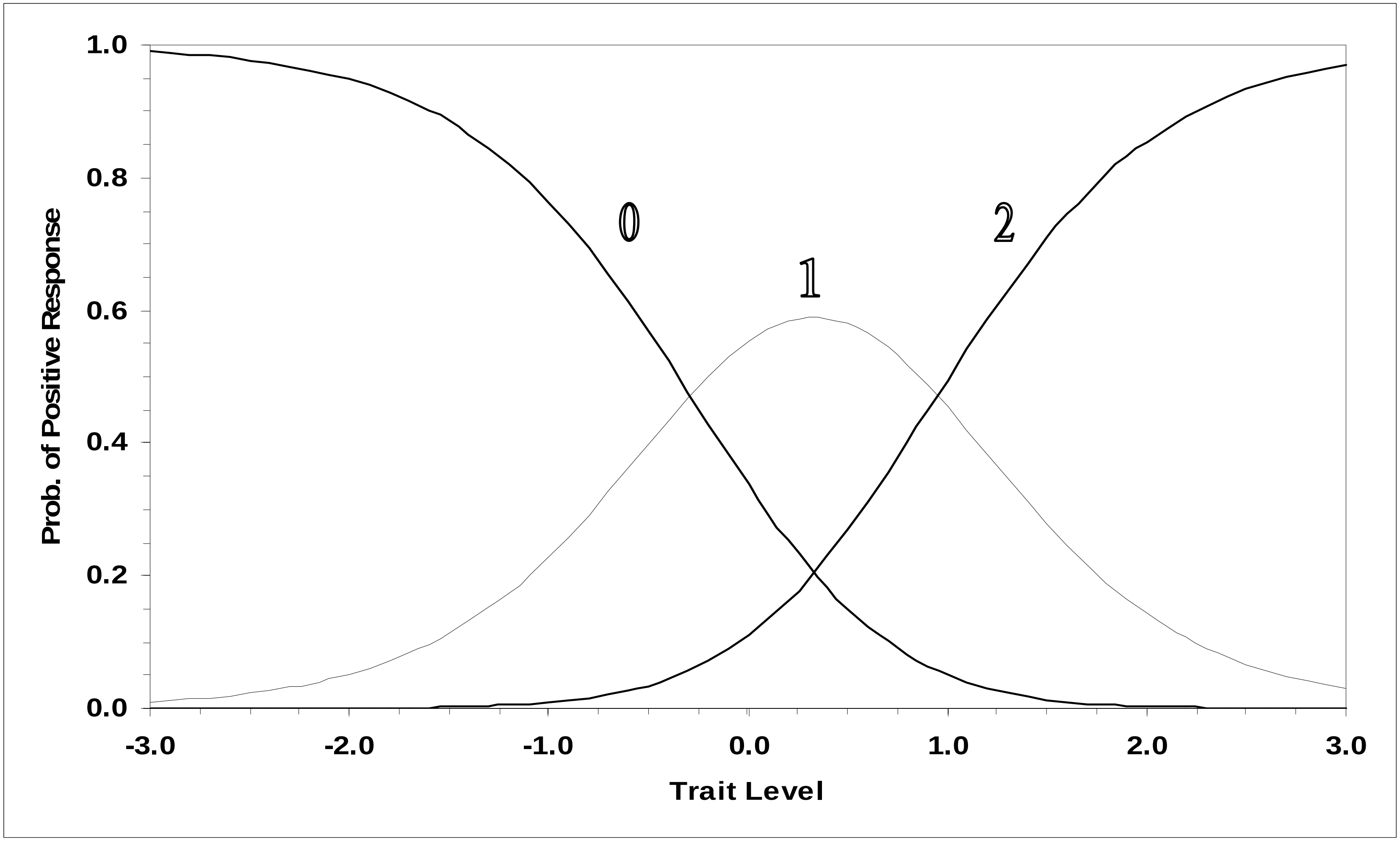


Figure 6. Category response curves for a three-category item under the partial credit model.

Generalized Partial Credit Model

Just as the 2PL could be obtained by relaxing the assumption of uniform item discrimination from the 1PL, the same maneuver can obtain the generalized partial credit model from the partial credit model, as proposed by Muraki (1992). However, it *is* assumed that the categories within an item are uniformly discriminating thus requiring a single discrimination parameter for each item. The generalized partial credit model *(see Figure 7)* is given by

$$P_{ix}(\theta) = \frac{\exp\sum_{k=0}^{x} a_i(\theta - b_{ik})}{\sum_{h=0}^{m_i}\exp\sum_{k=0}^{h} a_i(\theta - b_{ik})}, x = 0,1,\ldots,m_i, \qquad (11)$$

where m_i and b_{ik} are defined as they were in the partial credit model and a_i indicates "the degree to which the categorical responses vary among items as θ level changes" (see Muraki, 1992) and is analogous to item discrimination in dichotomous IRT.

Models Not Described

There are other polytomous IRT models that will not be described in detail here. However, these models will be introduced so that interested readers may consult with the original citations. Bock (1972) introduced the nominal response model to characterize item responses when the responses are not ordered along a trait continuum. Andrich (1978) developed a rating scale model which could be derived from the partial credit model. This model provides a scale location parameter for each item as well as category intersection thresholds for the entire set of items. Muraki (1990) proposed a rating scale model, a "modified" graded response model, to analyze questionnaires with

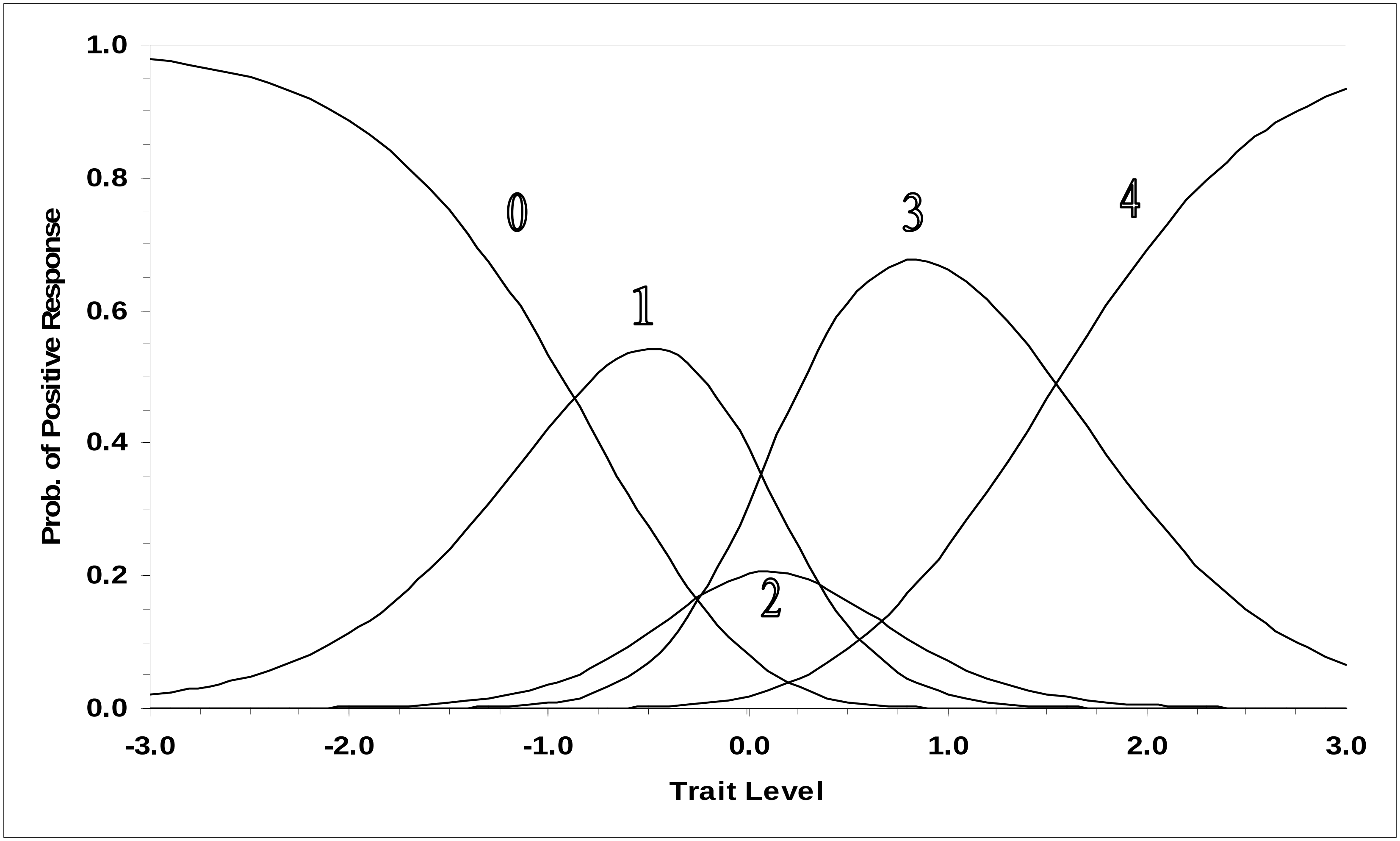


Figure 7. Category response curves for a five-category item under the generalized partial credit model.

rating-scale type response formats. This model breaks the category threshold parameter of the original graded response model into a location parameter for each item and a category threshold parameter for the entire scale. Finally, Rost (1988) designed the successive intervals model that combines features of the partial credit and rating scale models.

Item and Test Information for Polytomous IRT Models

Although similar in concept, item information in polytomous IRT is quantified differently than in dichotomous cases. Here, item information is composed of the individual contributions that each of the score categories of the polytomous item provides. These individual contributions were first defined by Samejima (1969) as

$$I_{ix}(\theta) = \frac{\left[P'_{ix}(\theta)\right]^2}{P_{ix}(\theta)} - \frac{P''_{ix}(\theta)}{P_{ix}(\theta)} \qquad (12)$$

otherwise known as the item-category information function (Muraki, 1993). In equation 12, $P_{ix}(\theta)$ is the probability of responding in category *x* for item *i* given θ, $P'_{ix}(\theta)$ is the first derivative of $P_{ix}(\theta)$ with respect to θ, and $P''_{ix}(\theta)$ is the second derivative of $P_{ix}(\theta)$ with respect to θ. From this, item information, $I_i(\theta)$, is defined as

$$I_i(\theta) = \sum_{x_i=0}^{m_i} I_{ix}(\theta) P_{ix}(\theta) \ . \qquad (13)$$

Substituting Equation 12 into Equation 13 and simplifying the resulting equation shows item information defined as

$$I_i(\theta) = \sum_{x=0}^{m} \frac{\left[P'_{ix}(\theta)\right]^2}{P_{ix}(\theta)} - \sum_{x=0}^{m} P''_{ix}(\theta). \qquad (14)$$

It was shown by Samejima that the second term in equation 14 equals zero and therefore can be removed from the equation.

As with information functions in dichotomous IRT, test information in the polytomous context can be obtained through the summation of the item information functions, and that test information is inversely related to the precision with which ability is estimated. (*Refer to Equations 5 and 6.*)

Adaptive Testing

Adaptive testing, as we know it today, is based on the principles of intelligence testing by Alfred Binet (1857-1911). This form of testing follows the simple logic that if an examinee is asked a question and answers it correctly, then the next question asked should be more difficult. Conversely, if an examinee is asked a question and answers it incorrectly, then the next question should be easier. This approach appears logical since a correct response to an item should not be rewarded with an item in which the chance of another correct response is unjustifiably greater. Similarly, an item with an increased probability of a wrong response should not follow a wrong response to an item. Adaptive tests that do not utilize this mechanism of administering items run the risk of obtaining inadequate information about an examinee’s ability. For a sufficient adaptive test, this mechanism of presenting items to the examinee should continue until the test administrator feels an accurate estimate of the examinee’s ability has been obtained.

The main advantage that adaptive testing has over conventional (non-adaptive) tests is that scores from adaptive tests are highly accurate for individuals of all ability levels, rather than just for the “average” ability individuals as in the conventional tests. This idea is based on the fact that most conventional tests are designed for the “average” individual. In other words, these tests provide very accurate scores for “average” abilities, but not for extreme high or extreme low abilities.

While adaptive testing was first introduced in the form of paper-and-pencil tests, the technological advancements of computers have made *computerized* adaptive testing a desired reality and will be discussed as such for the remainder of this paper.

Computerized Adaptive Testing

Studies have shown the psychometric efficiency of computerized adaptive testing. Urry (1977) used 57 Civil Service Job applicants on an adaptive verbal ability test to demonstrate the improved measurement of computerized adaptive testing over conventional testing. The CAT achieved an 80% reduction (compared to a conventional test) in the test length required to attain any of several specified levels of reliability. However, McBride and Martin (1983) claim that these results were based on indirect evidence since “the conventional test reliabilities were based on Spearman-Brown equation adjustments to the reliability obtained in an independent sample, and the adaptive test reliability was merely assumed, not empirically verified” (p. 225).

From this, McBride and Martin (1983) conducted two studies investigating the psychometric properties of CAT in comparison with a conventional test. The goal of the studies was to determine if CAT is more reliable *and* more valid than a comparable conventional test, holding test length constant between the two modes of testing. Both studies used 150 items that fit the guidelines set forth by Urry (1977), mentioned earlier, and a sample of male Marine recruits at the Marine Corps Recruit Depot, San Diego as the examinees. The difference in the studies is two-fold: 1) study II used a larger sample than study I, and 2) study II utilized a different computer system for administering the tests.

In terms of reliability, both studies demonstrated that the adaptive tests achieved higher reliabilities than the conventional tests, but only up to certain test lengths: 10 items for study 1 and 13 for study 2. However, the superiority of the adaptive tests was not clear in terms of validity. For these studies, validity was defined as the correlation between scores on adaptive and conventional tests created ("experimental" tests) and the scores from a concurrently administered 50-item criterion test, created from two obsolete operational test forms measuring word knowledge. The first study, based on a "pilot sample" did not show any significant differences in validity between the adaptive and conventional tests. However, the second study, based on a larger sample, showed significant differences in validity between the tests, again, up to a particular test length.

Green (1983) noted other advantages of computerized adaptive testing, apart from the psychometric efficiency already discussed, including:

1) Improved test security.
2) Elimination of answer sheets.
3) Immediate scoring and score reporting.
4) Each examinee can work at his/her own pace - staying busy, productively.
5) New items can be pre-tested without disrupting the flow of the testing program.

One other advantage listed as a *potential* advantage, given the year it was mentioned, that has certainly become a reality with the advancement in computing technology is the use of new types of items. For example, constructed-response items or items involving pictures or video clips can now be used in CATs.

Components of Computerized Adaptive Testing

It is through all of these advantages that computerized adaptive testing has continued to flourish in measurement testing. There are, however, important issues to analyze when considering the implementation of a CAT. These issues include item pool size and characteristics, item selection algorithms and stopping rules, ability estimation procedures, content balancing, and item/test security. Throughout this section, each of these issues is examined in greater detail.

Item Pool

Compared to traditional paper-and-pencil tests, in which the items used give the best measurement for the average examinee, CAT item pools provide the best measurement precision at *all* levels of proficiency. Along with this, CAT item pools need to be large enough to offset 1) the uneven item exposure of CAT item selection algorithms and 2) the number of occasions - within a short period of time - which some CATs may be administered. Stocking (1994) conducted a study investigating optimal item pool sizes for five adaptive tests that mirrored high-stakes tests. The CATs included two measures of verbal reasoning, two measures of mathematical reasoning, and one measure of analytical reasoning. CAT simulations found that across the five adaptive tests, an item pool 12 times the size of an average CAT was sufficient for fixed length adaptive testing. This finding was consistent across content and statistical considerations of the CATs. Stocking went on to conclude that an item pool with six to eight linear test forms would support fixed-length adaptive testing, when the fixed length is approximately one-half that of the linear paper-and-pencil test.

Way (1998) concluded that the results obtained in Stocking's study generalized best to admissions testing programs (i.e., GRE, SAT, ACT, etc.). These testing programs are considered high-stakes, but there is no control over the testing populations. For licensure or certification tests in which the testing population is controlled, the required item pool size is not as straightforward. Stahl and Lunz (1993) recommended item pools between 600 and 800 items, with 500 items being a minimum, for certification exams that ranged from 50 to 100 items.

Way (1994) and Way, Zara, and Leahy (1996) conducted simulations on two medical licensure exams in which one ranged from 60 to 250 items and the other 60 to 180 items. The operational item pool for the exam with a maximum of 250 items was found to have 1,300 to 1,800 items while the other exam had an item pool of approximately 1,100 to 1,500 items. The size of these item pools is necessary for two reasons. First, examinees are usually allowed to take a licensure exam once after a failed attempt, thereby exposing another set of items to the examinee. Also, it is suggested that an item pool be able to accommodate the content coverage of four maximum length exams (Way, 1994).

Item Selection

Item selection techniques govern how a CAT starts, continues, and stops for an examinee. Several approaches have been proposed for each of these functions. A general progression of a computer adaptive testing system is illustrated below:

1) How to start: Specify an initial estimate of proficiency; this specifies an initial item.
2) How to continue: Estimate proficiency $(\hat{\Theta})$ after each item response. Choose the next item that is most-informative near Θ to be administered.
3) How to stop: Stop when the precision of Θ is adequate, or when some number of items has been administered.

Starting a CAT. A widely used method of selecting the initial item of a CAT is to assume that each examinee is of “average” ability, therefore prompting the item selection algorithm to choose an item of medium (average difficulty) as the initial CAT item. Also, starting a CAT with an easy item can serve as a “warm-up” and confidence enhancer for the examinees. These methods are typically used when no prior information is obtained on the tested population. However, when prior knowledge *is* known about the testing population – for example, ability estimates obtained through previous testing – it can be used to specify an actual ability estimate to start a CAT.

The issue of test security is very important to discuss here and will be discussed throughout this paper. Choosing the most informative item to begin a test could result in over exposure of some items. For example, a single group of examinees, defined by previous background information, could receive the same first item – it being the most informative item for an age group, for example. This would lead to a large number of examinees being exposed to same item, leading to high exposure of that item. High exposure lends itself to the possibility that examinees could get an item right based on their previous knowledge of the item, not on the skill that was supposed to be measured.

In turn, ability estimates obtained through the examinees' previous knowledge of an item is highly inaccurate and provides false information about their performance and ability.

Continuing a CAT. Three goals are inherent in item selection for continuing a CAT: 1) maximize test efficiency, retain appropriate content balancing, and maintain test security. These goals are often in conflict resulting in a compromise among them, in practice. Maximum information and Bayesian item selection methods for maximizing test efficiency will be discussed next while content balancing and test security will be discussed later in this paper.

Maximum information item selection (Lord, 1977) selects the item that has the largest information value at the examinee's current ability for administration. In this procedure, item information may be determined using the IRT model that is assumed to underlie the examinee's responses to the test items. When the next item is to be selected, the item information for all items not yet administered is determined and the item with the highest information value at the individual's current level of ability is chosen for administration. This procedure may lend itself to estimation error in that items that appear to measure well at the *estimated* ability level may not do so at the *true* ability level.

Under this method, the item information at the true ability level represents the efficiency of the item for estimating θ. However, it can be the case that the estimated ability level $\hat{\theta}$ is substantially different from θ, disallowing the item information value from being a good indicator of efficiency in estimating θ. The discrepancy between $\hat{\theta}$ and θ is prevalent in the early stages of a CAT when a small number of items are used to estimate ability. This problem has led researchers to investigate other mechanisms for selecting the next item in a CAT, specifically for the early stages of the test.

Chang and Ying (1996) made a distinction between the local and global information around θ - local information being the information around a small region of θ and global information being the information outside the region. Global information was defined as the expected value of the log-likelihood ratio between $\hat{\theta}$ and θ. The researchers argued that the global information is valuable for use when the location of θ is not sufficiently known. Furthermore, simulation studies on using global information showed improvements on minimizing the differences between estimated and true abilities at the early stage of simulated tests (see Chang & Ying, 1996 for complete study).

Much more computationally intensive than the maximum information method, the Bayesian item selection method (Owen, 1975) selects items based on how much the variance of the posterior ability distribution will decrease. Items that have the potential of decreasing the variance of resulting ability distribution will have greater chances of being chosen as the next item. Although this procedure can often choose an item that may not be the most informative at a particular ability level, this procedure does yield superior results to maximum information method (Parshall et al., 2002).

Stopping a CAT. CATs can be group into two categories according to their lengths: fixed length or variable length. Fixed length CATs are specified by a fixed number of items to be administered to each examinee. In this case, each examinee receives the same number of items regardless of ability level. The problem with this is that the ability estimates of some examinees will not be as precise as others, depended on item response patterns. The more items an examinee can respond to, the more accurate the resulting ability estimate will. Occasionally, fixed length CATs are not long enough for all examinees to obtain precise ability estimates (Thissen & Mislevy, 2000). Also, simply increasing the number of items to be administered would lessen the advantage of gathering information about ability using the fewest items possible.

Variable length CATs prescribe the amount of precision in ability estimation that is needed before terminating the test. This allows the testing program to continue administering items until such a precision is met. In doing so, examinees will complete the test with different numbers of items each responded to, preventing imprecise estimates of ability that could occur if a fixed stopping rule was used. Because a CAT could run out of items in the pool before the specified precision of ability estimation is met, a combination of "target precision" and "maximum number of items" should be utilized as a stopping rule (Thissen & Mislevy, 2000).

Ability Estimation

Throughout a CAT, an examinee's ability estimate is updated after each response to give the best guess of the ability at that point. This *provisional* ability estimate represents the hypothesized ability, not the true (actual) ability. Two types of ability estimation procedures are commonly used in adaptive testing: those based on maximum likelihood estimates and those based on Bayesian strategies.

Maximum likelihood estimation procedures (Andersen, 1972; Bock, 1972; Lord, 1968; Samejima, 1969) give the maximum value of the likelihood function – the maximum probability that an examinee produces a given pattern of correct and incorrect answers to a set of items, given that the ability is at a fixed value. Hambleton, Swaminathan, and Rogers (1991) denotes the likelihood function is given as

$$L(u_1, u_2, ..., u_n \mid \Theta) = \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j} \qquad (15)$$

where u_j represents the observed response to item *j*, 0 for correct and 1 for incorrect;

$Q(\Theta) = 1 - P(\Theta)$.

The likelihood function is also a function of the item parameters, which could be explicitly specified in the likelihood function as

$$L(u_1, u_2, ..., u_n \mid \Theta, \beta_j) = \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j} \qquad (16)$$

where β_j represents the item parameter vector (a_j, b_j, c_j) for item *j*.

The primary disadvantage of maximum likelihood estimation procedures is that the estimation process cannot begin until there is at least one correct *and* one incorrect response obtained from the examinee. This means that a maximum likelihood estimate of ability cannot be determined for the examinee that gets every item correct or for the examinee that gets every item incorrect, the dichotomous case. Therefore, for the high ability examinee that continues to get items correct, difficult items continue to be administered until an incorrect response is obtained. Similarly, for low ability examinee that continues getting item incorrect, easier items continue to be administered until a correct response is obtained. Until this mixed set of responses has been obtained, a variable step-size estimation approach is used to update the ability estimate after each item. In most cases, this approach involves placing the trait estimate at half the distance between the current ability estimate and the most extreme *b*-value within the appropriate content area.

However, in the polytomous case, maximum likelihood estimation can occur after an examinee’s first response provided that the response is not in the extreme response categories. Variable step-size estimation is also used for ability estimation when maximum likelihood estimation cannot occur. As with dichotomous items, the variable step-size estimation procedure in the polytomous case assigns a new ability estimate that is half the distance between the current ability estimate and the most extreme item difficulty parameter, within the appropriate content area.

The maximum likelihood estimation procedures assume that no prior information on the examinees is being used to help estimate abilities. Researchers argue that using such information for ability estimation is a more proper approach in adaptive testing. Prior information may be obtained through previous experiences with the testing population and can provide information leading to an ability distribution before testing. This prior ability distribution can be integrated into the ability estimation procedure through the application of Bayes’ theorem, which gives the probability of an event occurring given that another event has occurred (Weiss, 1974).

Bock and Aitkin’s (1981) maximum a posteriori (MAP) and expected a posteriori (EAP) were introduced to distinguish between the Bayes modal and the Bayes estimators, respectively, that had been previously proposed as variants on Bayesian ability estimation methods. Bock and Mislevy (1982) adapted the EAP to computerized adaptive testing.

In adaptive testing, beginning after the first response has been obtained, a posterior distribution is determined that specifies a range in where an examinee’s ability may lie. However, the EAP and MAP procedures are designed to summarize this range using only one value. Therefore, the EAP uses the mean (expected value) of the posterior distribution as the point estimate of ability while the MAP uses the mode (maximum), given the prior information on the examinee.

An advantage of the EAP method over the maximum likelihood estimation methods is that EAP estimates *always* exist, even for an all-correct or all-incorrect response vector. With this, the EAP estimator is stable at all adaptive test lengths.

Weiss and McBride (1984) highlight one major disadvantage of the Bayesian strategies of ability estimation. In a simulation study, it was found that only unbiased measurement and measurements of equal precision resulted when an accurate prior of ability estimate was used. In ability testing, the actual ability of an examinee is never known; otherwise the testing would be essentially pointless. Therefore, accurate priors are never accomplished in Bayesian strategies. Given this, it was found that bias in ability measurement and measurement precision occurs when inappropriate priors of ability estimate is used. Specifically, when the prior is above the actual ability, then positive bias will occur; on the other hand, negative bias will occur when the prior is below the actual ability. Weiss and McBride recommend using maximum likelihood estimation (over the Bayesian method) when there is no differential prior information available for examinees.

Content Balancing

Up to this point, it has been stressed that item selection mostly relies on the amount of information that an item may provide toward estimating ability. However, in practice, the item selection techniques that were previously discussed often are forced into a compromise with the actual item content that is required by the test. Conventional and adaptive tests are often designed from a “table of specifications” or blueprint of some sort, outlining the breakdown of specific item types and content needed for the test. The compromise in adaptive testing is that when a CAT is selecting the next item for administration it may have to forego the item with the most information in favor of the item that represents what is actually needed to adhere to the test blueprint. This is commonly known as content balancing.

Positive effects of content balancing in measurement testing can be seen in several ways. First, it can serve as a validity check in that the items on a test represent what the test blueprint says they will test. For example, suppose a math test has three sections: algebra, geometry, and calculus. The algebra section is to cover 40% of the test with geometry and calculus each representing 30% of the test. Each CAT constructed from the math item pool should represent this breakdown of item types in order for it to achieve content validity.

Secondly, content balancing provides a means of developing alternate forms of the test that are parallel to each other. Again, this guarantees that each CAT constructed from the item pool, although possessing different questions, will have approximately the same percentage of item types, as specified by the blueprint. Thirdly, depending on the difficulty of specific contents, any lack of content balancing would prevent the more difficult contents from being administered to everyone even though the *content* needs to be tested on everyone. For example, if the calculus items were difficult, then without content balancing less proficient examinees would never see these items, thus removing calculus as a part of overall math ability for those examinees. Similarly, highly proficient examinees may only take these difficult items, removing the other (easier) skills from overall math ability for these examinees.

Three methods of content balancing an computerized adaptive testing have been proposed: the constrained CAT (CCAT) method (Kingsbury & Zara, 1989), modified multinomial model (MMM) approach (Chen & Ankenmann, 2004), and a modified CCAT approach (Leung, Chang, and Hau, 2001). The CCAT method selects items from content area that is the farthest below its targeted ideal administration percentage. The MMM method was designed to prevent the order effects that could occur using the CCAT procedure. This method uses a random number from a uniform distribution to determine the content area where the next optimal item would be selected. However, sampling errors could prevent target percentages from being met during testing. Therefore, once a content area has fulfilled its targeted percentage, a new multinomial distribution is generated using the unfulfilled percentages of the remaining content areas. The modified CCAT approach selects the optimal item from all of the unfulfilled content areas, avoiding the order effect of the CCAT method.

Exposure Control (Test Security)

A major goal of the exposure control techniques used in CATs is maintaining test security throughout the test's progression. As mentioned earlier, proficiency estimates are highly inaccurate when examinees have prior knowledge of test items. The effect of prior knowledge, as shown by Kaplan, typically results when several examinees respond to some of the same items, although they test on different occasions, and pass on their "knowledge" to individuals who will test at a later time. When this occurs, the only logical solution would be to discard the compromised items from the item pool. The problem with this is that item pools are expensive and time-consuming to develop and discarding items may render the item pool temporarily inoperable. A disabled item pool, in this case, may result from a sharp decrease in number of available items and/or a significant change in the content coverage – and psychometric properties – of the item pool, due to the content of items that are discarded.

The maximum information and maximum posterior precision procedures have received attention in CAT research when dealing with test security. Recall that the maximum information procedure selects the item that has the largest information value at the examinee's current ability estimate for administration while the maximum posterior precision procedure selects items based on how much the variance of the posterior ability distribution will decrease. These procedures can, and most often will, select the same initial item for every examinee, when no auxiliary information is used to start the CAT. Furthermore, the second item chosen for administration would be based upon whether the examinee got the first item right or wrong; the second item would be chosen out of two possibilities. Therefore, over time, the three most informative items could be over-used and examinees could eventually share these items with future examinees, thus arriving at the possibility of inaccurate ability estimates.

Item exposure control methods influence exposure rates through statistical algorithms incorporated into the item selection procedures. Parshall et al. (2002) mentioned item exposure control as a more direct alternative to 1) using a "big pool" of items (more than 5,000 items) and 2) restricting the testing schedule, for maintaining test security. However, modifying item selection techniques with item exposure methods has shown a decrease in test precision.

The methods proposed for item exposure control can fall into three categories, as first proposed by Way (1998): a) randomization procedures, b) conditional procedures, and c) stratification procedures. It should be noted that Way discussed the first two categories of exposure control methods and not the stratification procedures since they were developed later. The following section will introduce some of the methods for controlling item exposure. Other methods that are not described here will be listed at the end of the section for interested readers.

Randomization procedures. Perhaps the earliest, and most simple, statistical method for controlling item exposure was proposed by McBride and Martin (1983). Aimed as reducing sequence predictability and the exposure of initial items, the method first called for the five most appropriate (informative) items from which one would be randomly chosen as the first item administered. From there, the second item to be administered would be randomly chosen from the remaining four most-appropriate items, third from the three best items, and so on until the fifth item is administered, representing the best item. Although this method is easy to implement, it should be obvious that it does not prevent the overexposure of the most "popular" items that could start a CAT. However, this method does prevent sequence predictability.

A slight, and effective, variation to this method is the randomesque item selection procedure introduced by Kingsbury and Zara (1989). This method chooses the next item of administration from a group of items, with different groups used for choosing each subsequent item. The random component of this method is that the algorithm could randomly choose the next item from a group of items. However, this technique also allows for the next item chosen to be, in fact, the best item of the group – non-random selection.

This method can be used with either maximum-information or Bayesian selection procedures and it accomplishes the two goals of reducing item sequence predictability and reducing the exposure of initial items. The random component of this method also prevents examinees with the same ability level from seeing the same items, which the previous method may allow. It was suggested that an item is chosen from among the 2,3,…, up to 10 best items. One may consider item groups larger than 5 items to become laborious to determine statistically, thus increasing processing time and slowing the overall pace of the CAT. However, the necessary computations for determining the next group of items to choose from are done while the examinee is answering the current question, preventing a lag in computer processing time.

A third randomization procedure, Revuelta and Ponsoda's progressive-restricted method (1998), was developed to control item exposure without a decrease in test precision. This method is a hybrid of two separate exposure control methods that both enhanced test security individually. The restricted maximum information procedure selects items using the maximum information procedure, but does not allow items to be administered in more than 100*k*% of the tests. Here, *k* represents an item's maximum exposure rate. The progressive method, proposed by Revuelta (1995) and Revuelta and Ponsoda (1996), also adds a random component to the maximum information procedure, but this random component's influence in item selection decreases as the CAT progresses.

In a simulation study (Revuelta & Ponsoda, 1998), it was found that for variable-length tests these methods yielded longer tests – 3 more items – needed to reach the same degree of precision as the maximum information, McBride and Martin's randomization procedure, and Kingsbury and Zara's randomesque procedures. As a result, no clear differences in precision rates emerged. For fixed-length tests, the progressive and maximum restricted procedures had fewer items with high exposure rates, compared to the other methods. Since item exposure depends on test length, results of exposure control for the variable-length tests were not reported.

From these individual results, Revuelta and Ponsoda found it beneficial to combine the methods into one procedure in order to achieve maximum precision and exposure control. Simulation studies revealed that the progressive-restricted method had control over maximum exposure rates and the number of unused items. Furthermore, the precision rates achieved by this combined procedure were similar to rates obtained using the restricted maximum information procedure (see Revuelta and Ponsoda, 1998).

Conditional procedures. In order to control item exposure more directly, Sympson and Hetter (1985) suggests a probabilistic algorithm that controls item exposure using a pre-specified rate of exposure. This procedure is computationally intensive requiring simulations for generating exposure parameters, *K*, before actual live testing. The exposure parameters are used in CATs to govern which items are administered to the examinees, after been selected. What follows are the steps that outlines this procedure.

1) Specify the maximum expected item exposure rate *r* for the test.

2) Construct an information table that lists all available items by ability level. Within each ability level, their information level ranks the items, with the most informative item on top.

3) Generate the first set of K_i values, specifying a vector of 1s to represent first set of exposure parameters.

4) Conduct simulated adaptive tests to a random sample of simulees. Use maximum information to select an item i then generate a random number *x* from the distribution (0,1). If *x* is less than or equal to K_i, administer item i. Exclude item i from being selected again, regardless of whether or not it was administered.

5) When the simulee sample has been tested, use the number of times item i was selected (NS) and the number of times it was administered (NA) to compute its probability of selection, P(S) and its probability of being administered, P(A), for all items.

$$\begin{aligned} P(S) &= NS/NE \\ P(A) &= NA/NE \end{aligned} \qquad (17)$$

(NE = total number of examinees)

6) Using r from Step 1, compute new exposure parameters as follows:

If P(S) > *r*, then new $K_i = r/P(S)$

If P(S) $\leq$ *r*, then new $K_i = 1.0$

7) If n represents test length, then there should be at least *n* items in the pool where $K_i = 1.0$. These items are always administered when selected. If there are not enough of these items, then set the *n* largest K_i to 1.0.

8) With the new K_i, repeat steps 4-7 until the maximum value of P(A) obtained in step 5 is slightly above *r*.

The final K_i estimates are used in the live CAT testing, shown in the following steps:

1) Select the most informative item for the current ability estimate.
2) Generate a random number *x* from the distribution (0,1).
3) If *x* is less than or equal to K_i, administer the item; if *x* is greater than K_i, do not administer the item. Repeat steps 1-3 with the next most-informative item. Items selected but not administered are withheld from future selection.

Disregarding the disadvantage that the Sympson-Hetter methodology (SH) requires time-consuming simulations to generate the exposure parameters need for operational testing, the exposure parameters can only be used for the expected ability distribution from which they were based on. In other words, the ability distribution that was specified as the starting point for generating the exposure parameters may, in fact, not be the same as the actual ability distribution of the actual examinees that will be tested. Therefore, the exposure parameters obtained through one ability distribution cannot be used for one of different characteristics.

A solution to this problem is to generate exposure parameters for different ability levels. The conditional Sympson-Hetter (Parshall, Davey, & Nering, 1998) controls item exposure for examinees with similar abilities, making them independent of a specific ability distribution. The steps of the original SH approach are used for the conditional Sympson-Hetter (CSH) with one modification. The ability distribution is divided into distinct groupings allowing the frequencies of item selection to be recorded for each ability level.

Stratification procedures. Based on the issues pointed out by Chang and Ying (1996) and Chang and Stout (1993), a new procedure of item exposure control was developed that considers an item’s discrimination power as the criterion for selecting items. The *a*-stratification procedure (Chang & Ying, 1999) was designed to use items with high *a*s more efficiently than in the case of information-based selection procedures. Items with high discrimination provide more information than those with low discrimination, thus leading to the greater use of high discriminating items. However, Chang and Ying showed mathematically that high discriminating items might not be useful if the difficulty levels of these items are not near the estimated ability θ. The conclusion reached was that high discriminating items should not be used at the beginning of a CAT, when little information is known about θ, but as the test progresses when more information is obtained (see Chang & Ying, 1996).

The a-stratified selection (AS) method is described in the following steps:

1) Partition the item pool into K levels (strata) according to *a* values;
2) Partition the test into K stages;
3) For each stage, select n_k items from the kth level based on the similarity between b and $\hat{\theta}$, then administer the items;
4) Repeat step 3 from k = 1, 2, …, K.

The number of levels needed to partition an item pool depends on the spread of item discrimination values within the item pool, how well the range of item difficulty matches that of the expected ability, test length, and item pool size. It should be apparent that when the item pool consists of little variability among the values, then few levels are necessary. However, greater variability among the values requires a larger number of levels. Also, item pools in which the difficulty range corresponds to the expected ability distribution can be divided into more levels. Furthermore, larger item pools can be partitioned into more levels than smaller ones.

Chang and Ying argued that the *a*-stratification was most appropriate when there was not a correlation between item discrimination and item difficulty. However, previous research indicates that is seldom the case. For example, Lord (1975) illustrated how the use of the ability, θ, scale almost invariably leads to an undesirable situation in which item discrimination and item difficulty become positively correlated with each other. Although, a transformation on the ability scale was proposed by Lord to overcome this condition, the relevant issue to the study proposed in *this* paper is that of the correlation between the item parameters. Lord's study analyzed six separate sets of test data and found a positive correlation between item discrimination and item difficulty, leading him to believe that this phenomenon occurs more often than it should. Other authors have also experienced this phenomenon in their studies (Lord & Wingersky, 1984; Parshall, Hogarty, & Kromrey, 1999). Furthermore, Stocking (1998) argued that the positive correlation between item discrimination and item difficulty that is typically found in item pools may interfere with the technique of stratification hampering the mechanisms of the CAT. Given this, a modification to *a*-stratification is necessary to overcome the issue of a correlation between item discrimination and item difficulty.

Weiss (1973) presented a stratification approach that involved portioning an item pool according to item difficulty. This method was based upon the design of the

Stanford-Binet Scales. In short, test items were organized in sets of “mental age” levels in which 50% of the norm group of the corresponding chronological age responded correctly to those items. There was a distinction made between an examinee’s “basal age” and “ceiling age”. The basal age referred to the level of difficulty in which all items, presumably, would be answered correctly by the examinee. The idea is that those items would not provide any information on the examinee’s ability since they were too easy. The ceiling age defined the upper limit of difficulty in which all items are answered incorrectly. Like those of the basal age, these items do not provide any information on ability, but the reason is because they are too difficult. Therefore, this testing structure defines the outer limits of ability, administering items from the appropriate *b* group that closely matches the examinee’s ability.

Given this structure by Weiss, Chang, Qian, and Ying (2001) proposed an improvement to the *a*-stratification procedure, the AS with *b*-blocking (BAS) method. This approach uses the original framework of the AS procedure, but includes a *b* grouping mechanism to prevent possible mismatches between *b* and θ during item selection. The BAS method is outlined below:

1) Divide the item pool into M blocks of *b* values, each block containing the same number of items.
2) Partition each of the M blocks into K strata of *a* values.
3) For k=1, 2, …, K, recombine the *k*th stratum items across M blocks into a single stratum. There are now K strata (analogous to K strata in the AS method).
4) Divide the test into K stages.

5) In the *k*th stage, select items from the *k*th stratum based on the closeness of the *b* values to current θ.
6) Repeat step 5 for k= 1, 2, …, K.

The result is obtaining K strata of item discrimination each consisting of the same range of *b* values.

Following the notion that successful adaptive tests benefit from simultaneously controlling the statistical and content properties of the items, Yi and Chang (2003) suggested adding a content balancing component to the *a*-stratification procedure. This new method takes the content specifications of the test *and* the relationship between the difficulty and discrimination of the items into consideration during item pool stratification. Simply, an item pool is first stratified according to content specifications then assembled into “difficulty” strata according to the BAS approach. Lastly, the items within a stratum are pooled across content groups to obtain the operational strata structure (see Yi & Chang, 2003 for details outlining this procedure).

Previous research investigating the utility of content balancing in stratification designs has supported its perceived usefulness. Leung, Chang, and Hau (2003) performed a simulation study comparing the three methods of stratification: *a*-stratification (AS) *a*-stratification with *b*-blocking (BAS), and BAS with content balancing (CBAS). The study found that the CBAS was best in terms of pool utilization and control of overexposed items. Yi and Chang (2003) found similar results, but also found that measurement precision was enhanced using content balancing within the stratification approach, yielding precision estimates similar to those of the maximum-information with SH exposure control procedure.

Research on Stratification Procedures

Hau and Chang (2001) notes three advantages of the *a*-stratification procedure. First, as pointed out in a previous study, the precision of ability estimation is comparable to the traditional maximum information approach. Secondly, stratification of the item pool results in more even item exposure control – less chance of some items becoming overexposed. Third, it is easier to implement, compared to procedures such as those involving SH methodology in which numerous a priori computations are required.

The rest of this section will be devoted to examining some of the research that has been conducted investigating the mechanisms of the *a*-stratification procedures. A study investigating the optimum number of strata for these procedures is analyzed first, followed by a review of research that compares the *a*-stratification procedures to other common methods of item exposure control.

Optimum Number of Strata (Dichotomous Items)

When using the a-stratification procedures, it should be apparent that the required number of strata necessary for maximum efficiency is of great concern. As Hau, Wen, and Chang (2002) points out, having just one stratum produces lower efficiency in ability estimation than the maximum information approach, since it allows item selection based solely on difficulty. However, having too many strata can lower the chances of selecting items, within stratum, near the current ability estimate *if* the stratum does not have a sufficient range of difficulty.

The aforementioned researchers conducted a simulation study to investigate the optimum number of strata necessary to maximize the efficiency of the *a*-stratification procedures. The goal of the study was to determine the relationship between testing performance, in terms of efficiency and item pool usage, and the stratification process (number of strata used). This study utilized a 3x2x2 research design: three item pool sizes (200, 400, and 800 items); two item characteristics (no correlation between item difficulty and item discrimination and moderate correlation, 0.5, between item difficulty and item discrimination); and two item selection methods (maximum information and matched item difficulty with estimated ability). The items were calibrated according to the 2PL model and maximum likelihood estimation was implemented for estimating ability.

Two (fixed) tests lengths were examined each having a set of the number of strata used for the investigation. The 24-item test used 1, 2, 3, 4, 6, 8, 12, and 24 strata while the 48-item test used 1, 2, 3, 4, 6, 8, 12, 16, 24, and 48 strata. To reflect the conditions and analyses that will be used in the simulation study presented later in this chapter, the analyses from Hau et al. (2002) involving the conditions of maximum information item selection and the correlation between item difficulty and item discrimination being 0.05 will be discussed here. Since only one test length will be investigated later in this paper, the results presented here will reflect both test lengths used by Hau, Wen, and Chang.

For all three item pool sizes, the average bias in ability estimation did not exceed 0.01. Furthermore, the mean squared error (MSE) estimates of CAT accuracy not only decreased as the number of strata increased within each item pool size condition, but also decreased as the item pool size itself increased. Also, the correlations between the true and estimated abilities was above 0.97 across the three item pool size conditions.

Item exposure rates was analyzed in terms of underexposure – items exposed up to 5% of the time – and overexposure – items exposed at least 20% of the time. The number of underexposed items increased dramatically as the item pool size increased, for both test length conditions. For example, for the 24-item simulated CATs, the maximum number of underexposed items for the 200-item pool was 101, 289 for the 400-item pool, and 675 for the 800-item pool. However, the number of overexposed items did not show this dramatic trend. The maximum number of overexposed items for the 200-item pool was 55, 47 for the 400-item pool, and 41 for the 800-item pool, all for the 24-item simulated CAT. Finally, test overlap rates tended to decrease with increasing number of strata within each item pool size condition, as well as decreasing as the item pool size increased.

Presenting these findings illustrates how the stratification procedures work while varying the number of strata for dichotomous item pools. Although an optimum number of strata was not found, per se, patterns of the accuracy of the simulated CATs as well as the efficiency of item pool usage is shown. As will be detailed later in this paper, similar analyses will be compared to a traditional method of item exposure control to see how well the stratification procedures work in a polytomous context.

Comparisons to Other Exposure Control Methods (Polytomous Items)

Simulation studies have been done investigating the *a*-stratification procedures against the more traditional methods of exposure control using a polytomous item pool. These studies have looked at the how precision of ability estimation and item pool usage of the stratification procedures compares to these other methods. The results of the studies have been surprising and have called for further research on the stratification procedures *if* they are to become the appropriately used for exposure control in adaptive testing.

Pastor, Dodd, and Chang (2002) conducted a study comparing six methods of exposure control: no exposure control, the *a*-stratification procedure, the Sympson-Hetter approach, the conditional Sympson-Hetter methodology, the enhanced stratified design, and a conditional enhanced stratified design – exposure conditioned on ability. These methods were assessed in terms of item exposure control, item pool utilization, and measurement precision using two pools of items calibrated according to the generalized partial credit model, with five strata. The results indicated that the *a*-stratification method is promising in terms of providing exposure control, but perhaps in low- to medium-stakes testing, since it did not achieve the same level of control as the other methods.

Another study found some surprising results regarding the use of stratification procedures. Davis (2004) used a pool of 157 polytomous items to compare nine exposure control methods: no exposure control, four randomization procedures, two conditional procedures, and two stratification procedures (a-stratified and enhanced a-stratified). Again, the items used in this study were calibrated according to the generalized partial credit model, and were divided into five strata. However, this study differed from the previous study in that content balancing was used, taking into account the content area *and* number of categories for each item.

It was found that the a-stratification procedure produced high values of unused items, contradicting what was thought to be the strength of the procedure. However, Davis revealed that meeting the content *and* category restraints most likely had an effect on these results. Therefore, the multiple stages of stratification prevented the basic benefits of stratification from being realized, paving the way for more refinements on this procedure.

Research on Randomesque Item Exposure Control

Burt, Kim, Davis, and Dodd (2003) conducted a study comparing six exposure control methods in a polytomous CAT using generalized partial credit items. These methods were no exposure control, randomesque-3, randomesque-6, within .10 logits-3, within .10 logits-6, and Sympson-Hetter. Item groups of three and six were used for the randomesque-3 and randomesque-6 procedures, respectively, as well as for both within .10 logits procedures. An item pool of 210 items from the NAEP 1996 Science test was used which contained three- and four-category items covering three areas of science: physical, earth, and life science. With these content areas, Kingsbury and Zara's content balancing procedure was used for the CAT simulations.

The results of this study showed that the Sympson-Hetter method yielded the smallest maximum exposure rates and the fewest numbers of non-convergent cases compared to the other exposure control procedures. A non-convergent case is defined as having a trait estimate greater than or equal to 4.0 or less than or equal to -4.0 or if the maximum likelihood estimation was never reached (Davis, 2004). The randomesque-6 and within .10 logits-6 procedures yielded a high number of non-convergent cases due to a high number of inappropriate items that had been administered. Also, the randomesque-6 method produced the largest mean standard error associated with the ability estimates. However, this method (along with the within .10 logits-6 method) utilized more of the item pool than the other methods.

The previously mentioned study by Davis (2004) also investigated the randomesque and within 0.10 logits randomization procedures using two item group sizes of three and six. As with the Burt et al. study, the randomesque-6 procedure yielded a large number of non-convergent cases and seemed to provide high item pool usage. The difference between these studies is that Davis did not find overwhelming support of the Sympson-Hetter methodologies over the randomization procedures. This conclusion provides the basis for using the randomesque procedure to compare against the findings of the stratification procedures to be investigated in this study.

Statement of Purpose

The stratification procedures proposed by Chang and Ying were designed for use in CATs that utilized the theories of dichotomous IRT. Most of the CATs used today still employ dichotomous IRT structures, using the “right vs. wrong” scoring mechanism. However, there has been a growing desire for CATs that use more “performance” items that can give more information about examinees’ abilities than the traditional dichotomously scored items. Over the past decade, research has shifted into investigating ways of incorporating polytomous item structures into computerized adaptive testing. Although great strides have been made in this area of IRT, there still is a lot to be accomplished before polytomous CATs can become widely implemented.

The research studies outlined in the previous section of this paper have demonstrated the continued need for investigations concerning stratification procedures and polytomous CATs. However, the previous research studies in polytomous IRT have only investigated the *a*-stratification procedure. In addition to that procedure, this dissertation will also investigate the *a*-stratification with *b*-blocking procedure due to its previously mentioned advantage over its predecessor.

Hau, Wen, and Chang (2002) perhaps provided the initial impetus for investigating the stratification procedures within polytomous IRT. Although, an optimum number of strata was not found for dichotomous item pools, their study did reveal that increasing the number of strata produced CATs that also increased in efficiency and decreased in overlap rates. However, since dichotomous item pools are typically larger than polytomous item pools, using this many strata on a polytomous item pool would not be feasible. Therefore, it is necessary to perform a similar simulation study to investigate the performance of the stratification procedures using a smaller polytomous item pool. This leads to the first question of the study presented in this dissertation: *Is there an optimum number of strata to employ when using the a-stratification and a-stratification with b-blocking procedures on a polytomous item pool?*

As previously discussed, there have been studies done comparing the stratification procedures to other established methods of exposure control within polytomous IRT, but with mixed results. Pastor, Dodd, and Chang (2002) found that the stratification procedures could work, but not necessarily on the same level as the Sympson-Hetter methodologies. Davis (2004) found that the stratification procedures did not work well, showing large numbers of unused items across the generated CATs. Both of these studies show the weak performance of the stratification procedures in relation to other methods. However, both of these studies did utilize one fixed number of strata for the stratification procedures – five. From this, this dissertation is concerned with judging the performance of the stratification procedures, with a varying number of strata, against another procedure of exposure control: *Using polytomous items, does varying the number of* *strata within the a-stratification and a-stratification with b-blocking procedures help achieve the same level of exposure control as the randomesque procedure?*

Hau, Wen, and Chang (2002) and Pastor, Dodd, and Chang (2002) both examined the effects of item pool size on the performance of the stratification procedure. Both studies found that for the smaller item pools used, the *a*-stratification procedure did produce lower numbers of unadministered items (compared to the no-exposure control condition), but higher indices of item/test overlap. In other words, for Hau, Wen, and Chang, the 200-item pool resulted in the lowest numbers of "underexposed" items over the 400- and 800-item pools, but higher values of test overlap. By the same token, Pastor, Dodd, and Chang found that the 60-item pool produced a lower number of un-administered items than the 100-item pool, but higher values of item overlap. These same results are to be expected in this dissertation for the *a*-stratification, however the *a*-stratification with *b*-blocking procedure will also be investigated to determine the nature of its performance across two item pool sizes: *Does the a-stratification with b-blocking procedure show a similar trend in exposure and overlap rates across two item pool sizes? How do the overlap rates of the stratification procedures compare to the randomesque procedure?*

This dissertation presents a simulation study conducted to investigate the *a*-stratification and *a*-stratification with *b*-blocking procedures of exposure control on polytomous CATs, with items calibrated according to Muraki's generalized partial credit model. The goal of this study was to analyze these stratification procedures of exposure control in terms of the preceding questions.

CHAPTER THREE: METHODOLOGY

Two polytomous item pools were used to compare three exposure control procedures in a simulated CAT: randomesque, a-stratification, and *a*-stratification with *b*-blocking, as well as a “no exposure control” condition. CATs were simulated utilizing these various conditions of these exposure control methods to determine when item exposure is best controlled and the item pool is used most effectively for items calibrated according Muraki’s generalized partial credit model.

Item Pool

The items used in this study were taken from the 1996 National Assessment of Education Progress (NAEP) science assessment (Allen, Carlson, & Zelenak, 1999). The items investigated had three response categories, allowing for two step-difficulty values per item. Although the original set of items contained four-category items, they were not used in this study because of the insufficient number of these types of items and because of the pitfall of meeting both content and category restraints when using the stratification procedures, as outlined by Davis (2004). There were 208 available three-category items within the index of NAEP item parameters, however only 175 of them were used in this simulation. Those items that possessed step-difficulty values greater than +4 and/or less than -4 were cut from the item pool that was to be used in this simulation. This resulted in an item pool that contained items that fit the goal of determining ability estimates between -4 and +4. These items (n=175) covered three content areas of science: physical, earth, and life science. Table 1 shows the number of items in each content area.

Table 1: Number of Items Per Content Area for NAEP Science Items

Content Area		
Physical Science	Earth Science	Life Science
50	63	62

The information in Table 1 was used when specifying the target values for administering items of each content area. A second item pool (n=85) was constructed by randomly selecting items from the 175-item pool so that this second item pool resembled approximately the same balance across content areas as the first. The purpose of the second item pool was to be able to compare analyses of this simulation across two sizes of item pools. Table 2 shows the number of items in each content area for the 85-item pool used in this study.

It is known that the NAEP assessment carries no weight on the actual academic standing of its examinees. In other words, there is no real motivation for the examinees to do their best since the exam does not determine whether or not they actually pass on to the next grade level. Therefore, this exam is considered “low-stakes” and results in item discrimination parameters that are lower than desired for a “high-stakes” exam. Given this, the item discrimination parameter of each item was increased by 0.40 (Burt et al., 2003) so that the items used for this study would reflect those typically used in high-stakes assessments.

Stratification of the Item Pools

This study used a range of the “number of strata” for the a-stratification (AS) and *a*-stratification with *b*-blocking (BAS) methods. Although, the process of stratification for this study is described later, the reasoning for determining the number of strata conditions is worth mentioning. It has been suggested through previous research that the number of strata to use on an item pool is influenced by the structure of the item pool itself (Hau et al., 2002). Since the BAS procedure involves more stratification steps than the AS procedure, the range of “number of strata” to use in the simulations was determined by finding the maximum number of strata for use with this procedure. For this study, the range of the number of strata for the 175-item pool was two to five, given that five was used in previous research studies (Pastor, Dodd, & Chang, 2002; Davis, 2004). Since having one stratum is analogous to the no-exposure control condition, it was not included in the study. Table 3 displays the breakdown of the item pool by content and difficulty level.

Table 2: Number of Items Per Content Area for NAEP Science Items

	Content Area	
Physical Science	Earth Science	Life Science
24	31	30

Table 3: Number of Items Per Content Area and Difficulty Level for NAEP Science Items in the 175-Item Pool

			Content Area	
		Physical Science	Earth Science	Life Science
Difficulty Level	Easy	7	8	21
	Average	11	16	47
	Difficult	32	39	47

The process of determining the difficulty of each item is described later in this section. Given this item pool structure, the first condition for the 175 items used two strata for the item pool, the second condition three strata and so on until the last condition used five strata to group the item pools, for both stratification procedures. In the cases where there was not an even number of items per strata – for example, when there are two strata – the “extra” items were distributed as evenly as possible in the strata of the lower discriminating items to help control the use of items in higher strata.

For the 85-item pool, the maximum number of strata was three since due to its smaller size. This appears to be a reasonable number of strata to use since anything greater than three strata would guarantee some strata would not have enough items to choose from based on difficulty *and* content. Table 4 displays the breakdown of the 85-item pool by content and difficulty level.

Table 4: Number of Items Per Content Area and Difficulty Level for NAEP Science Items in the 85-Item Pool

		Content Area		
		Physical Science	Earth Science	Life Science
Difficulty Level	Easy	3	5	3
	Average	5	8	4
	Difficult	16	19	23

Table 5 and Table 6 show the numbers of items per stratum for each condition of the AS procedure for the 175- and 85-item pool, respectively. This was determined by breaking down each content area into the appropriate number of strata so that each stratum has approximately the same number of items, per content area. Table 7 and Table

Table 5: Number of Items Per Stratum (AS) for the 175-Item Pool

		“Number of Strata” Condition			
		2	**3**	**4**	**5**
Stratum Number	**1**	88	59	45	36
	2	87	59	45	36
	3		57	43	35
	4			42	34
	5				34

Table 6: Number of Items Per Stratum (AS) for the 85-Item Pool

		“Number of Strata” Condition	
		2	**3**
Stratum Number	**1**	43	29
	2	42	28
	3		28

8 show the numbers of items per stratum for each condition of the BAS procedure for 175- and 85-item pool, respectively.

Table 7: Number of Items Per Stratum (BAS) for the 175-Item Pool

		"Number of Strata" Condition			
		2	**3**	**4**	**5**
Stratum Number	**1**	90	61	46	40
	2	85	59	45	37
	3		55	44	34
	4			40	33
	5				31

Table 8: Number of Items Per Stratum (BAS) for the 85-Item Pool

		"Number of Strata" Condition	
		2	**3**
Stratum Number	**1**	45	32
	2	40	28
	3		25

The stratification with content balancing procedure described by Yi and Chang (2003) was used for both stratification procedures in this simulation to ensure that each stratum contains the appropriate balance of science content. More specifically, for the *a*-stratification procedure, the item pool was first grouped by content area, then partitioned

as equally as possible into strata according to the item discrimination values. Similarly, for the *a*-stratification by *b*-blocking procedure, the item pool was first grouped by content area, but then sorted by item difficulty within each content area.
Next, within each difficulty level, the items were sorted into strata according to their item discrimination values.

As previously mentioned, the stratification procedures developed as item exposure control mechanisms were not designed for polytomous items. More important to this study, the stratification of polytomous items becomes difficult when the items have more than one parameter that could be used in the stratification process. For example, polytomous items may have more than one "difficulty" parameter associated with them – a "step difficulty" for each step of the item. Given this, classifying items according to their difficulty is not as straightforward as in the dichotomous case. Also, there is a lack of research on the ways to classify polytomous items according to difficulty. Therefore, the *a*-stratification with *b*-blocking procedure presents a challenge for polytomous items.

To deal with this challenge, the items chosen for this study were classified into three groups: easy, average, and difficult. This distinction was made based on the signs of both step difficulties of an item. If an item had two positive step difficulties, then the item was considered to be "difficult". Conversely, if an item has two negative step difficulties, then it was considered to be "easy." If an item has one positive and one negative step difficulty, then it was considered an item of "average" difficulty. However, it was not an automatic decision to classify the item as "average." This is due to the fact that an item could have had a small positive step difficulty *and* a large negative step difficulty or vice versa. Therefore, when an item was being considered as “average” the difference in distance-from-zero between both step difficulties was analyzed. If this difference was greater than one, then the item was classified according to the sign of the step difficulty furthest from zero. If this difference was less than one, then the item was considered as “average.” For example, if the value of the first step difficulty of an item was -0.085 and that of the second step difficulty was 2.437, then this item was considered “difficult” since the second step difficulty value is further from zero by a difference of 2.352. However, if the first step difficulty value was -0.74 and the second was 0.456, then this item was considered “average” since neither step difficulty value had a distance-from-zero greater than one. Since the BAS method had not been previously tested in research on polytomous items, classifying the difficulty of the items in the manner specified here will provide the information necessary to judge its utility for implementing into operational CAT algorithms.

Simulated Data Generation

Item responses were simulated for samples of n = 1,000 simulees using IRTGEN (Whittaker, Fitzpatrick, Williams, & Dodd, 2003), a computer simulation program for polytomous items. The simulees were obtained from a normal distribution of ability, N(0,1). The ability of each simulee was determined through random assignment prior to the generation of the response vector. To prevent “chance” results among the CAT conditions, each condition used a different sample of n = 1,000 simulees, also drawn from a normal distribution of ability. Also, ten replications were simulated per condition for statistical stability Therefore, for each replication of each CAT condition a different simulee population of 1,000 was used for the item response generation and the subsequent CAT simulation.

CAT Simulations

SAS computer programs (Chen, Hou, & Dodd, 1998) were used in this study to simulate 20-item computer adaptive tests (Davis, 2004; Burt et al., 2003) to each sample of 1,000 simulees. For each simulee, an initial ability level, θ, of zero was specified for all CAT conditions. Maximum likelihood estimation was used to estimate ability levels once a mixed set of responses was obtained (i.e., response in two different categories). Prior to the mixed pattern of responses being obtained, a variable step-size approach was used to estimate the ability level, moving the trait estimate to a quarter of the distance to the most extreme item, within the appropriate content area. Although previous research on the variable step-size suggests using half the distance between the current ability estimate and the most extreme *b*-value, this study used a quarter of the distance between the ability and difficulty indices to prevent high numbers of simulees whose ability levels cannot be estimated through the simulated CATs.

Kingsbury and Zara's CCAT method was utilized for all conditions of this study ensuring that each CAT consisted of the appropriate balance of physical science, earth science, and life science items, as specified by Tables 1 and 2. This procedure selected items from the content area that was farthest below its targeted ideal administration percentage.

Also, Table 9 and Table 10 show the number of items administered from each stratum for the item pools. The number of items to be administered per stratum was determined by the desire to use an even number of items in each stratum, whenever

Table 9: Number of Items Administered Per Stratum for the 175-Item Pool

		“Number of Strata” Condition			
		2	**3**	**4**	**5**
Stratum Number	**1**	10	7	5	4
	2	10	7	5	4
	3		6	5	4
	4			5	4
	5				4

Table 10: Number of Items Administered Per Stratum for the 85-Item Pool

		“Number of Strata” Condition	
		2	**3**
Stratum Number	**1**	10	7
	2	10	7
	3		6

possible. When this was not the case, then the “extra” items were selected evenly from the lower discriminating strata, leaving fewer items to be administered from the higher discriminating strata.

The no-exposure condition utilized maximum information item selection for selecting items to administer during the simulation. The randomesque procedure randomly chose the next item for administration from a group of the *six* most-informative items within the appropriate content area. This number of items was based on the fact that Burt et al. (2003) and Davis (2004) found that the randomesque-6 utilized the item pool as well as or better than the Sympson-Hetter method, a commonly used procedure of exposure control. For each stratum of the stratification procedures, the next item to be administered will be the most informative item from the content area that is furthest below its ideal target rate.

Data Analyses

Several statistical indices were used to investigate the accuracy of the CATs in terms of estimating the latent trait. Descriptive statistics are used to describe the nature of the item discrimination and step difficulty parameters achieved in the simulations. Pearson product-moment correlations between known and estimated thetas describe the accuracy of the computerized adaptive tests. Other statistics were also used to evaluate the accuracy of the CATs in this study. The measure of bias calculated using

$$Bias = \frac{\sum_{k=1}^{n}\left(\hat{\theta}_k - \theta_k\right)}{n} \qquad (18)$$

and the root mean squared error (RMSE)

$$RMSE = \left[\frac{\sum_{k=1}^{n} \left(\hat{\theta}_k - \theta_k \right)^2}{n} \right]^{1/2} \qquad (19)$$

also portray the accuracy of the simulated CATs. In equations 18 and 19 $\hat{\theta}_k$ is the estimate of θ for simulee *k*, θ_k is the known ability for simulee *k*, and *n* represents the total number of simulees. Along with this, mean conditional bias is also used to portray the accuracy of estimation within the simulated CATs (Gorin, Dodd, Fitzpatrick, & Shieh, 2005).

Item exposure rate is computed by dividing the number of times the item was administered by the total number of simulees. These rates are analyzed through frequency distributions and descriptive statistics. Item pool utilization is evaluated through the percentage of items that were never administered throughout the CAT conditions.

Item overlap among the simulees, another indicator of exposure control, was investigated through the use of the simulees' audit trails – the records of items that each simulee was administered. The audit trails were used to compare one simulee's record of items with that of every other simulee. The number of items shared among the simulees was stored in file along with the difference between the known and estimated ability levels for pairs of simulees. This investigation made the distinction between "similar" simulees – simulees whose difference in known ability level was equal to or less than two logits – and "different" simulees – those simulees whose known ability levels differed by more than two logits (Pastor, Chiang, Dodd & Yockey, 1999; Davis, Pastor, Dodd, Chiang, & Fitzpatrick, 2003; Pastor, Dodd, & Chang, 2002). This analysis also made the distinction between “similar” simulees those whose difference in known ability levels is equal to or less than one logit and “different” simulees whose difference in known ability levels differed by more than one logit (Boyd, 2003).

CHAPTER FOUR: RESULTS

Descriptive Statistics for the Item Pools

Table 11 lists the mean, standard deviation, minimum and maximum values for the item discrimination parameter, A, as well as the first and second step difficulty values - SD1 and SD2 - for the 175-item pool. The mean item discrimination parameter of one reflects the adjustment that was made to the item discrimination parameters of the items in the pool prior to their use in this simulation. Table 12 provides the same descriptive statistics for the item parameters according to item content.

Table 13 and Table 14 give descriptive statistics of the item discrimination parameter estimates for each stratum of the *a*-stratification and *a*-stratification by *b*-blocking procedures, respectively. It is important to note the overlap of the item discrimination parameter estimates between the various strata within each strata condition beginning with the *a*-stratification-with-two-strata condition (AS-2). When using the stratification procedures, it is ideal to have the least amount of overlap of item discrimination among the strata (Davis, 2004). Intuitively, this ensures that the items selected from the higher strata are items with higher item discrimination values than those selected from the lower strata.

As an example of this, the AS-2 condition shows that the range of item discrimination values for stratum one is 0.64 to 0.97 while the second stratum has the range of 0.92 to 2.27. This may appear to be minimal overlap among the item

Table 11: Descriptive Statistics for the Item Parameter Estimates for the 175-Item Pool

	A	SD1	SD2
Mean	1.00	0.60	1.09
Standard Deviation	0.24	1.35	1.48
Minimum	0.64	-3.95	-3.63
Maximum	2.27	3.73	3.92

Table 12: Descriptive Statistics for the Item Parameter Estimates of the 175-Item Pool By Content Area

	A	SD1	SD2
Physical Science			
Mean	0.98	0.60	0.82
Standard Deviation	0.19	1.46	1.40
Minimum	0.64	-3.17	-2.26
Maximum	1.64	3.73	3.41
N	50		
Earth Science			
Mean	1.05	0.54	0.73
Standard Deviation	0.28	1.16	1.47
Minimum	0.64	-2.74	-3.63
Maximum	2.27	2.76	3.83
N	63		
Life Science			
Mean	0.97	0.66	1.68
Standard Deviation	0.21	1.47	1.38
Minimum	0.66	-3.95	-1.56
Maximum	1.83	3.49	3.92
N	62		

Table 13: Descriptive Statistics for the Item Discrimination Parameter Estimates of the 175-Item Pool By Strata Condition of the AS Procedure

	Strata	N	Mean	Standard Deviation	Minimum	Maximum
AS-2						
	Stratum 1	88	0.85	0.08	0.64	0.97
	Stratum 2	87	1.15	0.24	0.92	2.27
AS-3						
	Stratum 1	59	0.81	0.08	0.64	0.94
	Stratum 2	59	0.96	0.04	0.88	1.03
	Stratum 3	57	1.24	0.26	1.02	2.27
AS-4						
	Stratum 1	45	0.79	0.07	0.64	0.90
	Stratum 2	45	0.91	0.03	0.85	0.97
	Stratum 3	43	1.01	0.04	0.93	1.09
	Stratum 4	42	1.31	0.27	1.05	2.27
AS-5						
	Stratum 1	36	0.77	0.07	0.64	0.89
	Stratum 2	36	0.89	0.03	0.83	0.96
	Stratum 3	35	0.96	0.03	0.90	1.01
	Stratum 4	34	1.04	0.04	0.99	1.14
	Stratum 5	34	1.36	0.27	1.10	2.27

Table 14: Descriptive Statistics for the Item Discrimination Parameter Estimates of the 175-Item Pool By Strata Condition of the BAS Procedure

	Strata	N	Mean	Standard Deviation	Minimum	Maximum
BAS-2						
	Stratum 1	90	0.86	0.09	0.64	1.01
	Stratum 2	85	1.15	0.25	0.92	2.27
BAS-3						
	Stratum 1	61	0.82	0.08	0.64	0.94
	Stratum 2	59	0.96	0.05	0.87	1.07
	Stratum 3	55	1.25	0.26	0.99	2.27
BAS-4						
	Stratum 1	46	0.80	0.08	0.64	0.93
	Stratum 2	45	0.92	0.05	0.81	1.01
	Stratum 3	44	1.02	0.07	0.92	1.20
	Stratum 4	40	1.31	0.28	1.00	2.27
BAS-5						
	Stratum 1	40	0.78	0.07	0.64	0.92
	Stratum 2	37	0.90	0.04	0.81	0.99
	Stratum 3	34	0.96	0.04	0.90	1.04
	Stratum 4	33	1.08	0.09	0.97	1.35
	Stratum 5	31	1.35	0.30	1.02	2.27

discrimination parameter estimates, therefore not posing any substantial deficiency in item selection among the strata. Unfortunately, there is not an accepted standard for item discrimination overlap among strata, therefore, without any *substantial* overlaps, one cannot accurately speculate to when any overlaps affect the item selection processes of the stratification procedures.

Table 15 lists the mean, standard deviation, minimum and maximum values for the item parameter estimates for the 85-item pool while Table 16 provides the same descriptive statistics for the item parameters according to item content for the same item pool.

Table 17 and Table 18 show descriptive statistics for the item discrimination parameter estimates across the strata conditions for the *a*-stratification and *a*-stratification by *b*-blocking procedures. Even though it was mentioned earlier that it is hard tell when item discrimination overlap among the strata may be greater than desirable, the data presented in these two tables highlight overlap that just may count as far from desirable.

For example, for the BAS-2 condition, stratum one has an item discrimination range of 0.64 to 1.03 while stratum two possesses the range from 0.72 to 2.27. Now this may be considered as a substantial amount of overlap of item discrimination estimates between the two strata since a) the spread of item discrimination parameters in stratum one is large *and* b) the lowest item discrimination value of stratum two, 0.72, is not far from the minimum value of stratum one. Therefore, this condition possesses two strata that contain a large number of the same item parameter estimates, leading to the speculation that items selected from stratum two may possess the same item

Table 15: Descriptive Statistics for the Item Parameter Estimates of the 85-Item Pool

	A	SD1	SD2
Mean	1.01	0.72	1.08
Standard Deviation	0.25	1.34	1.54
Minimum	0.64	-3.95	-3.63
Maximum	2.27	3.73	3.83

Table 16: Descriptive Statistics for the Item Parameter Estimates of the 85-Item Pool By Content Area

	A	SD1	SD2
Physical Science			
Mean	1.01	0.62	1.19
Standard Deviation	0.20	1.39	1.44
Minimum	0.68	-1.48	-2.46
Maximum	1.64	3.73	3.41
N	24		
Earth Science			
Mean	1.04	0.61	0.55
Standard Deviation	0.32	1.15	1.72
Minimum	0.64	-2.73	-3.63
Maximum	2.27	2.76	3.83
N	31		
Life Science			
Mean	0.96	0.91	1.54
Standard Deviation	0.19	1.50	1.28
Minimum	0.67	-3.95	-0.95
Maximum	1.45	3.49	3.72
N	30		

Table 17: Descriptive Statistics for the Item Discrimination Parameter Estimates of the 85-Item Pool By Strata Condition of the AS Procedure

	Strata	N	Mean	Standard Deviation	Minimum	Maximum
AS-2						
	Stratum 1	43	0.86	0.09	0.64	0.97
	Stratum 2	42	1.16	0.27	0.93	2.27
AS-3						
	Stratum 1	29	0.82	0.09	0.64	0.94
	Stratum 2	28	0.96	0.04	0.86	1.03
	Stratum 3	28	1.24	0.30	1.02	2.27

Table 18: Descriptive Statistics for the Item Discrimination Parameter Estimates of the 85-Item Pool By Strata Condition of the BAS Procedure

	Strata	N	Mean	Standard Deviation	Minimum	Maximum
BAS-2						
	Stratum 1	45	0.90	0.10	0.64	1.03
	Stratum 2	40	1.13	0.31	0.72	2.27
BAS-3						
	Stratum 1	32	0.88	0.10	0.64	0.99
	Stratum 2	28	0.97	0.10	0.71	1.10
	Stratum 3	25	1.21	0.36	0.72	2.27

discrimination characteristics as those from stratum one. If this holds to be true, then the stratification procedure has not chosen items by its basic principle: selecting items of higher discrimination values from the higher strata. The BAS-3 condition also shows substantial overlap in item discrimination.

Descriptive Statistics of the CAT Simulations

Tables 19 and 20 provide the grand mean, standard error of the mean, minimum and maximum vales of the estimated thetas for the 175- and 85 item pool, respectively. These values are based on the ten replications conducted on each of the exposure control conditions. The number of nonconvergent cases, across all ten replications, is also shown for each exposure control condition. These cases are defined as simulees whose final ability estimates were not obtained through the simulated CATs. This could occur for one of two reasons: 1) the ability estimate terminated at value exceeding the -4.0 to +4.0 interval; or 2) the precision of ability estimation became 9.9 during the simulated CAT, thus causing the ability estimation process to terminate for that simulee. In either case, final ability estimates were not obtained for these simulees and were not used in the statistical analysis of the obtained parameters. From the information in these tables, the grand means of estimated thetas estimated occurred near zero for each of the exposure conditions, across both item pools.

Tables 21 and 22 provide the descriptive statistics of the standard deviations of the estimated thetas for both item pools used in this study. For both item pools, the mean standard deviations of estimated thetas achieved by the exposure control conditions are close to 1, with the greatest mean at 1.073. Therefore, these statistics along with those of

Table 19: Descriptive Statistics of the Estimated Thetas and the Number of Nonconvergent Cases of the Exposure Control Conditions for the 175-Item Pool Across Ten Replications

Exposure Control Condition	Grand Mean	Standard Error of the Mean	Minimum	Maximum	Nonconvergent Cases
No Exposure Control	-0.012	0.011	-0.062	0.028	1
Randomesque-6	-0.018	0.008	-0.067	0.019	15
AS-2	-0.018	0.010	-0.082	0.022	2
AS-3	-0.005	0.009	-0.043	0.048	1
AS-4	0.003	0.007	-0.049	0.032	2
AS-5	-0.010	0.011	-0.048	0.061	2
BAS-2	0.007	0.011	-0.040	0.060	2
BAS-3	-0.012	0.010	-0.070	0.039	1
BAS-4	-0.007	0.013	-0.082	0.055	1
BAS-5	0.001	0.011	-0.039	0.067	1

Table 20: Descriptive Statistics of the Estimated Thetas and the Number of Nonconvergent Cases of the Exposure Control Conditions for the 85-Item Pool Across Ten Replications

Exposure Control Condition	Grand Mean	Standard Error of the Mean	Minimum	Maximum	Nonconvergent Cases
No Exposure Control	0.002	0.015	-0.058	0.081	1
Randomesque-6	-0.030	0.013	-0.091	0.027	13
AS-2	-0.019	0.014	-0.099	0.028	1
AS-3	-0.037	0.009	-0.076	0.010	1
BAS-2	-0.015	0.008	-0.050	0.037	12
BAS-3	-0.002	0.008	-0.032	0.056	1

Table 21: Descriptive Statistics of the Standard Deviations of the Estimated Thetas for the 175-Item Pool Across Ten Replications

Exposure Control Condition	Mean	Minimum	Maximum
No Exposure Control	1.034	1.003	1.079
Randomesque-6	1.043	0.997	1.081
AS-2	1.044	1.002	1.078
AS-3	1.037	0.994	1.080
AS-4	1.050	1.004	1.086
AS-5	1.043	1.011	1.083
BAS-2	1.042	1.002	1.086
BAS-3	1.034	0.988	1.067
BAS-4	1.038	0.989	1.073
BAS-5	1.038	1.017	1.070

Table 22: Descriptive Statistics of the Standard Deviations of the Estimated Thetas for the 85-Item Pool Across Ten Replications

Exposure Control Condition	Mean	Minimum	Maximum
No Exposure Control	1.049	1.003	1.102
Randomesque-6	1.073	1.023	1.105
AS-2	1.058	1.043	1.077
AS-3	1.042	1.023	1.101
BAS-2	1.058	1.016	1.118
BAS-3	1.060	1.030	1.098

the mean estimated thetas show that the distribution of the ability estimates obtained through the simulations of this study is normal – possessing a mean of 0 and a standard deviation of 1.

The standard errors of the estimated thetas show the degree of precision to which the thetas were estimated. Table 23 shows that the no exposure control condition produced the lowest grand mean of standard errors (0.261) while the BAS-4 and BAS-5 produced the highest (0.284). This means that the no exposure control condition had better measurement precision than the other conditions. However, the difference between these minimum and maximum values is not substantial enough to raise concern, leading to the conclusion that all of the exposure control conditions performed to about the same degree of measurement. Table 24 shows different results for the 85-item pool. For this item pool, the no exposure control condition did produce the lowest grand mean of standard errors (0.283), but the randomesque procedure yielded the highest grand mean of standard errors (0.309). Again, the difference between these values does not warrant any concern. Therefore, as with the 175-item pool, the exposure control conditions appear to have performed with the same degree of measurement precision.

One index of the accuracy of the simulated CATs is the correlation between the thetas estimated from the simulated CATs and the “known” thetas specified for each simulee prior to the CATs. Tables 25 and 26 show the descriptive statistics for these correlations for the 175- and 85-item pool, respectively. For the 175-item pool, the means of the correlations range from 0.96 to 0.97, while the means from the 85-item pool range

Table 23: Descriptive Statistics of the Standard Errors for the 175-Item Pool Across Ten Replications

Exposure Control Condition	Grand Mean	Minimum	Maximum
No Exposure Control	0.261	0.259	0.263
Randomesque-6	0.278	0.274	0.280
AS-2	0.275	0.274	0.278
AS-3	0.281	0.279	0.284
AS-4	0.283	0.281	0.285
AS-5	0.287	0.285	0.289
BAS-2	0.274	0.282	0.277
BAS-3	0.279	0.276	0.281
BAS-4	0.284	0.283	0.286
BAS-5	0.284	0.282	0.286

Table 24: Descriptive Statistics of the Standard Errors for the 85-Item Pool Across Ten Replications

Exposure Control Condition	Grand Mean	Minimum	Maximum
No Exposure Control	0.283	0.280	0.286
Randomesque-6	0.309	0.305	0.311
AS-2	0.295	0.293	0.298
AS-3	0.301	0.299	0.303
BAS-2	0.294	0.291	0.296
BAS-3	0.298	0.296	0.299

Table 25: Descriptive Statistics of the Pearson Correlations Between Known and Estimated Thetas for the 175-Item Pool Across Ten Replications

Exposure Control Condition	Mean	Minimum	Maximum
No Exposure Control	0.965	0.960	0.696
Randomesque-6	0.956	0.940	0.966
AS-2	0.962	0.958	0.966
AS-3	0.962	0.959	0.968
AS-4	0.963	0.959	0.968
AS-5	0.958	0.955	0.962
BAS-2	0.958	0.915	0.967
BAS-3	0.961	0.956	0.966
BAS-4	0.957	0.926	0.963
BAS-5	0.960	0.957	0.964

Table 26: Descriptive Statistics of the Pearson Correlations Between Known and Estimated Thetas for the 85-Item Pool Across Ten Replications

Exposure Control Condition	Mean	Minimum	Maximum
No Exposure Control	0.960	0.957	0.964
Randomesque-6	0.946	0.935	0.953
AS-2	0.957	0.955	0.961
AS-3	0.955	0.952	0.960
BAS-2	0.960	0.955	0.966
BAS-3	0.958	0.956	0.962

from 0.95 to 0.96. Given these ranges of correlations between the estimated and known thetas, it appears that the exposure control conditions all performed to the same degree of accuracy, between the item pools as well as within them.

Bias and root mean squared error (RMSE) are other indicators CAT accuracy used in this study. Tables 27 and 28 display the mean, minimum, and maximum values of these indices for both item pools. The mean bias estimates were very similar across both item pools, being 0.01 or 0.02 – near zero. Since bias quantifies the difference between the known and estimated thetas, the values presented in tables 27 and 28 show relatively no bias in the estimation of ability. Appendix A contains plots of conditional bias for each of the CAT conditions simulated for both item pools. These plots portray the accuracy of the CATs at sixteen discrete intervals along the ability continuum. The general impression from these plots is that ability estimation becomes more accurate toward the center of the ability scale and less accurate at the extreme values of ability.

Exposure Rates and Pool Utilization

As previously defined, the item exposure rate is the number of times an item is given divided by the total number of test takers, simulees in this case. Tables 29 and 30 provide the grand mean, mean minimum, and mean maximum exposure rates for the simulated CATs for both item pools. The grand mean of the exposure rates for each CAT condition is of little importance since they are all the same - 0.114 for the 175-item pool and 0.235 for the 85-item pool. This occurs because the mean exposure rate simply reflects the ratio of test length to item pool size. Also, as the results reveal, the minimum exposure rate achieved through all of the CAT conditions, in both item pools, is 0.00.

Table 27: Descriptive Statistics of the Bias and Root Mean Squared Error (RMSE) for the 175-Item Pool Across Ten Replications

Exposure Control Condition	Mean Bias (Min, Max)	Mean RMSE (Min, Max)
No Exposure Control	0.013 (0.008, 0.020)	0.270 (0.261, 0.277)
Randomesque-6	0.011 (-0.002, 0.036)	0.304 (0.276, 0.371)
AS-2	0.005 (-0.024, 0.016)	0.285 (0.270, 0.299)
AS-3	0.005 (-0.014, 0.019)	0.283 (0.269, 0.294)
AS-4	0.004 (-0.015, 0.020)	0.284 (0.273, 0.295)
AS-5	0.007 (-0.002, 0.018)	0.298 (0.285, 0.313)
BAS-2	0.004 (-0.011, 0.015)	0.297 (0.268, 0.430)
BAS-3	0.012 (0.001, 0.026)	0.285 (0.271, 0.297)
BAS-4	0.010 (-0.001, 0.018)	0.290 (0.282, 0.298)
BAS-5	0.010 (0.003, 0.016)	0.291 (0.279, 0.304)

Table 28: Descriptive Statistics of the Bias and Root Mean Squared Error (RMSE) for the 85-Item Pool Across Ten Replications

Exposure Control Condition	Mean Bias (Min, Max)	Mean RMSE (Min, Max)
No Exposure Control	0.013 (0.004, 0.025)	0.293 (0.281, 0.307)
Randomesque-6	0.014 (-0.004, 0.023)	0.348 (0.321, 0.386)
AS-2	0.018 (0.007, 0.035)	0.307 (0.292, 0.316)
AS-3	0.013 (-0.012, 0.026)	0.309 (0.289, 0.321)
BAS-2	0.008 (-0.003, 0.018)	0.298 (0.290, 0.312)
BAS-3	0.005 (-0.008, 0.022)	0.304 (0.297, 0.313)

Table 29: Descriptive Statistics of the Exposure Rates for the 175-Item Pool Across Ten Replications

Exposure Control Condition	Grand Mean	Mean Minimum Exposure Rate	Mean Maximum Exposure Rate
No Exposure Control	0.114	0.000	0.859
Randomesque-6	0.114	0.000	0.608
AS-2	0.114	0.000	0.778
AS-3	0.114	0.000	0.799
AS-4	0.114	0.000	0.799
AS-5	0.114	0.000	0.690
BAS-2	0.114	0.000	0.787
BAS-3	0.114	0.000	0.817
BAS-4	0.114	0.000	0.740
BAS-5	0.114	0.000	0.822

Table 30: Descriptive Statistics of the Exposure Rates for the 85-Item Pool Across Ten Replications

Exposure Control Condition	Grand Mean	Mean Minimum Exposure Rate	Mean Maximum Exposure Rate
No Exposure Control	0.235	0.000	0.933
Randomesque-6	0.235	0.000	0.672
AS-2	0.235	0.000	0.857
AS-3	0.235	0.000	0.817
BAS-2	0.235	0.000	0.832
BAS-3	0.235	0.000	0.893

This means that some items were never administered in the simulated CATs. Therefore, the important information from these tables is that of the maximum exposure rates.

For the 175-item pool, the no exposure control condition produced the largest mean maximum exposure rate (0.859), as expected, while the randomesque-6 procedure yielded the lowest maximum exposure rate (0.608). Similarly, the 85-item pool yielded the largest mean maximum exposure rate for the no exposure control condition (0.933) and the lowest for the randomesque-6 condition (0.672). What should be surprising is that the BAS procedures yielded larger maximum exposure rates than the AS procedures. The reason that this is surprising is because the BAS procedures were designed to reduce the exposure rates, in comparison to the AS procedure, by taking into account any correlation that might exist between the discrimination and difficulty parameters. However, this should not be surprising given that Davis (2004) found that employing multiple stratification techniques leads to the poor performance of the stratification procedures.

The standard deviation of the exposure rates reveals how evenly the items within the pool were used. Tables 31 and 32 give the descriptive statistics of the standard deviations of the exposure rates for the 175- and 85-item pool, respectively. From table 30, it appears that the randomesque-6 procedure yielded the most even use of the items with a standard deviation of exposure rates of 0.155, while the no exposure control condition yielded the most uneven use of items with a standard deviation of exposure rates of 0.197. The high standard deviation of exposure rates for the no exposure control procedure is to be expected. Similarly for the 85-item pool, the randomesque-6 condition had the lowest standard deviation of exposure rates (0.206) while the no exposure control

Table 31: Descriptive Statistics of the Standard Deviations of the Exposure Rates for the 175-Item Pool Across Ten Replications

Exposure Control Condition	Mean	Minimum	Maximum
No Exposure Control	0.197	0.194	0.199
Randomesque-6	0.155	0.154	0.157
AS-2	0.180	0.118	0.188
AS-3	0.182	0.179	0.184
AS-4	0.177	0.174	0.179
AS-5	0.181	0.179	0.183
BAS-2	0.189	0.187	0.190
BAS-3	0.182	0.180	0.185
BAS-4	0.183	0.179	0.188
BAS-5	0.186	0.184	0.189

Table 32: Descriptive Statistics of the Standard Deviations of the Exposure Rates for the 85-Item Pool Across Ten Replications

Exposure Control Condition	Mean	Minimum	Maximum
No Exposure Control	0.271	0.268	0.277
Randomesque-6	0.206	0.200	0.211
AS-2	0.257	0.254	0.259
AS-3	0.241	0.235	0.245
BAS-2	0.258	0.253	0.262
BAS-3	0.239	0.234	0.246

condition had the highest (0.271). Since the stratification procedures yielded mean standard deviations much greater than those of the randomesque-6 procedure, it is inferred that the randomesque-6 procedure used the items more evenly than the stratification procedure.

This result is quite surprising given that the stratifications procedures were meant to use items more evenly than a randomized approach. By placing items into strata, the procedures were design to “guarantee” a more even use of the items over random approaches. Another surprising result is that there was not a clear pattern of item usage as the number of strata changed within the stratification procedures using the 175-item pool. However, for the 85-item pool, the items appear to be more evenly used with three strata rather than two, for both stratification procedures.

Tables 33 and 34 provide information on the mean item pool usage for the item pools for each of the CAT conditions, respectively. This table shows the distribution of mean exposure rates for each item within the item pools, the mean number of un-administered items, and the mean percentage of the item pool that was not administered for each of the CAT conditions. Perhaps the most important piece of information within these tables is the number, and percentage, of items that were never administered. For the 175-item pool, the randomesque-6 procedure outperformed the other procedures given that it achieved the lowest grand mean number of unused items, 57 (33%). Although the no exposure control condition produced the largest mean number of unused items, 84 (48%), the numbers achieved by the stratification procedures were not that far behind, ranging from 75 to 83 (43 to 47%). Similar results were found using the 85-item pool.

Table 33: Mean Frequency of Exposure Rates for the 175-Item Pool Averaged Across Ten Replications

Exposure Rate	No Exposure Control	Randomesque-6	AS-2	AS-3	AS-4	AS-5	BAS-2	BAS-3	BAS-4	BAS-5
1	0	0	0	0	0	0	0	0	0	0
.91-.99	0	0	0	0	0	0	0	0	0	0
.81-.90	3	0	0	1	0	2	0	1	1	1
.71-.80	2	0	3	2	1	0	2	1	2	2
.61-.70	1	1	5	2	2	4	6	3	3	4
.51-.60	6	3	3	7	9	8	5	8	5	5
.41-.50	9	9	8	9	5	7	9	9	8	5
.36-.40	4	8	3	2	7	3	2	1	3	7
.31-.35	3	7	8	3	5	7	3	3	5	5
.26-.30	5	12	7	7	7	6	7	6	6	4
.21-.25	8	6	4	8	7	5	5	8	7	6
.16-.20	5	7	7	9	6	6	9	8	8	9
.11-.15	7	9	5	7	6	6	6	7	7	6
.06-.10	8	15	11	11	15	13	12	13	11	13
.01-.05	31	42	32	27	30	30	27	28	34	26
0.0	84	57	81	81	77	80	82	79	75	83
% Not Administered	48	33	46	46	44	46	47	45	43	47

Table 34: Mean Frequency of Exposure Rates for the 85-Item Pool Averaged Across Ten Replications

Exposure Rate	No Exposure Control	Randomesque-6	AS-2	AS-3	BAS-2	BAS-3
1	0	0	0	0	0	0
.91-.99	1	0	0	0	0	0
.81-.90	3	0	3	2	3	1
.71-.80	3	0	3	3	4	3
.61-.70	4	3	6	5	6	4
.51-.60	7	9	3	4	2	5
.41-.50	6	11	8	7	5	6
.36-.40	3	5	3	3	4	5
.31-.35	2	4	3	6	7	4
.26-.30	4	5	3	7	3	6
.21-.25	2	3	5	3	5	4
.16-.20	4	4	5	5	3	4
.11-.15	7	8	8	7	10	8
.06-.10	7	9	15	14	14	14
.01-.05	15	22	16	15	17	15
0.0	19	3	16	15	17	15
% Not Administered	22	3	19	17	20	17

The randomesque-6 procedure yielded the smallest grand mean number of unused items, 3 (3%), while the no exposure control condition yielded the largest grand mean, 19 (22%). Again, the stratification procedures yielded item pool usage rates near that achieved by the no exposure control condition.

The poor performance of the stratification procedures, in relation to the randomesque-6 procedure, reveals that substantial portions of the 175-item pool were not used for the CATs – nearly 50%. In an operational CAT, this can be quite disappointing given the amount of work, time, and money that goes into planning, designing and maintaining an item pool. However, as seen with the 85-item pool, the stratification procedures did utilize at least 75% of the item pool, although this was not at the same level as the randomesque-6. From this, it can be inferred that the stratification procedures do a better job at utilizing items from smaller polytomous item pools, a finding that is consistent with results from previously mentioned studies.

Item Overlap

As another indicator of item exposure, item overlap was analyzed to determine which exposure control condition presented the highest (and lowest) degree of item overlap among the simulees. As well as determining the overall average item overlap among all of the simulees, this investigation also looked at the item overlap rates for four comparisons: 1) simulees whose known abilities differed by more than two logits; 2) simulees whose known abilities differed by less than two logits; 3) simulees whose known abilities differed by more than one logit; and 4) simulees whose known abilities differed by less than one logit. The comparisons among the simulees were such that each simulee was compared with every other simulee for each exposure control condition.

Table 35 provides the average item overlap results of the 175-item pool, for when simulees' known abilities differed by two logits.

The lowest average of overall item overlap for the 175-item pool occurs for the randomesque procedure, six, while the no exposure control condition presents the highest average overall item overlap, nine. The stratification procedures show that eight items, on average, were shared among the simulees. Similar results were also found when the known abilities differed by less than two logits, The randomesque-6 procedure revealed seven items shared on average while the no exposure control condition produced ten. Nearly all of the stratification procedures showed nine items shared among the simulees, with the BAS-2 procedure showing ten (within rounding). When the known abilities differed by more than two logits the average number of shared items drops, but most of the conditions revealed the same number of shared items. Within rounding, the no exposure control, randomesque-6, AS-2, AS-3, AS-4, AS-5, BAS-2, and BAS-3 show that two items, on average, was shared among the simulees. However, the BAS-4 and BAS-5 procedures revealed three items shared, on average.

Table 36 shows the presents the same descriptive information for the 85-item pool. Again, the randomesque-6 procedure showed the lowest grand mean overall item overlap, the no exposure control condition revealed the highest, and the stratification procedure fell in the middle. For the simulees whose abilities differed by less than two logits, the randomesque-6 still performed better than the other exposure control methods by at least two items, within rounding. However, when the abilities differed by more than two logits,

Table 35: Descriptive Statistics of Item Overlap for the 175-Item Pool Across Ten Replications When Defining Ability Groups by Two Logits

Exposure Control Condition	Overall Overlap Grand Mean (Min, Max)	Similar Abilities Grand Mean (Min, Max)	Different Abilities Grand Mean (Min, Max)
No Exposure Control	9.010 (8.813, 9.010)	10.211 (10.122, 10.305)	2.367 (2.223, 2.579)
Randomesque-6	6.448 (6.373, 6.565)	7.259 (1.226, 7.296)	1.981 (1.788, 2.170)
AS-2	8.328 (8.210, 8.456)	9.380 (9.156, 9.484)	2.203 (2.022, 2.337)
AS-3	8.049 (7.844, 8.195)	9.104 (8.968, 9.210)	2.203 (2.064, 2.364)
AS-4	7.662 (7.184, 7.867)	8.780 (8.680, 8.923)	2.071 (1.909, 2.260)
AS-5	7.989 (7.876, 8.069)	9.008 (8.871, 9.096)	2.336 (2.120, 2.544)
BAS-2	8.457 (8.343, 8.584)	9.590 (9.547, 9.689)	2.294 (2.178, 2.412)
BAS-3	8.098 (7.913, 8.789)	9.107 (9.006, 9.253)	2.173 (1.928, 2.389)
BAS-4	8.095 (7.855, 8.419)	9.075 (8.819, 9.310)	2.578 (2.290, 2.847)
BAS-5	8.309 (8.163, 8.490)	9.229 (8.223, 9.489)	3.257 (2.518, 8.168)

Table 36: Descriptive Statistics of Item Overlap for the 85-Item Pool Across Ten Replications When Defining Ability Groups by Two Logits

Exposure Control Condition	Overall Overlap Grand Mean (Min, Max)	Similar Abilities Grand Mean (Min, Max)	Different Abilities Grand Mean (Min, Max)
No Exposure Control	10.861 (10.711, 11.139)	12.165 (12.012, 12.350)	3.386 (3.573, 3.827)
Randomesque-6	8.265 (8.069, 8.420)	9.077 (8.949, 9.209)	3.945 (3.778, 4.101)
AS-2	10.246 (10.130, 10.323)	11.475 (11.379, 11.573)	3.362 (3.447, 3.881)
AS-3	9.580 (9.349, 9.719)	10.654 (10.512, 10.829)	3.586 (3.205, 3.756)
BAS-2	10.295 (10.084, 10.452)	11.525 (11.336, 11.639)	3.627 (3.353, 3.918)
BAS-3	9.504 (9.312, 9.773)	10.630 (10.493, 10.835)	3.594 (3.326, 3.780)

the no exposure control condition and the AS-2 both revealed that three items were shared on average, the lowest of the exposure control condition.

Tables 37 and 38 present the descriptive statistics for item overlap for both item pools when the differences in simulees' abilities are defined by one logit. For the 175-item pool and abilities differing by less than one logit, the randomesque-6 procedure outperformed the other conditions by showing only eight items shared, on average. The other conditions showed ranged from eleven to thirteen shared items. When the abilities differed by more than one logit for the same item pool, the randomesque-6 and AS-2 showed only four shared items while the other conditions showed five.

The 85-item pool revealed larger numbers of shared items than the 175-item pool as it did when the ability groups were defined by two logits. From table 38, the randomesque-6 showed only ten shared items, on average, when the abilities differed by less than one logit while the other exposure control conditions showed between thirteen and fifteen shared items. When the abilities differed by more than one logit, however, the range of shared items was only six to seven items.

Table 37: Descriptive Statistics of Item Overlap for the 175-Item Pool Across Ten Replications When Defining The Ability Groups by One Logit

Exposure Control Condition	Similar Abilities Grand Mean (Min, Max)	Different Abilities Grand Mean (Min, Max)
No Exposure Control	12.828 (12.714, 12.897)	4.819 (4.729, 4.926)
Randomesque-6	8.716 (8.650, 8.773)	3.951 (3.848, 4.029)
AS-2	11.721 (11.582, 11.816)	4.627 (4.454, 4.738)
AS-3	11.230 (11.118, 11.331)	4.504 (4.353, 4.637)
AS-4	10.874 (10.780, 11.027)	4.283 (4.112, 4.426)
AS-5	11.038 (10.922, 11.190)	4.628 (4.427, 4.765)
BAS-2	11.835 (11.744, 11.920)	4.771 (4.644, 4.908)
BAS-3	11.250 (11.118, 11.367)	4.500 (4.332, 4.714)
BAS-4	11.080 (10.911, 11.368)	4.775 (4.439, 5.175)
BAS-5	11.031 (8.234, 11.460)	5.334 (4.778, 8.193)

Table 38: Descriptive Statistics of Item Overlap for the 85-Item Pool Across Ten Replications When Defining The Ability Groups by One Logit

Exposure Control Condition	Similar Abilities Mean (Min, Max)	Different Abilities Mean (Min, Max)
No Exposure Control	14.527 (14.363, 14.619)	6.841 (6.698, 7.078)
Randomesque-6	10.168 (10.089, 10.285)	6.205 (6.019, 6.377)
AS-2	13.509 (13.437, 13.657)	6.706 (6.548, 6.925)
AS-3	12.510 (12.399, 12.699)	6.331 (6.063, 6.563)
BAS-2	13.628 (13.431, 13.931)	6.698 (6.469, 6.877)
BAS-3	12.509 (12.426, 12.640)	6.273 (6.027, 6.433)

CHAPTER FIVE: DISCUSSION

The simulations within this study have provided some useful information in terms of using stratification procedures as mechanisms of item exposure control. The benefit of what has been presented here is that, unlike previous research studies involving the utility of the stratification procedures, this study investigated the various number of strata that could be used within an item pool and then compared those results to an already established method of exposure control. Also, this study provided, perhaps, the first procedure of blocking polytomous items that have more than one difficulty parameter into groups of different difficulties, something that had not been explored in-depth before.

This chapter will discuss the results of the CAT simulations in relation to the three research questions outlined in chapter three. Following the results, a brief discussion of improvements that have been made to the stratification procedures will be presented along with previous research studies investigating them. Next, limitations of this study will be highlighted with possible solutions to overcome them in future studies. Lastly, this chapter will cover suggestions for future research in the area of stratification and polytomous items.

Research Questions

Is there an optimum number of strata to employ when using the a-stratification and a-stratification with b-blocking procedures on a polytomous item pool? In terms of CAT efficiency and item pool utilization, the results from this study indicate that there is not an optimum number of strata to use when using the stratification procedures on polytomous items. This is not surprising given that the same conclusion was reached when dichotomous items were used. The difference between the investigation of dichotomous items and this investigation of polytomous items is that patterns of increased efficiency and overlap rates were found using dichotomous items but not polytomous items. It is possible that from this finding that the stratification procedures investigated in this study were not designed to handle the characteristics of the polytomous items used.

Using polytomous items, does varying the number of strata within the a-stratification and a-stratification with b-blocking procedures help achieve the same level of exposure control as the randomesque procedure? In general, the randomesque-6 procedure performed better than the stratification procedures regardless of the number of strata that was used. This is quite surprising given that the randomesque procedure is based on a randomized approached of selecting items for administration and the stratification procedures are not. The stratification procedures were designed to control the item selection mechanisms better by placing the items into groups and specifying, within the testing algorithm, how to select items from each group. In short, this mechanism should have produced better exposure rates than that found using the randomesque procedure.

A surprising result from this study is the performance of the *a*-stratification with *b*-blocking procedure itself. As discussed before, this procedure was designed to control item exposure better than the a-stratification procedure since it takes into account any positive correlation that might exist between the item discrimination and item difficulty parameters of an item pool. However, results of this study did not necessarily reveal that. In some cases, this procedure performed about the same as the AS procedure, sometimes worse, but not better. This supports the result of Davis (2004) in which it was found that using multiple stratification procedures on polytomous items can lead to the poor performance of the stratification procedure. Even though Davis was referring to stratification in terms of item content and the number of score categories the items had, the premise still holds here: stratifying on item content and item difficulty probably had an effect on the overall performance of the procedure.

When using the a-stratification and a-stratification with b-blocking procedures, does the smaller item pool produce lower numbers of un-administered items but higher overlap rates than the larger item pool? How do the overlap rates of the stratification procedures compare to the randomesque procedure? Both stratification procedures exhibited better rates of item pool utilization for the 85-item pool over the 175-item pool. For the *a*-stratification procedure, the average percentage of un-administered items for the 175-item pool was 46% while it was 18% for the 85-item pool. Similarly, for the *a*-stratification with *b*-blocking procedure, the 175-item pool yielded an average of 46% items never administered, but only 19% for the 85-item pool. The number of strata used for both item pools does not affect these results since the size of the item pools was taken into account when deciding upon the maximum number of strata to use on both item pools. Item overlap rates for the 175-item pool were generally lower than the 85-item pool across all comparisons.

Generally, the randomesque-6 procedure yielded better item overlap rates than the stratification procedures. In terms of overall item overlap, abilities differing by less than two logits, and abilities differing by less than one logit, the randomesque-6 procedure handled item overlap better than the stratification procedures. When the abilities differed by more than two logits (and, subsequently, by more than one logit) the superiority in item overlap was not achieved by the randomesque-6 procedure. Therefore, the randomesque-6 procedure outperformed the stratification procedures in item overlap when investigating simulees that were considered to be “similar” in abilities but not when simulees were considered to be “different.”

A possible explanation of the superior performance of the randomesque procedure over the stratification procedures may lie in the availability of items throughout the CATs. In the randomesque procedure, the entire item pool is available when selecting the *n* most informative items from which one will be randomly chosen for the simulee. This is the case throughout the entire CAT. However, the stratification procedures are restricted to which items can be used for selection. In other words, if there are not any appropriate items to administer to a simulee from a particular stratum, the CAT must still use items from the stratum until the maximum number of items administered from that stratum has been reached. Therefore, the CAT cannot proceed to another stratum in the event that there are not any appropriate items available for a simulee. This possibly allows the randomesque procedure to use items more effectively than the stratification procedures.

Although this study did reveal a superior performance of the randomesque procedure in terms of item exposure control and item overlap, this does not mean that the randomesque procedure is the *best* procedure of exposure control. Other research studies using polytomous CATs have found that other methods of exposure control can perform at the same level or better than the randomesque procedure. For example, Burt et al. found that a modified within .10 logits procedure, with an item group size of 6, used approximately the same amount of the item pool as the randomeque-6 procedure. In fact, these two procedures utilized the item pool more than the other procedures. Davis (2002) found similar results for the randomesque and modified within .10 logits procedures. However, the conditional Sympson-Hetter procedure also produced a high rate of item pool utilization, similar to the previous two procedures.

Boyd (2003), however, found different results regarding the best method of exposure control. Investigating the optimal method of exposure control for CAT systems based on the three-parameter logistic testlet response theory and systems based on the partial credit model, it was found that the progressive restricted procedures, restricted to a maximum exposure rate of .20 or .30 yielded the best results. These procedures utilized the entire the set of items in both CAT systems.

Stratification Procedures: New Developments

As shown in these simulations, the issue of exposure control was not resolved with the stratification procedures used. Leung, Chang, and Hau (2002) attributed findings such as those found in this study to a small ratio of item pool size to test length. In this dissertation, the ratios were approximately nine and 4 for the 175- and 85-item pool, respectively. Given this persistent dilemma, Leung, Chang, and Hau incorporated Sympson-Hetter methodologies into the *a*-stratification design and proposed the enhanced stratified to help overcome the effects of small item pool size to test length ratios. This procedure involves setting exposure control parameters through the Sympson-Hetter procedure with a stratified item pool. Following this development, Pastor, Dodd, and Chang (2002) proposed the conditional enhanced a-stratified design to further control exposure control by conditioning it on ability.

These stratification procedures are more advanced than the ones investigated in this paper, therefore the specific mechanisms of these procedures will not be discussed in this chapter. However, it is important to mention them given that they were designed to control item exposure better than what was found in this study. Using dichotomous items, Leung, Chang, and Hau found that the enhanced stratified procedure produced lower item overlap rates and lower rates of un-administered items than that found with the *a*-stratification procedure. In other words, the enhanced stratified method provided a stronger approach at exposure control than the original *a*-stratification procedure. Pastor, Dodd, and Chang (2002) found support for the conditional enhanced stratified procedure in terms of item exposure, but not in terms of the precision of ability estimation. Given the weak performance of the *a*-stratification and *a*-stratification with *b*-blocking procedures on polytomous items in this dissertation, the advanced stratification methods may prove beneficial to investigate more fully in the polytomous context.

However, it should be noted that adapting to these more advanced models of stratification inhibits the original simplicity of the stratification procedures as methods of exposure control. When these methods were first proposed, the idea was to be able to utilize exposure control procedures without the intensive calculations involved with the conditional procedures (e.g., Sympson-Hetter) and without leaving item selection up to chance (e.g., randomization). However, with the more advanced models of stratification, the simplicity is somewhat removed since they involve the methodologies of the Sympson-Hetter, calculating exposure control parameters by using a priori CAT simulations. If this type of procedure is more fruitful than the original stratification procedures, then, perhaps, it is more beneficial to just utilize the Sympson-Hetter without stratification. This is the main reason that the advanced methods of stratification were not investigated in this study.

Limitations

One limitation of this study involves the characteristics of the items used for the CAT simulations. As previously mentioned, the assessment from which the items came from is a low-stakes assessment. Therefore, since the motivation to perform well is low then the subsequent item parameters might not be ideal, which was the case for this study. An attempt to rectify this problem involved adjusting the item discrimination parameters. This adjustment might not have reflected reality. As previously discussed, when stratifying items by discrimination it is necessary that the overlap of item discrimination across strata be minimal. This is to ensure that each stratum contains a unique set of information, leading to an optimal selection of items to administer. In other words, items chosen from higher strata should not reflect the same amount of information as items chosen from previous strata.

Tables 13, 14, 17 and 18 reveal that the items used in this study did not always allow for minimal overlap in item discrimination among the strata. This could allow the test algorithm to choose items not appropriate from certain strata since they could be contributing the same information characteristics as items that have already been administered. Since there is not a specific standard of allowable overlap in item discrimination, making judgments about what is acceptable, given a stratified item pool, can sometimes be difficult.

Another limitation of this study related to the characteristics of the items is the fact that some items exhibited a reversal of step difficulties. In other words, the second step-difficulty value was sometimes lower than the first step-difficulty. Although this is not a violation of the item response theory model used in this study, it does present a challenge to the optimal characteristics of items. As pointed out by Andrich (1988), when the step difficulty values are in reversed order, there is not a region of ability level relative to the item's difficulty for which a certain score category is most likely. Perhaps easier to understand, when the step difficulty values are in appropriate order, from smallest to largest, then there will be a region of ability level relative to item difficulty for which each score category is most likely.

The manner in which the polytomous items were classified by item difficulty presented another challenge within this study. As discussed before, there is no previous research that lends support in classifying polytomous by difficulty when the items have more than one difficulty parameter. Therefore, the procedure of classification used in this study was designed without any reference to previous research literature. This suggests that there may be a more optimal way (e.g., using the width of the items' information functions) that these items could have been classified according to difficulty.

Some might argue that using an average value of item difficulties would, in fact, help solve the issue of classifying items with more than one difficulty parameter. Although, this could be done, it does present a major disadvantage. Using an average value of difficulty for items with more than one difficulty parameter would be analogous to adapting to a different item response theory model altogether. In the case of this study, that would mean going from the generalized partial credit model to Andrich's rating scale model, which assumes only one index of difficulty per item. Therefore, even though averaging difficulty parameters within an item can work, it does change the assumptions of the model originally being used.

Directions for Future Research

The issue of an optimum number of strata to use when utilizing mechanisms of stratification on polytomous item pools is far from being resolved, despite what has been presented in this dissertation. There is still research to be done to continue refining the stratification procedures and finding support for their use in polytomous CATs. Two key aspects of future research investigating the stratification procedures with polytomous items are discussed next.

First, this study investigated the stratification procedures using their original methodologies, not incorporating the more advanced techniques developed years later. In other words, future research should be directed at analyzing the enhanced stratified and conditional enhanced stratified procedures to continue the search of an optimum number of strata. Since the enhanced stratified method was designed to overcome a small item pool size to test length, this procedure may be able to handle the small sizes typical of polytomous item pools.

Second, this study used fixed-length CATs to investigate the stratification procedures. Fixed-length CATs have the advantage of administering the same number of items to everyone regardless of ability. However, it can also result in different measurement errors of ability estimation across examinees. Variable length CATs have the advantage of producing approximately the same measurement error for all examinees despite the different numbers of items each examinee is administered. From this, Wen, Chang, and Hau (2000) adopted the stratification procedures into variable length CATs and found that the *a*-stratification procedure, as utilized in this dissertation, performed well in terms CAT accuracy and item pool usage, using dichotomous items. From this, future research should investigate the stratification procedures – varying the number of strata – with variable length CATs.

APPENDIX A: CONDITIONAL BIAS

Figure A1: Conditional Bias for No-Exposure Control with the 175-Item Pool

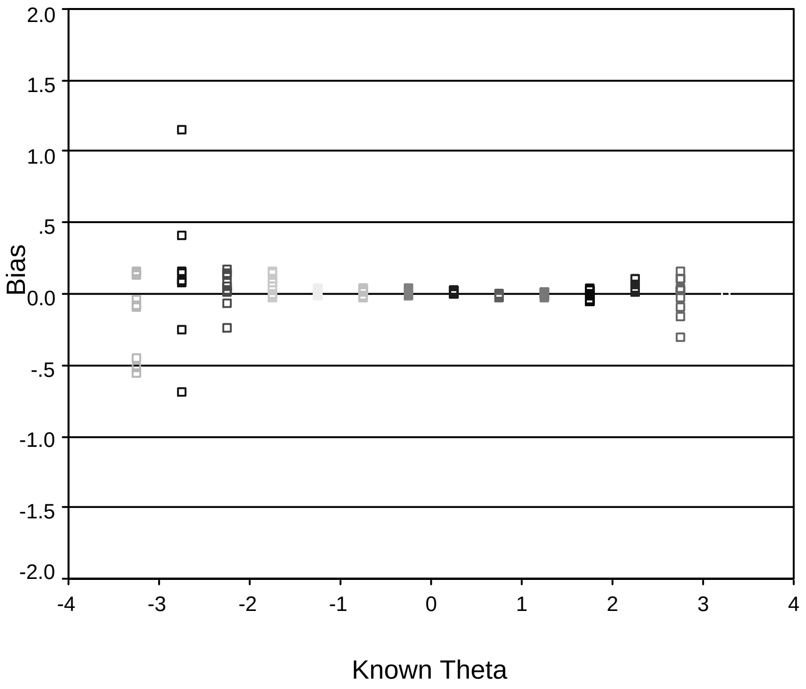


Figure A2: Conditional Bias for Randomesque-6 with the 175-Item Pool

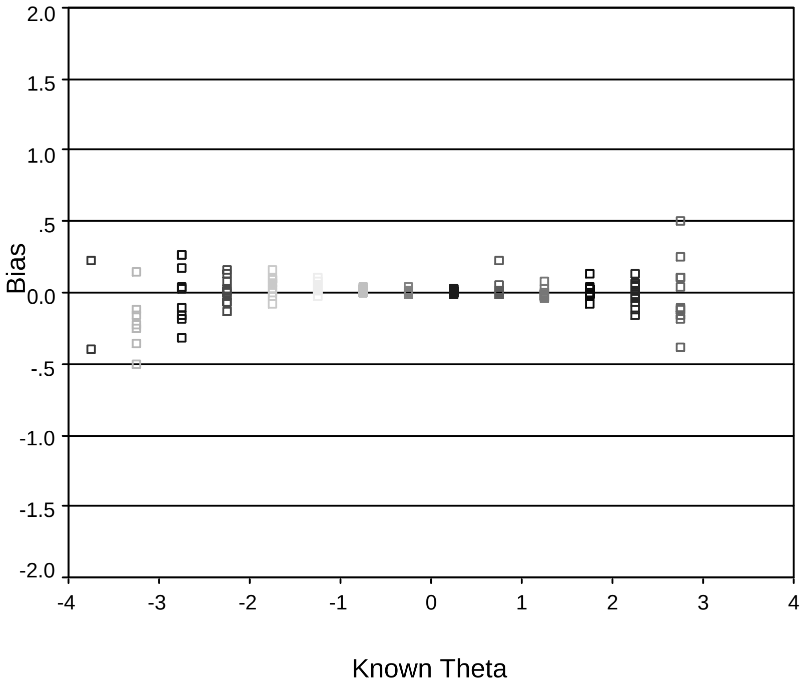


Figure A3: Conditional Bias for AS-2 with the 175-Item Pool

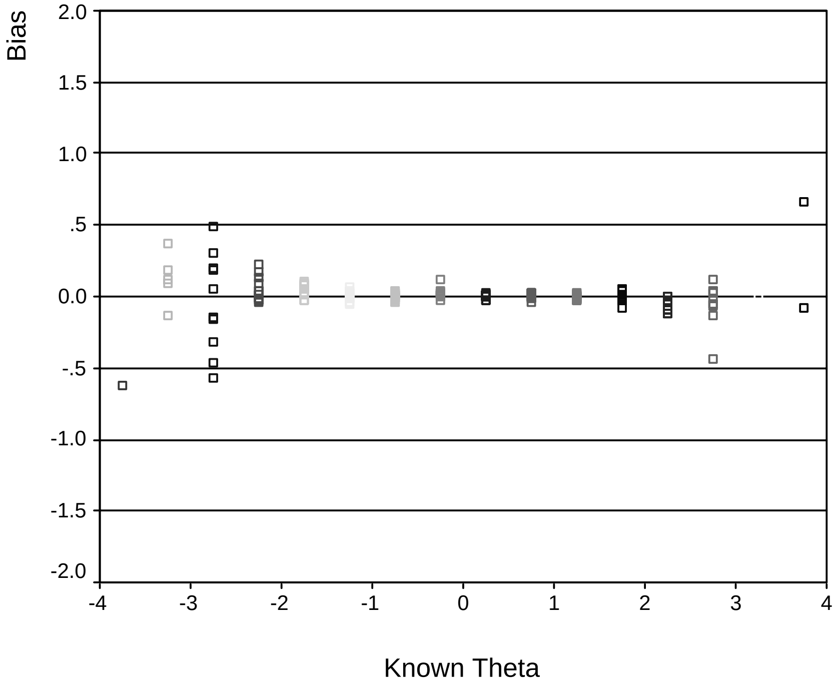


Figure A4: Conditional Bias for AS-3 with the 175-Item Pool

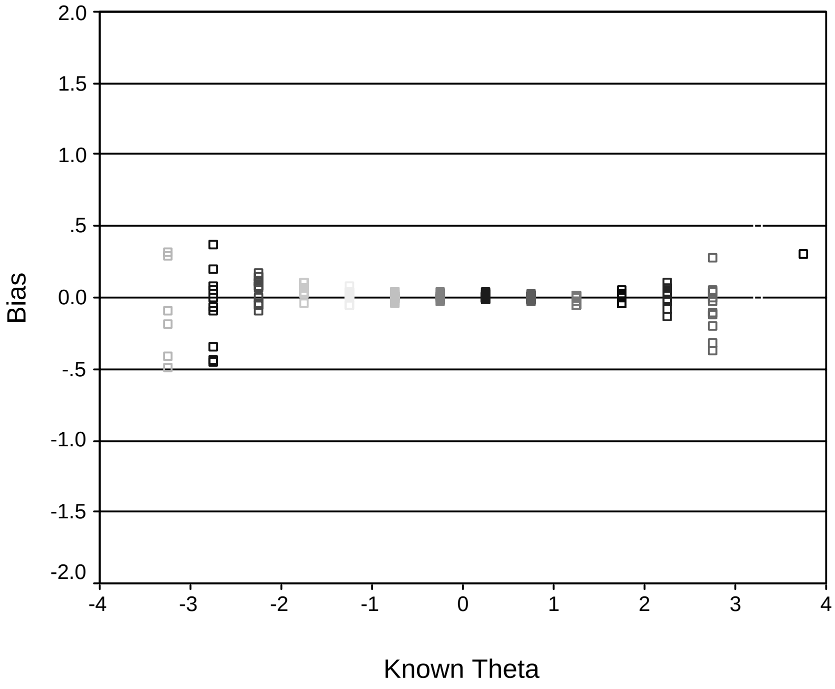


Figure A5: Conditional Bias for AS-4 with the 175-Item Pool

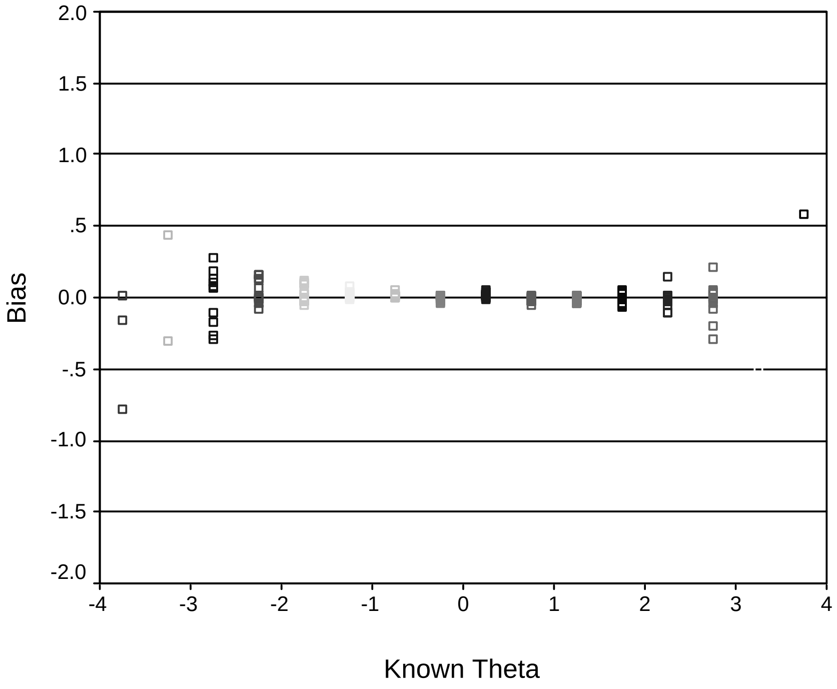


Figure A6: Conditional Bias for AS-5 with the 175-Item Pool

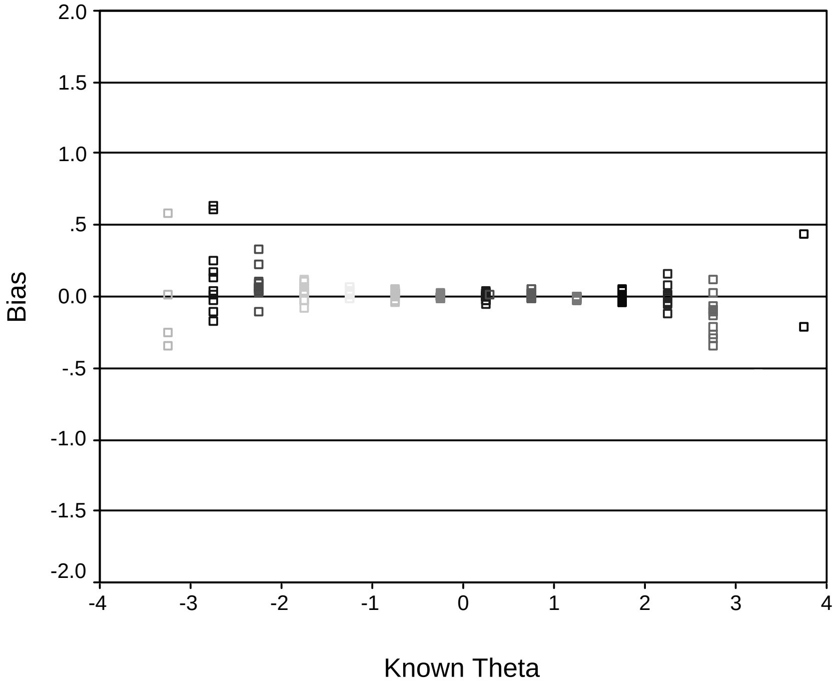


Figure A7: Conditional Bias for BAS-2 with the 175-Item Pool

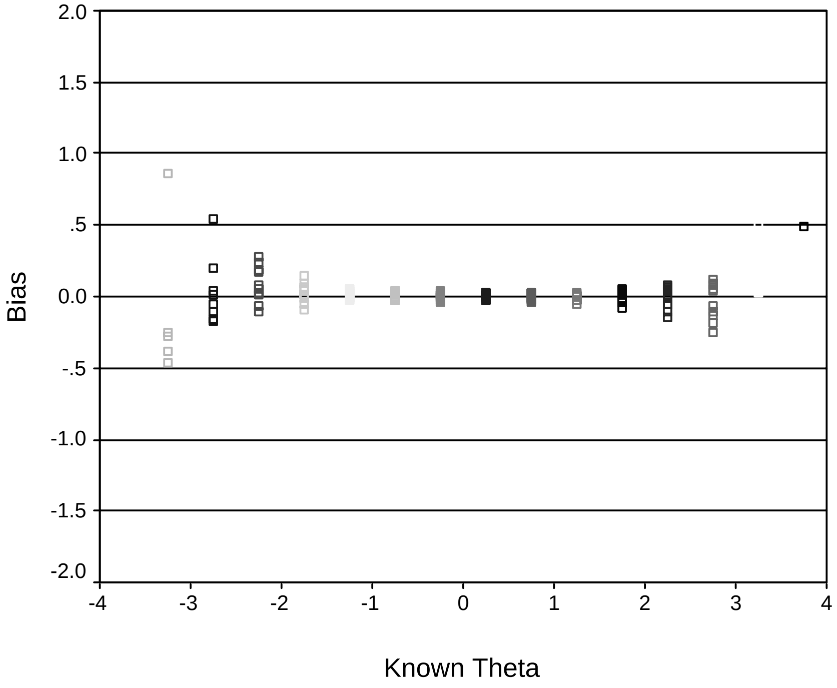


Figure A8: Conditional Bias for BAS-3 with the 175-Item Pool

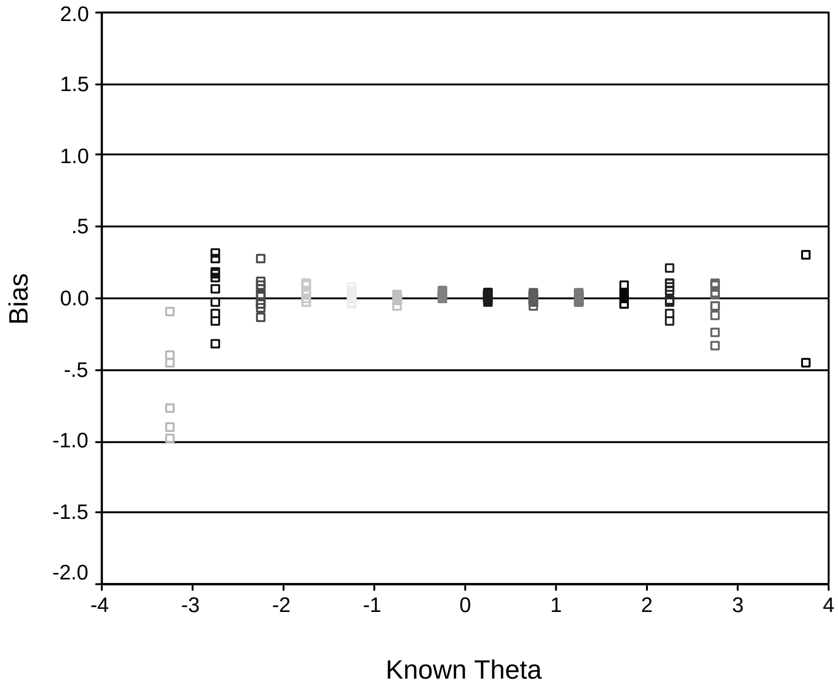


Figure A9: Conditional Bias for BAS-4 with the 175-Item Pool

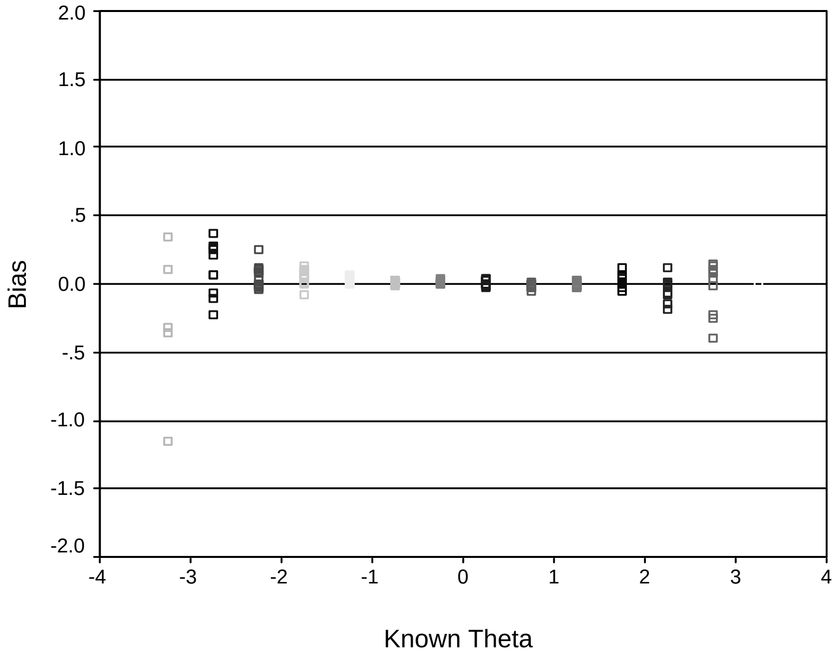


Figure A10: Conditional Bias for BAS-5 with the 175-Item Pool

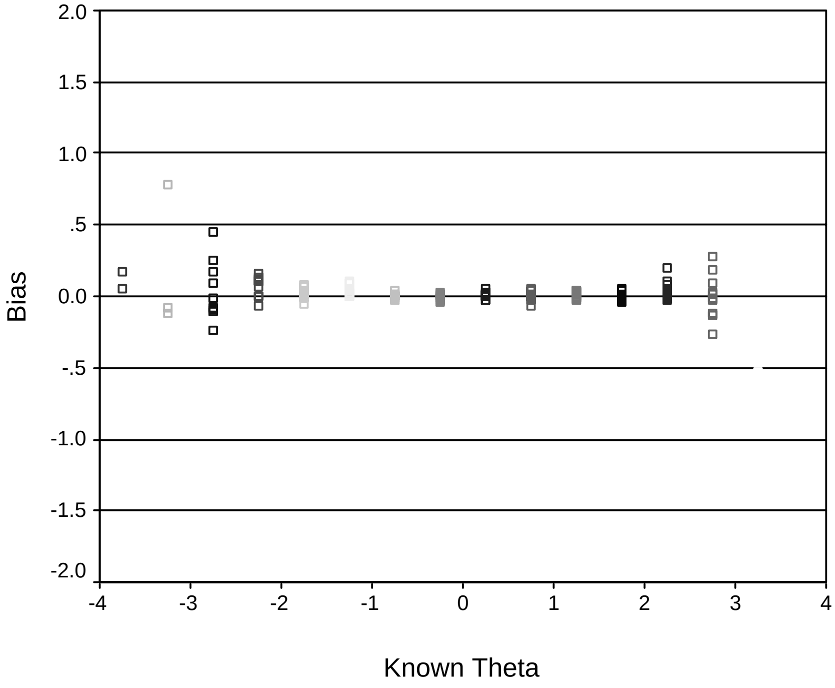


Figure A11: Conditional Bias for No-Exposure Control with the 85-Item Pool

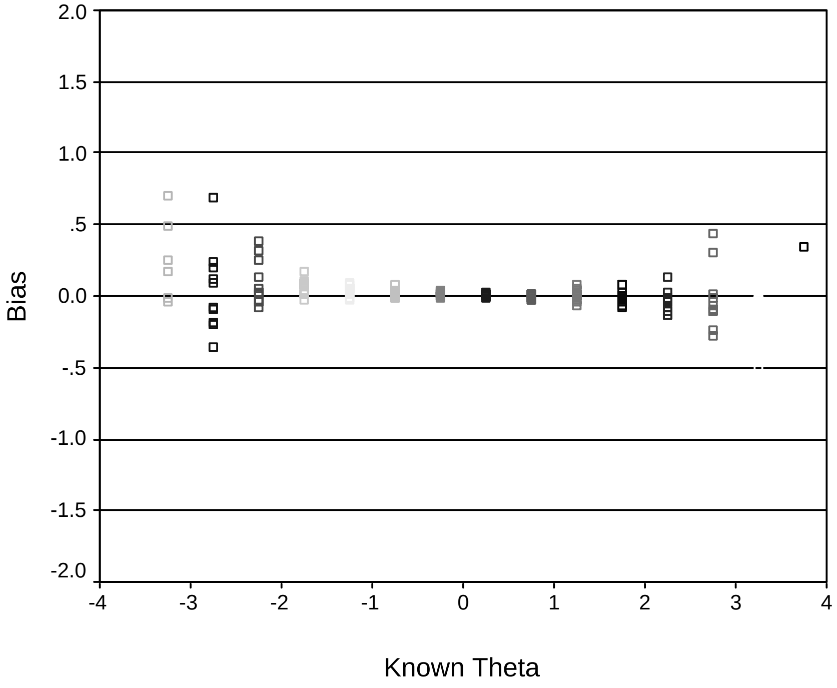


Figure A12: Conditional Bias for Randomesque-6 with the 85-Item Pool

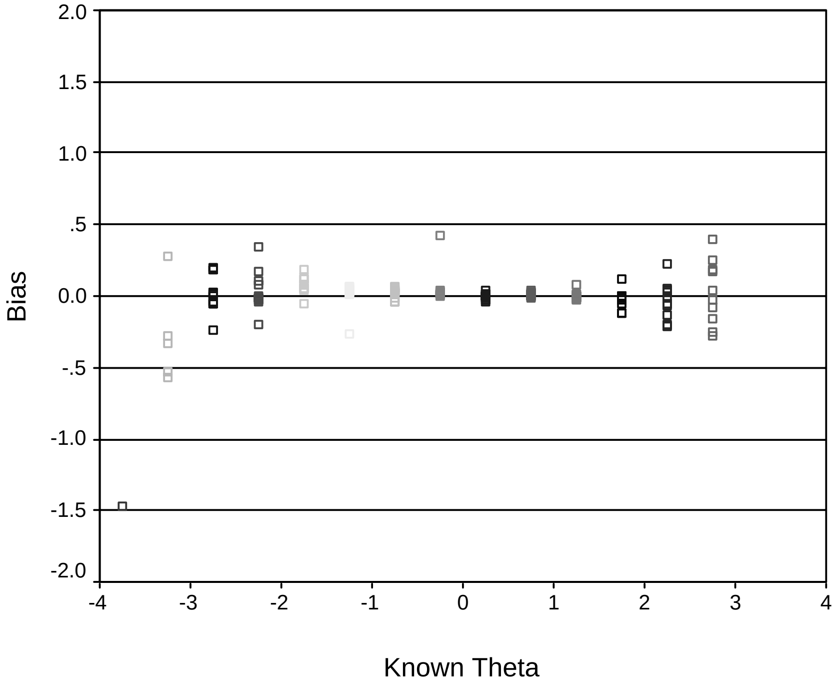


Figure A13: Conditional Bias for AS-2 with the 85-Item Pool

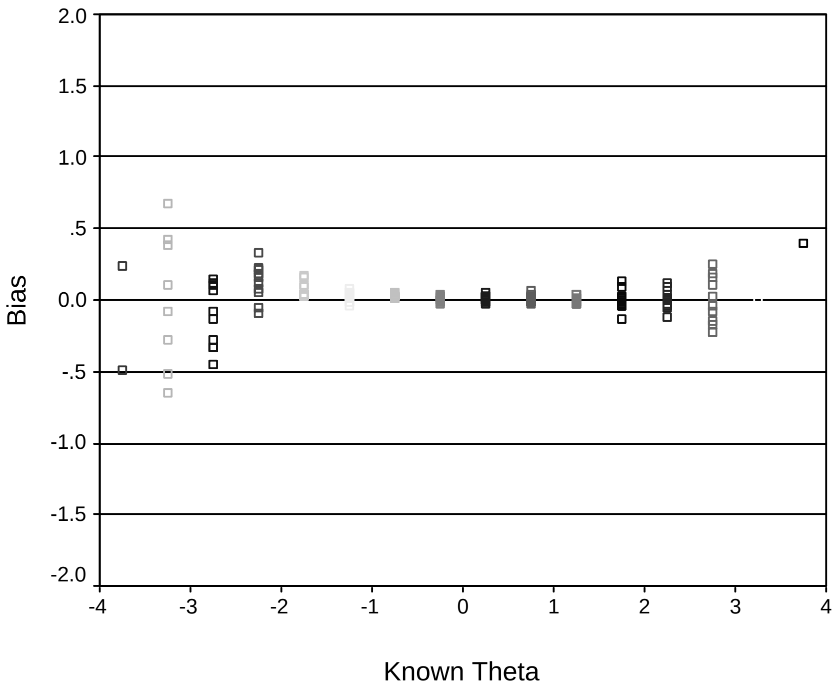


Figure A14: Conditional Bias for AS-3 with the 85-Item Pool

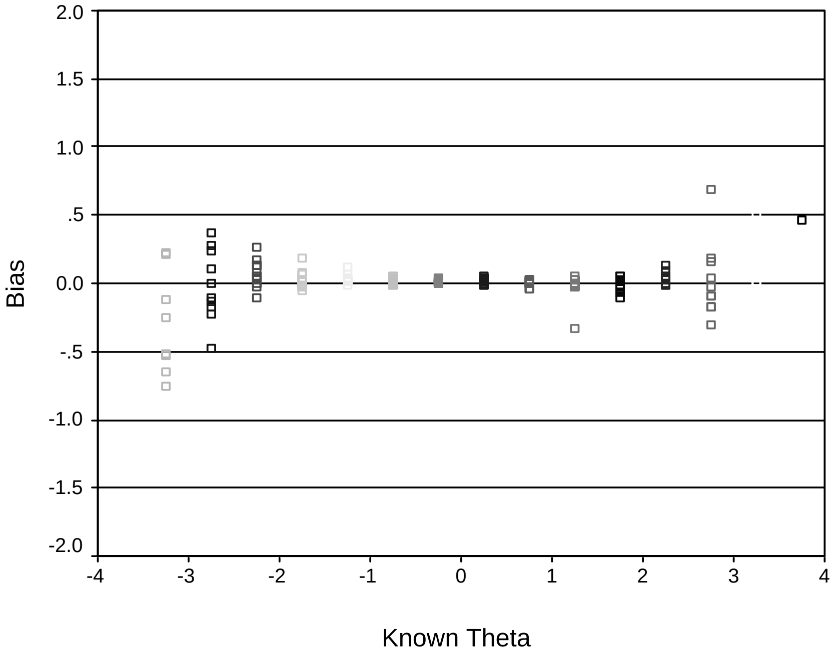


Figure A15: Conditional Bias for BAS-2 with the 85-Item Pool

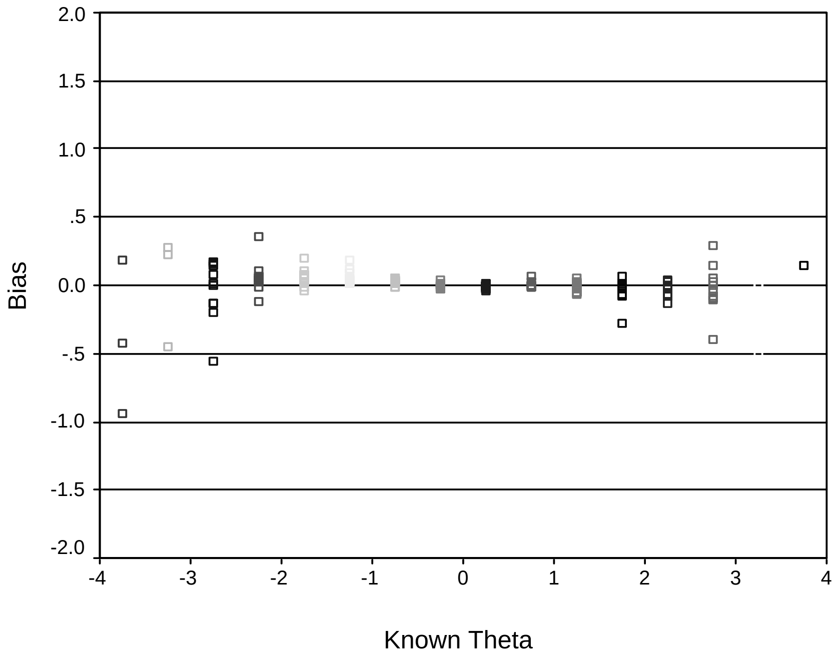
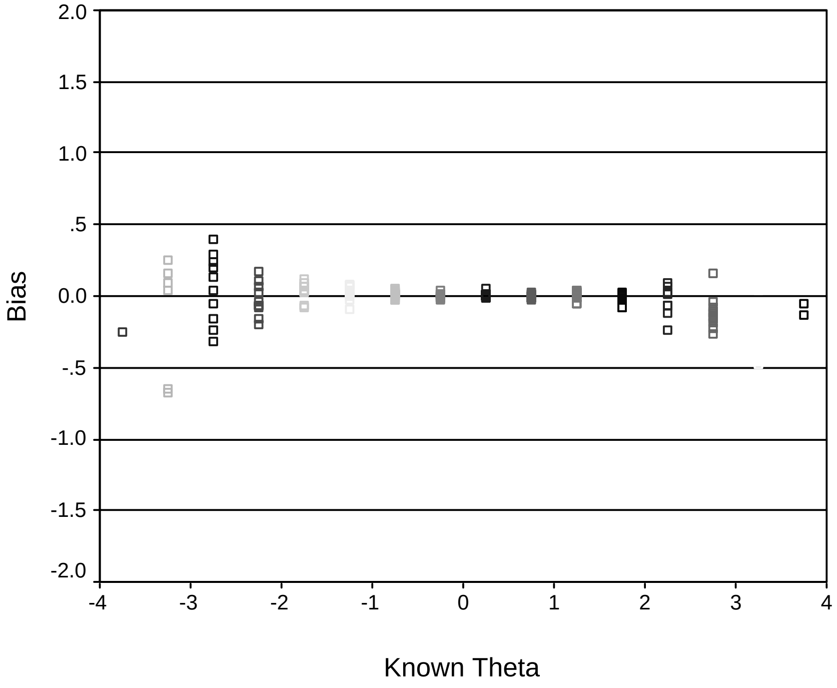


Figure A16: Conditional Bias for BAS-3 with the 85-Item Pool



REFERENCES

Allen, N.L., Carlson, J.E., & Zelenak, C.A. (1999). The NAEP 1996 technical report. Washington, DC: National Center for Educational Statistics.

Andersen, E.B. (1972). The numerical solution of a set of conditional estimation equations. The Journal of the Royal Statistical Society, Series B, 34, 42-54.

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43(4), 561-573.

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

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

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika, 46(4), 443-459.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6(4), 431-444.

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

Burt, W., Kim, S., Davis, L.L., & Dodd, B.G. (2003). A comparison of item exposure control procedures using a CAT system based on the generalized partial credit model. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Chang, H.H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. Psychometrika, 58(1), 37-52.

Chang. H.H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. Applied Psychological Measurement, 20(3), 213-229.

Chang, H.H., & Ying, Z. (1999). *a*-Stratified multistage computerized adaptive testing. Applied Psychological Measurement, 23(3), 211-222.

Chang, H.H., Qian, J. & Ying, Z. (2001). *a*-Stratified multistage computerized adaptive testing with *b*-blocking. Applied Psychological Measurement, 25(4), 333-341.

Chen, S., Hou, L., & Dodd, B.G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. Educational and Psychological Measurement, 58, 569-595.

Chen, S., & Ankenmann, R.D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. Journal of Educational Measurement, 41(2), 149-174.

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

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

Eignor, D.R., Stocking, M.L., Way, W.D., & Steffen, M. (1993, April). Case studies in computer adaptive test design through simulation. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

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. (1983). The promise of tailored tests. In H. Wainer and S. Messick, Principles of modern psychological measurement. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

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

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

Hau, K.T., Wen, J.B., & Chang, H.H. (2002, April). Optimum number of strata in the a-stratified computerized adaptive testing design. Paper presented at the annual meeting of the American Educational Research Association (New Orleans).

Hau, K.T., & Chang, H.H. (2001). Item selection in computerizes adaptive testing: should more discriminating items be used first? Journal of Educational Measurement, 38(3), 249-266.

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

Leung, C.K., Chang, H.H., & Hau, K.T. (2001, April). An examination of item selection rules by stratified CAT designs integrated with content balancing methods. Paper presented at the annual meeting of the American Educational Research Association, Seattle.

Leung, C.K., Chang, H.H., & Hau, K.T. (2002). Item selection in computerized adaptive testing: improving the *a*-stratified design with the Sympson-Hetter algorithm. Applied Psychological Measurement, 26(4), 376-392.

Leung, C.K., Chang, H.H., & Hau, K.T. (2003). Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. Educational and Psychological Measurement, 63(2), 257-270.

Lord, F.M., & Wingersky, M.S. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. Applied Psychological Measurement, 8, 347-364.

Lord, F.M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 28, 989-1020.

Lord, F.M. (1975). The 'ability' scale in item characteristic curve theory. Psychometrika, 40(2), 205-217.

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

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

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

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

McBride, J.R., & Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D.J. Weiss (Ed.), New Horizons in testing (pp. 223-236). New York: Academic Press.

Muraki, E. (1990). Fitting a polytomous item response theory model to Likert-type data. Applied Psychological Measurement, 14(1), 59-71.

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

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

Parshall, C., Hogarty, K., & Kromrey, J. (1999, June). Item exposure in adaptive tests: An empirical investigation of control strategies. Paper presented at the annual meeting of the Psychometric Society, Lawrence, KS.

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

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: The Danish Institute for Educational Research.

Rost, J. (1988). Measuring attitudes with a threshold model drawing on a traditional scaling concept. Applied Psychological Measurement, 12(4), 397-409.

Revuelta, J. (1995). El control de la exposicion de los items en tests adaptivosa informatizados (Item exposure control in computerized adaptive tests). Unpublished master dissertation. Universidad Autonoma de Madrid, Spain.

Revuelta, J., & Ponsoda, V. (1996). Metodos sencillos para el control de las tasas de exposicion en tests adaptativos informatizados (Simple methods for item exposure control in CATs). Psicologica, 17, 161-172.

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

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. Psychometrika Monograph, No. 17.

Stahl, J.A., & Lunz, M.E. (1993, April). Assessing the extent of overlap of items among computerized adaptive tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Stocking, M.L. (1994). Three practical issues for modern adaptive testing. (Research Report 94-5). Princeton, NJ: Educational Testing Service.

Stocking, M.L. (1998). A framework for comparing adaptive test designs. Unpublished manuscript.

Sympson, J.B., & Hetter, R.D. (1985, October). Controlling item exposure rates in computerized adaptive testing. Paper presented at the annual meeting of the Military Testing Association, Navy Personnel Research and Development Center, San Diego.

Thissen, R.J., & Mislevy, D. (2000). Testing Algorithms. In H. Wainer (Ed.) Computerized adaptive testing: A primer (2nd ed., pp. 101-134). Mahwah, NJ: Lawrence Erlbaum.

Urry, V.W. (1977). Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 14(2), 181-196.

Wainer, H., & Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In H. Wainer (Ed.), Computerized adaptive testing: A primer (2nd ed., pp. 271-299). Mahwah, NJ: Lawrence Erlbaum.

Way, W. (1994, April). Psychometric results of the NCLEX™ beta test. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Way. W., Zara, A., & Leahy, J. (1996, April). Modifying the NCLEX™ CAT item selection algorithm to improve item exposure. Paper presented at the annual meeting of the American Educational Research Association, New York.

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

Weiss, D.J. (1973). The stratified adaptive computerized ability test. (Research Report 73-3). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D.J. (1974). Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D.J. & McBride, J.R. (1984). Bias and information of Bayesian adaptive testing. Applied Psychological Measurement, 8(3), 273-285.

Wen, J.B., Chang, H.H., Hau, K.T. (2000). *Adaptation of a-stratified method in variable length computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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

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

VITA

Marc Anthony Johnson was born in Austin, Texas on March 11, 1978, the son of James Louis and Lauren Jo Johnson. After completing his work at Lyndon Baines Johnson High School, Austin, Texas, in 1996, he entered The University of Texas at Austin. He received the degree of Bachelor of Arts in Psychology, with a minor in Educational Psychology from The University of Texas at Austin in December 2001. In August 2002 he entered the Graduate School of The University of Texas at Austin.

Permanent Address: 5304 Robinsdale Lane
Austin, Texas 78723

This dissertation was typed by the author.