The Dissertation Committee for Laurie Laughlin Davis
Certifies that this is the approved version of the following dissertation:

Strategies for Controlling Item Exposure in Computerized Adaptive Testing with Polytomously Scored Items

Committee:

Barbara G. Dodd, Supervisor

William R. Koch

Steven J. Fitzpatrick

Susan N. Beretvas

Hua-Hua Chang

Thomas E. Brooks

Strategies for Controlling Item Exposure in Computerized Adaptive Testing with Polytomously Scored Items

by

Laurie Laughlin Davis, 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

May 2002

Dedicated to my grandparents

Harry and Doris Tobin

and

William and Lorraine Laughlin

Acknowledgements

I would first like to thank my advisor Barbara Dodd, not only for her invaluable support and feedback in preparing this dissertation, but also for her encouragement, enthusiasm, and friendship throughout the entire graduate school process. Barbara was responsible for recruiting me into the quantitative methods program and her belief in my abilities has carried me through even the toughest times.

I would also like to acknowledge the other members of my dissertation committee. Thanks to Tasha Beretvas for helping me to keep things manageable and for being so involved despite all her competing obligations. Thanks to Hua-Hua Chang for his support and consultation when things did not go as planned. Thanks to Tom Brooks for his valuable questions and comments. Thanks to Steve Fitzpatrick for his awe-inspiring insights and technical support on this and numerous other projects over the past five years. And thanks to Bill Koch for asking the interesting questions and for all of his wonderful classes which prepared me to answer them.

I must also thank my support system of friends and colleagues. First, I owe thanks to my cousin, Dan Laughlin, for giving me the best advice about making it through the dissertation process in one piece. Next, I must thank Carol George for her friendship, advice, and motivational pep talks, without which, I would still be on page two. I would like to thank Lori Richardson for her continued friendship and support throughout this process, even in times of self-doubt and sloppy sentence structure. Lastly, I must thank Dena Pastor for having the impeccable timing to share this life altering experience with me. Dena understands better than anyone the highs and lows, the euphorias and the anxieties, and the blood, sweat, and tears poured into this degree because she has felt them right alongside me. Dena has become not only an indispensable friend, but someone for whom I have the utmost respect and admiration. I will continue to call her a friend long after dissertation becomes just another word in the dictionary.

Finally, I would like to thank my parents, Donald and Roseann Laughlin, for raising me to aspire to achieve and for giving me the conscientious spirit which has allowed me to succeed in all that I have tried. And last, but in no way least, I thank my husband, Dustin, for all his love and support, for his interest and patience, for his hard work and sacrifice, and for bringing me chocolate and a warm cat when I really needed it.

Strategies for Controlling Item Exposure in Computerized Adaptive Testing with Polytomously Scored Items

Publication No._____

Laurie Laughlin Davis, Ph.D.

The University of Texas at Austin, 2002

Supervisor: Barbara G. Dodd

Choosing a strategy for controlling the exposure of items to examinees has become an integral part of test development for computerized adaptive testing (CAT). Item exposure can be controlled through the use of a variety of algorithms which modify the CAT item selection process. This may be done through a randomization, conditional selection, or stratification approach. The effectiveness of each procedure as well as the degree to which measurement precision is sacrificed has been extensively studied with dichotomously scored item pools. However, only recently have researchers begun to examine these procedures in polytomously scored item pools.

The current study investigated the performance of six different exposure control mechanisms under three polytomous IRT models in terms of measurement precision, test security, and ease of implementation. The three models examined in

the current study were the partial credit, generalized partial credit, and graded response models. In addition to a no exposure control baseline condition, the randomesque, within .10 logits, Sympson-Hetter, conditional Sympson-Hetter, a-Stratified, and enhanced a-Stratified procedures were implemented to control item exposure rates. The a-Stratified and enhanced a-Stratified procedures were not evaluated with the partial credit model. Two variations of the randomesque and within .10 logits procedures were also examined which varied the size of the item group from which the next item to be administered was randomly selected.

The results of this study were remarkably similar for all three models and indicated that the randomesque and within .10 logits procedures, when implemented with the six item group variation, provide the best option for controlling exposure rates when impact to measurement precision and ease of implementation are considered. The three item group variations of the procedures were, however, ineffective in controlling exposure, overlap, and pool utilization rates to desired levels. The Sympson-Hetter and conditional Sympson-Hetter procedures were difficult and time consuming to implement, and while they did control exposure rates to the target level, their performance in terms of item overlap (for the Sympson-Hetter) and pool utilization were disappointing. The a-Stratified and enhanced a-Stratified procedures both turned in surprisingly poor performances across all variables.

Table of Contents

List of Figures

List of Tables

Chapter I

Introduction

Wainer (1990) lists among the advantages of computerized adaptive tests (CATs) over paper and pencil tests that CATs provide improved test security. This is due to both the increased physical security of the computerized item pool and to the individualized nature of a CAT which makes it more difficult to "artificially boost one's score by merely learning a few items." Since each examinee takes a test individually tailored to his or her trait level, it was thought that it would be necessary to learn a large portion of the item pool in order for this preknowledge to have any impact on an examinee's score. While not entirely incorrect, a rather high profile case has shown this logic to be incomplete by not taking into account the continuous nature of CAT administrations, the manner in which items are selected for administration in a CAT, and the size of the item pool.

In the Fall of 1993, after five years of development, Educational Testing Service (ETS) launched the computerized adaptive version of the Graduate Record Examinations (GRE) (Mills, 1999). In December of 1994, Kaplan Educational Centers, one of the country's largest test preparation companies, instructed its employees to take the GRE CAT, to remember as many items as possible, and to report those items back to Kaplan. Subsequently, other employees would review the list of now compromised items (so as not to duplicate efforts) and repeat the process. Within a few weeks, Kaplan employees discovered that most of the items they were

administered already appeared on the list of compromised items. Kaplan had successfully compromised a large enough portion of the item pool that ETS, upon notification of the incident by Kaplan, temporarily shut down testing while new items were developed (Mills, 1999; Wainer & Eignor, 2000).

One reason that Kaplan was able to compromise the GRE CAT item pool in this manner has to do with the continuous nature of computerized adaptive testing administrations. When admissions tests were only administered using paper and pencil methods, students would gather by the hundreds in university auditoriums a limited number of times a year to take the examination. Because tests were administered on the same day to a large group of individuals and the number of items in common with the next form of the test was relatively small, test administrators and developers did not have to be so concerned that examinees would pass information about the test to one another that could affect a student's performance at a later time (i.e. artificially raise their score due to advanced knowledge of the test material). The primary concerns in terms of test security were physical—the security of the testing facility and the security of the test materials before, during, and after the examination. The largest security threat was that someone would obtain and distribute an advanced copy of the test, but with appropriate precautions this threat could be minimized. However, a more surreptitious security concern did exist in terms of students using the differences in administration times and time zones to obtain information about the test content. A student who took the test on the east coast, could inform students on

the west coast of which items were on the test, before the west coast administrations began.

With high stakes tests being moved increasingly to computerized format, exam administration procedures have changed dramatically causing many new security concerns to arise and old concerns to be exacerbated. Tests that had been printed in booklets and required little more than a chair and desk of some sort to administer, now required expensive computer equipment. Due to the relative expense of computerized testing over paper and pencil, it was quickly determined that testing en masse in large auditoriums would no longer suffice. Because of the need for appropriate computer equipment, test centers that could seat many fewer examinees (five to ten at a time originally), supplanted the traditional auditoriums (that could handle hundreds of students at the same time). To compensate for the fact that test volumes per administration would be so much lower, test centers would have to offer the test on a near continuous basis—multiple administrations per day throughout the week. Changing the venue and, consequently, frequency of test administration, has resulted in a much larger potential for examinees to share test related material with one another. Now, a student taking a test on Friday, can obtain information from a friend who took the test on the previous Wednesday.

While continuous testing has contributed to an increased threat to test security in CAT, the primary threat stems from the CAT algorithm itself. In CAT, items may be selected based on the relative information provided at a given trait level. We learn little about an examinee's trait level from items which are too hard or too easy. On

the other hand, we learn the most about an examinee's trait level when items are selected to be close to the estimated trait level of the examinee and maximize the item information at that estimated trait level (Wainer, 1990). Maximum information item selection may, therefore, be used to optimize measurement precision. While this makes the CAT quite efficient in terms of trait level estimation, it also produces variability in the frequency with which items are used. Because each item has its own set of characteristics definable by the parameters of a given model, items will differ in terms of the desirability of these characteristics for measuring an examinee's trait level. Items which are of average difficulty will be more often selected for administration because of the assumed normal distribution of examinee trait level. Additionally, items which are more informative will be selected more often because of the use of maximum information item selection. The converse of this situation would be one in which items are selected for administration randomly, thus producing uniform item usage, but poor measurement efficiency. Uneven item usage presents a problem for test security, because more popular items (those with the most desirable characteristics for measuring trait level) will be administered quite frequently, whereas some items may be administered less frequently or never be administered at all. This results in a differentiation between the available item pool (all items available for administration) and the functional item pool (those items which will most often be administered on examinee tests). Rather than having to memorize the available item pool (a daunting task), examinees may only have to have

preknowledge of the functional item pool (which is often much smaller) in order to artificially raise their scores.

One often proposed solution to this dilemma has been to create larger available item pools for CATs thus making the functional item pools larger as well. A related option would be to create multiple pools which can be frequently rotated in and out of use such that even if examinees had preknowledge of the functional item pool, the lifespan of its usefulness would be short. However, either of these solutions would require extensive item development which is both costly and time consuming. Test items are generally written by experienced content area experts (often teachers and retired educators) who must receive additional training in order to write items. It can be difficult to find and recruit item writers and is often costly since they must be paid for their expertise. Once items are written, they must then undergo a rigorous review process from committees looking for biased or poorly performing items. Any items which do not meet the statistical and bias requirements are rewritten or discarded, causing additional item development costs.

Regardless of the time and cost required for item development, simply enlarging the pool of possible items fails to address the real problem of differential item usage. Wainer & Eignor (2000) discuss this problem in terms of the relationship between the frequency of item usage and the rank order of item usage (Zipf, 1949; Wainer, 2000; Wainer & Eignor, 2000). In the case of item selection, an item with a low rank order (close to 1) is administered more frequently than an item with a high rank order. They demonstrated that typically, CAT item selection produces a

scenario whereby not only are the pool's most informative items most often administered, but also that a very small percentage of the pool's items account for a very large percentage of the administered items. In other words, the most popular items are popular by a large margin, resulting in an exponential decline in item usage as rank increases. Enlarging the available item pool will not alter this relationship.

In practical terms, this explains why Kaplan was able to succeed in compromising test security over such a short period of time. They did not need to capture the entire pool of items, only the functional pool. Wainer and Eignor (2000) report on the severity of this relationship in the GRE CAT and an experimental version of the SAT CAT. They found that as few as 12% of the available item pool can account for as much as 50% of the functional item pool (those items actually administered) and that as few as 33% of the available item pool can account for as much as 75% of the functional item pool.

Many algorithms exist which seek to control item exposure through constraining the administration of more popular items. These algorithms differ in terms of their complexity and the variables taken into consideration. Parshall, Davey, and Nering (1998) discuss the three often conflicting goals of item selection in CAT. First, item selection must maximize measurement precision, by selecting the item which maximizes information or posterior precision for the examinee's current trait level. Second, item selection must seek to protect the security of the item pool by limiting the degree to which items may be exposed. Third, item selection, must ensure that examinees will receive a content balanced test. Stocking and Swanson

(1998) add a fourth goal to this list, stating that item selection must also maximize item usage so that all items in a pool are used, thereby ensuring good economy of item development. Stocking and Lewis (2000) equate the item selection problem to an inflated balloon—pushing against one side, may address one issue, but will ultimately cause another problem to appear as a bulge on another side of the balloon.

Different approaches to the goals of item selection will produce different testing algorithms (Stocking & Lewis, 2000). Attempts to address the third goal are denoted exposure control methodologies (Parshall, Davey, & Nering, 1998). Way (1998) discusses two types of exposure control strategies—randomization and conditional selection. Randomization strategies randomly choose the next item for administration from a set of nearly optimal items rather than selecting the single most informative item. The conditional selection strategies are those in which the probability of an item being administered is controlled conditional on a given criterion, such as the expected frequency of item usage. Finally, a third approach to exposure control has recently been proposed by Chang and Ying (1996) in which items in the pool are stratified according to their statistical properties (item parameters) and items are constrained to be administered from certain strata.

While the research investigating the extent that measurement precision is affected when using exposure control procedures with dichotomous (right/wrong) scoring is extensive, only recently have researchers begun to address the effects of exposure control when using polytomous (partial credit) scoring. Polytomously scored items have properties unique and separate from dichotomously scored items

and must be separately studied under conditions of constrained item selection. Therefore, the purpose of the present dissertation was to examine the utility of several exposure control methods with three commonly used polytomous IRT models. Each procedure was evaluated in terms of its ability to successfully control item exposure and test overlap rates and make optimal use of the available item pool while minimizing impact on measurement precision. In addition, the ease or difficulty of implementation of each procedure was evaluated with respect to the gains made in each of these variables.

Chapter II

Literature Review

This literature review is divided into four main topics of discussion, culminating in the statement of problem. The first section discusses the assumptions and properties of item response theory and presents several IRT models appropriate for use with both dichotomous and polytomous scoring. Special attention is given to the polytomous IRT models as they are used in the current study. The second section provides an introduction to computerized adaptive testing with dichotomous items, discussing the advantages of CAT over paper and pencil testing and the components of a CAT system, including the item pool, item selection procedure, trait estimation method, stopping rule, content balancing, and exposure control procedures. The third section provides an in-depth discussion of the three broad approaches for controlling item exposure in CAT along with a detailed presentation of specific methods which exemplify each of these approaches. Finally, the fourth section addresses the special issues faced when CAT is implemented with polytomous items and provides a review of the research examining exposure control methods in polytomous CAT.

Item Response Theory with Dichotomous Items

The psychometric foundation of computerized adaptive testing lies in item response theory (IRT) which provides a measurement model that focuses on the individual item as the level of analysis as an attempt to address several of the deficits of classical test theory (CTT). Wainer (1990) explains IRT as "a mathematical characterization of what happens when an individual meets an item." IRT provides a

mathematical description of the probability of getting an item correct conditional on trait level. The properties of IRT and assumptions of some common IRT models are discussed in the next two sections. Following this, several dichotomous and polytomous IRT models are discussed in detail.

Properties of IRT

IRT describes a group of probabilistic models in which a set of parameters that define an item (i.e. difficulty, discrimination, guessing) interact with an examinee's trait level (θ) to determine the probability of a correct response when the examinee attempts the item. Examinee trait level and item difficulty are expressed on the same scale. If an examinee's trait level is high relative to the difficulty of a given item, the probability of a correct response would be high. Conversely, if an examinee's trait level is low relative to the difficulty of a given item, the probability of a correct response would be also be low. When the item difficulty equals the examinee's trait level, the probability of a correct response is equal to .50 for models which do not assume guessing (Wainer, 1990). This relationship can be displayed graphically using an item characteristic curve (ICC) where trait level forms the abscissa, probability of a correct response forms the ordinate, and the item parameters define the shape of the function which relates the two.

IRT models possess several features which make them useful for CAT. Among these, is the concept of parameter invariance which states that item parameters are independent of the group of examinees on which they are calibrated

(within a linear transformation) and that trait estimates are independent of the particular subset of items which an examinee receives (Hambleton & Swaminathan, 1985). This property is especially important because, in a CAT, examinees take different sets of items of varying difficulties, but their trait levels can still be expressed on a common metric. In addition, IRT provides a measure of precision for each level of the trait (Hambleton & Swaminathan, 1985). A separate standard error of measurement is available for each trait estimate and is often used as a criterion in ending CAT administration.

Assumptions of Common IRT Models

While IRT models which can handle the measurement of multiple traits or dimensions have been developed (Reckase & McKinley, 1991; Reckase, 1985), the majority of currently used IRT models make the assumption that a single trait or ability is being measured. Various statistical procedures such as factor analysis, multidimensional scaling, and DIMTest (Stout, 1987) have been implemented to test for unidimensionality. This assumption must be carefully considered before employing a unidimensional IRT model, as such factors as speededness or multiple content areas can introduce additional dimensionality into a test.

A second assumption of IRT which follows from unidimensionality is that of local independence of item responses within a given trait level. That is, for a given trait level, item responses should be uncorrelated. Only when this criterion is met can the probability of a response string be defined as the product of the independent item probabilities. Any time that a set of items refers back to the same stimuli (such as a

reading passage), this assumption is threatened. When this assumption is violated, trait estimates may be inflated due to overestimation of item information (Wainer & Lewis, 1990).

Finally, an assumption is made when a particular IRT model is chosen that the model will fit the data. In other words, it must be assumed that the mathematical function (ICC) which represents the relationship between trait level and item response will be an accurate reflection of that relationship for the data. While model fit statistics such as the likelihood ratio chi-square have been proposed (McKinley & Mills, 1985), the capacity to check this assumption is hindered both by issues with statistical power and by the fact that trait level is a latent variable and not directly measurable (Hambleton & Swaminathan, 1985).

Dichotomous IRT models

There are three models which are primarily used to describe data scored in a binary, right/wrong fashion: the one parameter logistical (or Rasch) model (1PL; Wright, 1968; Rasch, 1960), the two parameter logistic model (2PL; Lord, 1952; Birnbaum, 1968), and the three parameter logistic model (3PL; Lord, 1952; Birnbaum, 1968).

The one parameter logistic model defines the probability of correctly responding to an item as:

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

where θ represents the examinee's trait level, and b_i represents the difficulty of item *i*. The location of b_i is the point at which the examinee has a 50 percent chance of answering the item correctly (Hambleton & Swaminathan, 1985).

The two parameter logistic model defines the probability of correctly responding to an item as:

$$P_i(\theta) = \frac{e^{ai(\theta - bi)}}{1 + e^{ai(\theta - bi)}}, \quad (2)$$

where θ represents the examinee's trait level, and b_i represents the difficulty of item *i*, and a_i represents the discrimination of item *i*. The item discrimination is proportional to the slope of the ICC at the point where $\theta = b_i$. The steeper the slope at this point, the higher the item discrimination (Hambleton & Swaminathan, 1985).

The three parameter logistic model defines the probability of correctly responding to an item as:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{ai(\theta - bi)}}{1 + e^{ai(\theta - bi)}}, \quad (3)$$

where θ represents the examinee's trait level, and b_i represents the difficulty of item *i*, a_i represents the discrimination of item *i*, and c_i represents the pseudo-guessing parameter for item *i*. The psuedo-guessing parameter marks the lower asymptote of the ICC and represents the probability that low trait level examinees would have of answering the item correctly based on factors such as guessing. It should be noted that when the pseudo-guessing parameter is greater than zero, the probability of an

examinee with trait level θ_b answering the item correctly exceeds 50 percent (Hambleton & Swaminathan, 1985).

Each model has properties and assumptions which make it more or less attractive. The Rasch model has the advantage of requiring the fewest parameters, so it is easier to work with and provides fewer estimation problems than the other models. In addition, with the Rasch model, raw score is a sufficient statistic, meaning that every person with the same raw score will receive the same trait estimate. However, the Rasch model also assumes that guessing does not influence an examinee's response to an item and that all items are equally discriminating. Many researchers (Birnbaum, 1968; Traub, 1983; Hambleton & Swaminathan, 1985) have questioned the degree to which achievement test data can meet these assumptions. On the other hand, the 3PL model, while presenting the most flexibility in terms of modeling the data, also requires the most parameters. The pseudo-guessing parameter is especially hard to estimate because of sparse data conditions at low trait levels (Hambleton & Swaminathan, 1985).

Each item calibrated with an IRT model has an item information function, *I(θ)*, which reflects the precision of measurement conditional on trait level. Information can be defined as:

$$Ii(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)}, \qquad (4)$$

where $P_i(\theta)$ equals the probability of correctly responding to item i given θ , $P'_i(\theta)$ is the first derivative of $P_i(\theta)$ with respect to theta, and $Q_i(\theta)$ is equal to $(1-P_i(\theta))$. These bell shaped functions can be used to describe, compare, and select items. Information is inversely related to the standard error of measurement according to the formula

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}; \qquad (5)$$

where $SE(\theta)$ represents the standard error of measurement at θ and $I(\theta)$ represents the value of the test information function (which is the sum of all item information functions) at θ.

The maximum value and the location of that maximum value are a function of the particular item parameters. The item information function peaks at b_i for the 1PL and 2PL models. For the 3PL model, maximum information occurs slightly to the right of b_i depending on the value of the pseudo-guessing parameter. In the 1PL model, the value of maximum information is constant, with only the location at which it occurs shifting across items. In the 2PL model, the value of maximum information is proportional to the square of the discrimination. The steeper the slope of the ICC, the greater the discrimination, and the greater the information provided by the item. In the 3PL model, the smaller the pseudo-guessing parameter, the greater information.

Item Response Theory with Polytomous Items

While, to date, less commonly used than the dichotomous models, polytomous IRT models allow for the scoring of items when multiple response categories are allowed. Examples of polytomous scoring include Likert type scaling for attitudes, essay scoring in which different score values are awarded for different essay qualities, or any situation in which partial credit might be awarded to indicate differing levels of item performance. In short, any time a gradient that reflects varying amounts of the trait measured is applied to scoring rather than a simple right/wrong approach, polytomous models would be appropriate.

Polytomous IRT models are extended from the dichotomous models, but differ in that they use multiple parameters to represent the probability of responding in each category rather than a single item difficulty parameter. These parameters are called step difficulties or category boundaries, depending on the particular model chosen. Rather than having a single item characteristic curve (ICC) to represent the relationship between trait level and the probability of a correct response, polytomous models have multiple category characteristic curves (CCCs) which represent the relationship between trait level and the probability of responding in a given category. An examinee with a given trait level will be most likely to respond in the category whose curve is highest for their trait level.

In the family of polytomous models, there are two basic types: difference models and divide-by-total models. The difference models (Samejima, 1969; Muraki, 1990) artificially dichotomize the item to obtain the probability of responding in a given category or higher. For example, in an item with possible scores of 0,1, or 2,

the dichotomization would represent the probability of responding in category 0 versus in categories 1 or 2 and the probability of responding in categories 0 or 1, versus category 2. Adjacent probabilities are then subtracted to obtain the probability of responding in a particular category. The divide-by-total models (Bock, 1972; Muraki, 1992; Masters, 1982; Rost, 1988, Andrich, 1978) calculate the probability of responding in a given category by normalizing the probability space to sum to 1.0 This is accomplished by dividing the numerator which represents a response in a given category by the denominator which represents each possible response. While there are many different polytomous IRT models with different derivations and parameterizations, only the models examined in the current study will be discussed in detail. These models are the graded response model, the generalized partial credit model, and the partial credit model.

The Graded Response Model

The graded response model (Samejima, 1969) is an extension of the two-parameter logistic model to the polytomous case. The model assumes that item responses occur in two or more ordered categories. The graded response model is a difference model and requires two steps in which response categories are artificially dichotomized and then subtracted in order to obtain the probability of responding in a given category. The cumulative probability function of scoring in category x or higher on item i given the examinee's trait level, θ, for the graded response model is defined as:

$$P^*_{ix}(\theta) = \frac{\exp[ai(\theta - bix)]}{1 + \exp[ai(\theta - bix)]}, \quad (6)$$

where b_{ix} is the difficulty parameter associated with score category x and a_i is the item discrimination. The graded response model allows items to differ in discrimination, however, it is assumed that categories within an item are uniformly discriminating, therefore, each item has only one discrimination parameter. For an item with m+1 categories, there will be m ways to dichotomize the item, and, therefore, m $P^*_{ix}(\theta)$ equations. In order to obtain the probability of responding in a given category, adjacent cumulative probability functions are subtracted such that;

$$P_{ix}(\theta) = P^*_{ix}(\theta) - P^*_{i,x+1}(\theta). \quad (7)$$

For the lowest and highest categories (x=0 and x=m+1), the cumulative probabilities are $P^*_{i0}=1$ and $P^*_{im+1}=0$ respectively.

Samejima (1969) extended the formulation for the information function to the polytomous case. The information for a given item can be expressed as:

$$Ii(\theta) = \sum_{x=1}^{mi} \frac{[P'ix(\theta)]^2}{[Pix(\theta)]}, \quad (8)$$

where P_{ix} is equal to the probability of scoring in category x on item i, and P'_{ix} is the first derivative of P_{ix} with respect to theta, and m_i is the number of categories. Test information can be computed simply by summing the item information values. Samejima's formulation of the information function can be applied to all polytomous models.

The Generalized Partial Credit Model

The generalized partial credit model (Muraki, 1992) is an extension of the two parameter logistic model to the polytomous case. The generalized partial credit model is a divide-by-total model and allows for steps to be unordered (that a later step may be easier than a former step). The generalized partial credit model allows items to differ in discrimination, however, it is assumed that categories within an item are uniformly discriminating, therefore each item has only one discrimination parameter. The probability function of scoring in category *x* on item *i* given the examinee's trait level, θ, for the partial credit model is defined as:

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{k=0}^{x} a_i(\theta - b_{ik})\right]}{\sum_{h=0}^{m_i} \exp\left[\sum_{k=0}^{h} a_i(\theta - b_{ik})\right]}, \qquad (9)$$

where m_i is the number of score categories minus one, b_{ik} is the difficulty parameter associated with score category *x*, and a_i is the item discrimination. The generalized partial credit model simplifies to the partial credit model, when item discrimination values are uniform.

The Partial Credit Model

The partial credit model (Masters, 1982) is an extension of the one-parameter logistic (Rasch) model to the polytomous case. The probability function of scoring in category *x* on item *i* given the examinee's trait level, θ, for the partial credit model is defined as:

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{k=0}^{x}(\theta - b_{ik})\right]}{\sum_{h=0}^{m_i}\exp\left[\sum_{k=0}^{h}(\theta - b_{ik})\right]}, \quad (10)$$

where m_i is the number of score categories minus one and b_{ik} is the difficulty parameter associated with score category x. Like the generalized partial credit model, it is a divide-by-total model and allows for steps to be unordered.

Computerized Adaptive Testing with Dichotomous Items

Computerized adaptive testing (CAT) provides a bridge between the advantages of cost effective group administered testing and measurement efficient individually administered testing, by creating a hybridization which allows individualized testing in a group setting using the computer as a delivery medium. CAT is based on the idea that items which are too easy or too difficult contribute little to the information about an examinee's trait level. By eliminating the need to administer items of inappropriate difficulty, CAT can shorten testing time, increase measurement precision, and reduce measurement error due to boredom, frustration or guessing (Wainer, 1990). In CAT, an examinee's trait level is re-estimated after each item based on the responses to all previous items and testing ceases when a certain criterion of measurement precision is met (Bergstrom & Lunz, 1999). The following sections will provide a more detailed description of CAT. First, CAT will be contrasted with paper and pencil testing and the advantages of CAT outlined. Second, the components which make up a CAT will be discussed.

Advantages of CAT over Paper and Pencil Testing

CAT offers many advantages over the traditional paper and pencil testing format. The purported benefits of CAT include (Wainer, 1990):

1) Increased efficiency in testing. Examinees may take shorter tests without reduction in measurement precision because items are targeted to their trait level, so they do not waste time with inappropriate items.
2) Improved test security due to both the increased physical security of the computerized item pool and to the individualized nature of a CAT. Each examinee takes a test individually tailored to his or her trait level so that it would be necessary to learn a large portion of the item pool in order for this pre-knowledge to have any impact on an examinee's score.
3) Reduction in the negative effects of time constraints for some examinees. Since testing time can be individualized, within certain reasonable constraints, examinees can work at their own pace with the influence of speededness largely negated.
4) Reduction in examinee frustration and boredom. Since examinees see only those items appropriate for their trait level, they remain engaged and challenged.
5) Elimination of separate answer documents. All examinee responses are automatically recorded in a computer file, which eliminates the additional paperwork associated with separate answer documents, as well as eliminates the possibility of errors in recording examinee responses.

6) Immediate scoring and feedback to examinees. The computer can automatically translate the examinee's trait estimate into the chosen scaled score and report the results immediately upon completion of testing.
7) Simple pretesting of items. Since test forms are not preconstructed, new items can simply be inserted into an examinee's test form at any point.
8) Easy removal of faulty items. If a poorly performing or incorrect item is identified it may simply be removed from the item pool, without the need for expensive form reprints.
9) Ability to include new and innovative item types. The computer as a medium for test delivery offers many possibilities for the use of interactive audio, visual, or simulation based items, not possible when the test is administered on paper.

While many of these advantages remain, the past two decades of research have uncovered issues which bring into question whether certain of these benefits are recognized in operational testing. One specific example, relates to the supposed improvement in test security which a CAT provides. The third section of this literature review will address test security in a CAT environment with specific focus on methods for controlling item exposure and test overlap rates.

Components of a Computerized Adaptive Test

Reckase (1989) lists four major components of a CAT. These include, the item pool, the item selection method, the trait estimation method, and the stopping rule. In addition to these basic elements, two new components which have become

routinely adopted by testing programs using CAT are discussed. These are the content balancing and exposure control mechanisms.

Item Pool. Unlike paper and pencil testing, where items are developed and used for a particular test form, CAT requires the development of an item pool, from which all tests will be drawn. This pool will need to have a sufficient number of high quality items across a large range of difficulties to allow the CAT to estimate trait level for a broad range of examinees. In addition, care must be taken to ensure that the item pool provides a sufficient number of items in each desired content area to meet the test specifications (Wainer, 1990). The necessary size of an item pool is determined by test length, size of the examinee population, and target item exposure and overlap rates (Bergstrom & Lunz, 1999). Guidelines for the appropriate size of the item pool range from six to twelve times the number of items seen on a paper and pencil form (Stocking, 1993;Patsula & Steffan, 1997; Luecht, 1998). However, issues of item exposure, item retirement, and pool rotation may require this number to be much larger. Due to the continuous nature with which many CATs are administered, the useful life of an item or a item pool is limited. Luecht (1998) suggests that between 3,800 and 21,000 items may be needed to begin a CAT program when sufficient pool size, multiple pools, and item pretesting are taken into consideration. Strategies to extend the life of a pool such as drawing multiple overlapping pools from an item vat (Patsula & Steffan, 1997) have been proposed, however, the cost and effort to create and maintain a CAT item pool remains formidable and far exceeds that of paper and pencil testing.

Item Selection Procedure. In a paper and pencil test, examinees begin with the first item and proceed in a largely linear fashion, taking each item in sequence, until the end of the test is reached. In CAT, however, items are selected adaptively based on the examinee's trait level. The trait estimate for an examinee is updated after each item and the next item administered is one that maximizes a given mathematical function (McBride, 1997). The two most popular methods of item selection in CAT are maximum information and Owen's Bayesian (Owen, 1969). Items selected according to the maximum information criterion are those which provide the most information at the examinee's current trait estimate. Information values are calculated for each item at the estimated trait level and the item with the largest information is selected for administration. Items selected according to the Owen's Bayesian criterion are those which maximize the expected posterior precision of the trait estimate, or, conversely, that will minimize the expected posterior variance of the trait estimate. After each item is administered, the posterior distribution of trait level is computed. All items are evaluated and the item which will maximally reduce the posterior variance is selected. The Owen's Bayesian method is computationally easier than the maximum information method, which made it more popular when computing power was limited. However, increases in available computing power and the fact that with the Owen's Bayesian method, estimated trait level varies as a function of item order have made maximum information more widely used (Wainer, 1990).

Trait Estimation Method. Two issues need to be addressed in terms of estimating trait level for examinees. First, since no information is known about an examinee at the beginning of the test administration, an initial value for trait level must be supplied. This value is commonly the expected mean trait level of the testing population, however, other values may be used when prior information is available or when it is desirable to limit the exposure of an initial item.

Second, after each item is administered an examinee's trait level is re-estimated based on his or her responses to all previously answered items. This can be accomplished either through maximum likelihood estimation (MLE) or one of the Bayesian estimation approaches. MLE determines the most likely trait level for an examinee given the response string to items with specified parameters, by multiplying together the individual probabilities of a correct response given theta to compute a cumulative probability with the function,

$$L(\theta i \mid x_i) = P(x_i \mid \theta i, \beta) = \prod_j Pj(\theta i)^{xij} Qj(\theta i)^{1-xij}, \qquad (11)$$

where x_i is the vector of item responses for examinee i to items 1 to j and β is the matrix of item parameters. $L(\theta_i|x_i)$ represents the likelihood that a particular response vector, x_i, would be observed if θ were the true value for examinee i. The function is plotted across values of theta and the modal value is used as the new trait estimate. Mathematically, the most likely value of θ can be found by taking the derivative of the likelihood function, $L(\theta_i|x_i)$, setting the result equal to zero and solving for θ using the Newton-Raphson iteration procedure (Wainer, 1990).

The Bayes Modal estimation procedure works in the same fashion as described above except that the likelihood function is multiplied by a prior distribution of theta such that,

$$p(\theta \mid x_i) \propto L(\theta \mid x_i) p(\theta), \qquad (12)$$

where $p(\theta)$ represents the prior distribution of θ. In essence, the Bayes Modal procedure treats the prior as an additional item and multiplies it in with the product of the probabilities of the other item responses. Note that when no prior is used, $p(\theta)$ has the same uniform value for all θ and the Bayes Modal procedure simplifies to MLE (Wainer, 1990).

Expected a posteriori (EAP) estimation is popular variant on the Bayes Modal procedure which divides the prior distribution into many distinct quadrature points rather than evaluating it as a continuous distribution (Bock & Aitkin, 1981; Bock & Mislevy, 1982). EAP is mathematically easier to implement than the Bayes Modal procedure.

The Bayesian procedures have the advantage of smaller standard errors than with MLE for the same number of items administered because prior information is known. However, use of a bad prior can result in the need to administer more items to recover and a regression toward the mean in trait estimation tends to occur. With the Bayesian procedures a trait estimate can be obtained after the first item response. MLE cannot estimate trait level until one correct and one incorrect response are obtained. Thus, MLE cannot be used after the first item or in the case in which an

examinee answers all items correctly or incorrectly. Therefore with MLE, a variable stepsize—moving the trait estimate half the distance to the most extreme item—is typically recommended until estimation can occur.

Stopping Rule. A paper and pencil test ends when the examinee completes the final item or when time runs out. Several methods have been proposed to determine when to stop administering items in a CAT and compute a final trait estimate. One such method deemed "fixed length" requires that all examinees take some predetermined number of items. While this produces equity in terms of test length, measurement precision will differ across trait levels, with the worst precision typically being obtained for examinees of low or high trait level. Conversely, the "variable length" method requires that examinees continue to take items until some predetermined level of precision is reached, such as a target level of information (standard error) or posterior precision. While this produces equity in terms of measurement precision, examinees will take tests of different lengths (Wainer, 1990). Variable length tests tend to make better use of the item pool as they minimize test length (Bergstrom & Lunz, 1999). However, Lunz and Bergstrom (1994) found that it may be difficult to explain the equity of variable length tests to examinees. In practice, some minimum or maximum number of items may be applied with the variable length stopping rule.

Content Balancing. In a paper and pencil exam, appropriate content coverage can be ensured by writing items to match a table of specifications. Since all examinees take all items, all examinees are ensured a valid test covering all content

areas. In a CAT, however, each examinee is administered a unique set of items. Equivalent content coverage across examinees may be achieved in a number of ways through the use of what are commonly called "content balancing" techniques.

The first issue to be considered in content balancing is the number and type of content areas or item features that are of interest to the test developer. Stocking & Swanson (1993) discuss three levels of non-statistical item features that may be used in classifying and selecting items for administration in a CAT environment—intrinsic item properties, overlap constraints, and item set constraints. Intrinsic item properties are those features of an item which contribute to its content area, item format, etc. Overlap constraints refers to the fact that items may be similar to or inform each other in a way that is not relevant to the trait being measured. Item set constraints acknowledge that certain items may share a common stimulus or set of directions. Some researchers and testing programs have opted to consider all three levels of item features (Stocking & Swanson, 1993; Parshall, Davey, & Nering, 1998), while others have chosen to consider only the major content classifications represented by the intrinsic item properties (Morrison, Subhiyah, & Nungester, 1995; O'Neill, Lunz, & Thiede, 1998; Kalohn & Spray, 1998; Bergstrom & Lunz, 1999). The major difference between these two approaches is in the number of "contents" considered (as many as 50 in the former approach and as few as three to five in the latter) and the particular technique implemented for content balancing. The weighted deviations model (WDM) proposed by Stocking and Swanson (1993) may be necessary for large

numbers of content constraints, whereas less complex methods may be applied when fewer contents are of interest.

Equivalent content coverage across examinees may be achieved in a number of ways. When content areas are disparate or introduce additional dimensionality, one option is to split the item pool by administering separate tests, with separate trait estimations for each content area (Segall, Moreno, & Hetter, 1997). Another option is to use a multidimensional model and estimate separate trait levels within a single test (Parshall, Davey, & Nering, 1998). However, when content areas are shown to measure a single trait dimension it is possible to design the item pool with item quantities proportional to the desired content coverage for each examinee's test (Segal, Moreno, & Hetter, 1997). For example, in a test of mathematical ability it might be desirable to ensure equal coverage for the content areas of addition/subtraction, multiplication, long division, and fractions. Such a decision would dictate that 25% of the item pool would represent each content area. Programmatic restrictions to the item selection procedure would be necessary to ensure the desired content coverage during the CAT.

Kingsbury and Zara (1989) discuss a procedure whereby an item is selected for administration not only according to the properties of its parameters, but also by examining the discrepancies between the desired percentage of items to be administered from a given content area and the actual number of items administered from that content area. Segall, Moreno, and Hetter (1997) present a strategy used for the General Science section of the CAT-ASVAB (Armed Services Vocational

Aptitude Battery) whereby the three primary content areas are administered in a rotational fashion to mirror the proportion of items from each content area administered in the paper and pencil reference test. Within a specific content area, the next item to be administered can be selected based on maximum information as constrained by the level of exposure control. Content areas are then rotated through so that each content area will have an approximately equal representation in the examinee's test.

One possible drawback for any content balancing method is that the most informative item in the selected content area may not be the most informative item available in the item pool. This threatens measurement precision (in a fixed length test) and could result in longer tests (in a variable length test) due to administration of sub-optimal items.

Exposure Control. CAT administration creates special circumstances that lead to over-exposure of certain items within the item pool to examinees. Frequently exposed items will cease to be a valid measure of the trait because they may have been compromised (Parshall, Davey, & Nering, 1998). Any time an examinee has prior knowledge of an item, his/her response will not be an accurate measure of his/her true trait level. One source of prior knowledge comes from the necessity of pretesting items. An examinee may see an item in its pretest and again in operation. In order to minimize the damage from this exposure, items that are pretested together can be put into separate item pools (Stocking & Lewis, 2000).

Continuous test administration poses another security risk, as examinees may seek to share information with one another (Stocking & Lewis, 2000). Luecht (1998) describes a threat from what he calls "Examinee collaboration networks (ECNs)." These ECNs are global groups of examinees who seek to pool their resources and test experience to discover a sufficient number of test items from an item pool to artificially increase scores. While not all examinees are associated with an ECN, the higher the stakes associated with scores from a given test, the more tempting cheating becomes. Luecht (1998) compares the threat to test security to those threats faced by casinos and lotteries. Any time stakes are high, cheaters will look for anything that will give them an advantage. Stocking and Lewis (2000) note that the amount of effort a testing program should put into exposure control does, indeed, depend on the uses of the test scores and whether the testing program can be classified as high, medium, or low stakes.

Finally, the CAT item selection algorithm itself, constitutes a threat to item security. Under maximum information item selection, certain items will be administered to almost all examinees and a small proportion of items available will account for a large proportion of items actually administered (Parshall, Davey, & Nering, 1998). Wainer and Eignor (2000) describe the relationship between the frequency of item usage and the rank order of item usage as defined by a modified version of Zipf's law (Zipf, 1949; Wainer, 2000; Wainer & Eignor, 2000). Specifically, the log of the frequency of an event (item administered) is related to the rank order of the event in a linear fashion. In the case of item selection, an item with

a low rank order (close to 1) is administered more frequently than an item with a high rank order. They proposed using Zipf plots to examine the usage of items within a pool. A steeper slope indicates highly differential item usage, whereas a flatter slope would indicate more uniform item usage. They demonstrated that typically, CAT item selection produces a plot with a steep negative slope indicating that not only are the pool's most informative items most often administered, but also that a very small percentage of the pool's items account for a very large percentage of administered items. In other words, the most popular items are popular by a large margin, resulting in an exponential decline in item usage as rank increases. In addition to the Zipf relationship of CAT item selection, when the same initial trait estimate is used for all examinees, the initial item sequences are limited and predictable, becoming easily overexposed. Furthermore, with maximum information item selection, two examinees with the same trait estimate will see the same item (Hetter & Sympson, 1997).

Parshall, Davey, and Nering (1998) discuss the three often conflicting goals of item selection in CAT. First, item selection must maximize measurement precision, by selecting the item which maximizes information or posterior precision for the examinee's current trait level. Second, item selection must seek to protect the security of the item pool by limiting the degree to which items may be exposed. Third, item selection, must ensure that examinees will receive a content balanced test. Stocking and Swanson (1998) add a fourth goal to this list, stating that item selection must also maximize item usage so that all items in a pool are used, thereby ensuring

good economy of item development. Stocking and Lewis (2000) portray the item selection problem as a balloon—pushing in on one side will cause a bulge to appear on another.

Overview of Exposure Control Methods

Different approaches to the goals of item selection will produce different testing algorithms (Stocking & Lewis, 2000). Attempts to address the third goal are denoted exposure control methodologies (Parshall, Davey, & Nering, 1998). Way (1998) discusses two types of exposure control strategies—randomization and conditional selection. Randomization strategies randomly choose the next item for administration from a set of nearly optimal items rather than selecting the single most informative item. The conditional selection strategies are those in which the probability of an item being administered is controlled conditional on a given criterion, such as the expected frequency of item usage. Finally, a third approach to exposure control has recently been proposed by Chang and Ying (1996) in which items in the pool are stratified according to their statistical properties (item parameters) and items are constrained to be administered from certain strata. This section will provide a detailed description of exposure control methodologies which exemplify these three categories. The randomization procedures will be presented first, followed by the conditional selection procedures, and finally the stratification procedures.

Before beginning this discussion, however, a brief overview of the most popular method for controlling item exposure—the Sympson-Hetter procedure—is

provided, so that comparisons of other procedures to the Sympson-Hetter can be better understood. The Sympson-Hetter is an example of a conditional selection procedure which seeks to control exposure by assigning exposure control parameters to each item. These parameters are derived through a series of simulations in which the frequency with which an item is administered is recorded. Items which are administered more frequently are assigned more restrictive exposure control parameters. Items which are administered less frequently are assigned less restrictive exposure control parameters. In operational testing, items are selected according to the item selection procedure (maximum information), but an item will be administered only if the value of its exposure control parameter exceeds that of a random number. The Sympson-Hetter procedure (or one of its many variants) is the most frequently used method of controlling item exposure and represents the standard to which all other methodologies are compared.

Randomization Procedures

Randomization procedures are usually considered to be easy to understand and simple to implement, but provide no guarantee that item exposure will be constrained to a given level. This section will present five different types of randomization procedures and will also discuss possible variations or combinations for each of these.

5-4-3-2-1 procedure. Early research by the military for the CAT-ASVAB (Armed Services Vocational Aptitude Battery) spawned a randomization procedure which sought to control the exposure of items early in a CAT administration through

what has been called the 5-4-3-2-1 procedure (McBride & Martin, 1983; Hetter & Sympson, 1997). The first item to be administered in a CAT is not the single most informative, but rather, is randomly selected from among the five most informative items. The second item to be administered is selected from among the four most informative items, the third from among the three most informative, and so on, until the fifth item when maximum information item selection resumes (Hetter & Sympson, 1997). The number of items to which the random component may be applied can, of course, be customized to the test developer's needs, with maximum information item selection resuming after the third item or not until the eighth. The goal of the procedure is to alter the ordering of item administration among the earliest items based on the recognition that after the first few items, examinee trait estimates will become distinct enough that test overlap will be minimized (Stocking, 1992).

The procedure, while extremely simple to implement has taken much criticism. While it does appear to reduce the predictability of initial item sequencing, it still overexposes the pool's most informative items (Hetter & Sympson, 1997) and provides no specific guarantee that item exposure will be constrained to a given level (Parshall, Davey, & Nering, 1998). Items which are considered in the group for administration, but not administered are not blocked from further selection, so it is likely that the pool's most informative items will eventually be exposed (Davey & Parshall, 1995). Further research has even suggested that the procedure does not noticeably increase item security over the no exposure control condition (Chang, 1998).

Randomesque procedure. Kingsbury and Zara (1989) proposed the randomesque method for controlling item exposure. Similar to the mechanism employed by the 5-4-3-2-1 procedure, the randomesque procedure randomly selects the next item to be administered from a group of the most informative items, rather than selecting the single most informative item. However, rather than returning to maximum information item selection after the initial few items, the randomesque method continues to employ a random component throughout the test administration. The size of the group from which an item is randomly administered may be customized for each testing program. The goal of this procedure is not just to reduce item overlap in the beginning of the test, but also to control the exposure of items to examinees with similar trait levels later in the test (Stocking, 1992; Morrison, Subhiyah, & Nungester, 1995).

Morrison, Subhiyah, & Nungester (1995) implemented the randomesque procedure in a variable length CAT for the In-Training Examination in Internal Medicine using the Rasch model. They randomly selected the next item for administration from the five most informative items. Their results showed that the randomesque procedure was not effective for reducing item exposure; concurring with the results of Eignor, Stocking, Way, and Steffen (1993). They speculated that a group size of five may not have been large enough to affect changes in the exposure rate. This illustrates the largest criticism of the randomesque procedure, that there is no way to decide a-priori the size of the item group from which the item should be

randomly selected, leaving only trial and error which can be frustrating and time consuming (Stocking, 1992).

Revuelta and Ponsoda (1998) also investigated the randomesque procedure with a group size of five with fixed 35 item and variable length (standard error less than .22 or a maximum of 50 items) CATs. They found a modest increase in pool utilization and a slightly decreased maximum exposure rate over the maximum information condition with the use of the randomesque procedure at little to no expense to measurement precision, but overall evaluated the procedure poorly in comparison to other options which are discussed below.

Thomasson (1998) evaluated a modified randomesque procedure which he called "Choose 1 of 3." In this variant, an item is randomly chosen to be administered from among the three most informative items, with the two items not selected for administration being blocked from further consideration for the current examinee. He evaluated this procedure along with several variants of the Sympson-Hetter procedure, including one designed to mimic the one-third probability of item administration expected with the "Choose 1 of 3" technique with a 15 item fixed length CAT based on the CAT-ASVAB Math Knowledge item pool. He concluded that the modified randomesque procedure was robust to changes in distributions of trait level and that it appeared to use the item pool more efficiently than the comparable modified Sympson-Hetter alternative while still providing the same level of exposure control and measurement precision.

Within .10 logits procedure. Lunz and Stahl (1998) created a randomization procedure whereby rather than selecting the single most informative item for an examinee's trait estimate, all items within 0.10 logits of the needed item difficulty are available for selection and the item to be administered is randomly chosen from among them. This procedure, which was designed to work with the Rasch model, does not utilize information in item selection, but rather matches the current trait estimate directly to the difficulty values of the items. The number of items which will appear in any selection grouping will depend on the distribution of item difficulties in the pool and within a given content area. If there are no items within 0.10 logits of the required item difficulty, the algorithm will randomly select the item having the closest difficulty to the target.

Lunz and Stahl (1998) evaluated this procedure with five different item pool sizes ranging from a low of 183 to a high of 823 items in a content balanced variable length CAT. They concluded that the procedure worked well to control test overlap, especially with the larger pool sizes. They further point out that even though test overlap may be extensive in smaller pools and for candidates of similar trait levels, the procedure served to alter the sequential ordering of items presented such that each candidate had a unique testing experience. Changing item order has often been used to control security on paper and pencil tests, so it is reasonable to assume that different item orderings in CAT may also provide some security. Bergstrom and Lunz (1999) further demonstrate the utility of the procedure with a 900 item pool, concluding that maximum exposure rate is less than 30% for most items, with only a

few items near the pass point and in content areas where the number of items was insufficient to meet target constraints being administered with higher frequency.

Restricted Maximum Information procedure. Revuelta and Ponsoda (1998) sought to develop randomization methods which would be similar in exposure control performance to the Sympson-Hetter, but which would be less complex and more straightforward to implement. One of the methods they proposed is called the restricted maximum information procedure. With this method, items are removed from the item pool when their exposure rate exceeds some predetermined level and are only returned once the increase in tests administered forces the exposure rate back below the critical level.

Like the Sympson-Hetter, a target maximum exposure rate, r, is set that no item's frequency of administration should exceed. The number of tests on which an item has appeared is divided by the total number of tests which have been administered. This proportion is compared to the target exposure rate and if it exceeds this rate, the item is removed from the pool for the next test administration. After each test administration, all items, both those included and excluded from the current pool are evaluated and their probability of administration compared to the target. As an excluded item lies dormant, its probability of administration decreases as the total number of test administrations increase, but the number of tests on which it has been administered remains unchanged. Eventually the probability of administration for the item will drop below the target value and the item may again be included for selection in the item pool. In this way, items will constantly be moving

in and out of the available pool. Within the available pool, items are selected according to maximum information. Exposure is controlled simply by the removal of overexposed items from consideration for administration.

Revuelta and Ponsoda (1998) evaluated the restricted maximum information method using a target exposure rate of 0.40 with both fixed (35 item) and variable (standard error less than 0.22 or maximum of 50 items) length tests with a 221 item pool. They compared its performance to other randomization procedures (5-4-3-2-1, randomesque, progressive) and to the Sympson-Hetter using the same 0.40 target maximum exposure rate. They found that the restricted maximum information procedure was comparable to the Sympson-Hetter in terms of both measurement precision and maximum observed exposure rate with both procedures requiring an average of 3-4 more items in the variable length conditions to attain the same level of precision as the optimal condition and both procedures successfully constraining exposure to the desired 0.40 level. While the restricted maximum information procedure had slightly better pool utilization than did the Sympson-Hetter, the percent the of pool that went unused remained greater than 15%.

Progressive procedure. Revuelta and Ponsoda (1998) proposed a second randomization procedure which they termed the progressive method. In this method a weight is computed for each item, *i*, according to the formula

$$W_i=(1-s)R_i + sI_i, \quad (13)$$

where *s* represents the serial position in the test (how many items have been administered divided by the total test length), *I* represents the item information at the

estimated trait level and *R* represents a random uniform number. The item with the largest weight is administered. Item information is weighted by the serial position in the test and so is less important early in the test, but increases in importance as the test progresses. Conversely, the random component is weighted by one minus the serial position in the test and so is most important for items early in the test, but becomes less important as the test progresses. The authors postulate that because initial trait estimates can differ substantially from final trait estimates that the progressive method should offer exposure control without greatly reducing measurement precision.

Revuelta and Ponsoda (1998) examined the progressive procedure using the same study design reported for the restricted maximum information procedure. The results showed that measurement precision was comparable to the Sympson-Hetter and restricted maximum information procedures, but slightly worse than optimal. The progressive procedure had 100% pool utilization, however, like the 5-4-3-2-1 and the randomesque procedures had a high maximum exposure rate (0.64).

Revuelta and Ponsoda (1998) concluded that since the progressive procedure had good pool utilization, but high maximum exposure rates and the restricted maximum information procedure had low maximum exposure rates, but poor pool utilization, that a combination of the two procedures might provide a good solution. The combined procedure would weight items according to the progressive strategy, but would only consider those items included in the available pool according to the restricted maximum information strategy. In a second study, they compared the

combined procedure under two different levels of target maximum exposure rate (0.40 and 0.15) to both procedures individually and the Sympson-Hetter. They evaluated three fixed test lengths (20, 40, and 60) and two item pool sizes (500 and 1000 items). The results revealed that while the combined procedure with the most restrictive target maximum exposure rate (0.15) had poor measurement precision, the other conditions performed well. Previous findings for the individual procedures in terms of pool utilization and maximum exposure rates were replicated and the combination procedure under the 0.40 target exposure rate was successful in producing both good pool utilization and low maximum exposure rates.

Conditional Procedures

The conditional selection strategies are those in which the probability of an item being administered is controlled conditional on a given criterion. The advantage of the conditional procedures is that they allow a preset target exposure rate (r) to be set and provide a reasonable guarantee that exposure will be constrained to this level. The next section presents four different types of conditional selection procedures. The Sympson-Hetter procedure is discussed first, followed by the conditional Sympson-Hetter procedure, the Stocking and Lewis multinomial procedures, and finally, the Davey-Parshall procedures.

Sympson-Hetter procedure. The Sympson-Hetter procedure (Sympson & Hetter, 1985; Hetter & Sympson, 1997) attempts to directly control the rate of item exposure through the use of probabilistically determined exposure control parameters assigned to each item. The goal of the Sympson-Hetter is to constrain the maximum

probability of administration for an item to a predetermined level. The advantage of the Sympson-Hetter is that it allows a preset target exposure rate (r) to be reasonably ensured as long as the characteristics of the sample to which the test is given are distributed the same as those from which the parameters were derived. The target exposure rate stipulates what maximum proportion of the population should see each item (Hetter & Sympson, 1997).

In effect, the Sympson-Hetter attempts to control exposure by differentiating between the probability of selecting an item, *P(S)*, and the probability of administering an item, *P(A)*. This is accomplished by implementing exposure control parameters to modify *P(S)* according to

$$P(A) = P(A|S) * P(S) , \quad (14)$$

where $P(A|S)$ is the probability of administering an item given that it is selected and is used as the exposure control parameter (K_i) (Stocking & Swanson, 1998).

The Sympson-Hetter uses the frequency with which an item is administered in a simulated sample of examinees to determine the rate of exposure for each item. This iterative process results in each item being assigned an exposure control parameter (K_i) with a value between zero and one. These parameters are then used in live testing to constrain the probability of administering an item. When an item is selected for administration by maximum information or other optimal selection strategy, the corresponding exposure control parameter must be compared to a random number drawn from a uniform distribution. If the value of the item's K_i exceeds the random number, the item is administered. Otherwise, the item is

blocked from further administration and the next most informative item is selected for consideration. This process then continues until an item is administered.

In practice, the Sympson-Hetter proceeds in two phases, the first of which establishes the exposure control parameters and the second of which uses the final parameters from the first phase in operational testing. To establish the exposure control parameters, all K_is are initially set to 1.0, indicating that if the item is selected, it will be administered. A simulated CAT is then administered to a set of simulees with known trait levels, making a comparison for each item selected between a random uniform number and the item's K_i. The proportion of times an item is selected and administered are computed for each item according to the following formulae:

$$P(S) = NS/NE, \quad (15)$$

$$P(A) = NA/NE, \quad (16)$$

where *NS* is the number of times an item was selected, *NA* is the number of times an item was administered, and *NE* is the number of examinees. Note that for the first iteration where all K_is equal 1.0, *P(S)* will equal *P(A)*. The probability of selection is then compared to the desired target value (r) and adjustments are made to the K_is as follows:

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

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

If an item is selected more often than desired, the new exposure control parameter will be more restrictive. If an item is selected less often than desired, the new exposure control parameter will be set to 1.0, indicating no restriction on its administration. This process is then iteratively repeated, using the K_is from the previous iteration, until the maximum probability of administration for any item approaches a limit slightly above *r*. After each iteration, it may be necessary to make some adjustment to the exposure control parameters to ensure that a sufficient number of items are available to allow for a complete test to be administered. To accomplish this, it is suggested that the 'n' largest (least restrictive) K_is be set to 1.0, where 'n' is equal to the test length. The final round of K_is are then used in operational testing (Hetter & Sympson, 1997).

Stocking (1992) extended the Sympson-Hetter to work with the more complex testing scenario of item blocks which correspond to the same stimulus and must be administered together by controlling passage exposure rates as well as item exposure rates. Using a target exposure rate of 0.20 for items and passages, she compared the Sympson-Hetter to a modification of the 5-4-3-2-1 randomization procedure. She concluded that the Sympson-Hetter provided better exposure control than did the randomization procedure.

The primary impact of the Sympson-Hetter, other than controlling item exposure, is that it forces the administration of sub-optimally informative items. While Hetter and Sympson (1997) found that the use of the Sympson-Hetter did not significantly impact measurement precision, other studies have found an increase in

standard error of measurement and a decrease in reliability over the optimal condition (Stocking, 1992; Parshall, Davey, & Nering, 1998; Chang, 1998).

Other criticisms of the method include the fact that the simulations to obtain the exposure control parameters are time consuming and that if even a single item is added to or deleted from the item pool, the simulations must be rerun (Stocking & Lewis, 2000). In addition, the exposure control parameters are dependent on the distribution of examinee trait level used in the simulation. If the distribution differs in operation, the exposure control parameters are no longer valid and cannot guarantee the expected level of exposure control (Parshall, Davey, & Nering, 1998; Stocking & Lewis, 2000). Parshall, Davey, and Nering (1998) also point out that the procedure tends to produce more relaxed exposure control parameters for items whose information peaks at the tails of the distribution, because of the sparseness of examinees at those trait levels. While the item may not be seen by more than the targeted proportion of the population, almost all examinees of low or high trait level will see it. This is especially problematic in test-retest scenarios as the person's true trait level does not change and they are likely to be exposed to the same items. Additionally, there is reason to believe that examinees are more likely to have friends of the same trait level and may share information with them. Finally, Stocking (1992) demonstrates that when implementing the Sympson-Hetter in conjunction with content constraints, the procedure may not converge if the item pool does not closely match the test specifications.

Conditional Sympson-Hetter. Even though an item's global exposure rate may be 20%, it may, in fact, be seen by most or all test takers of high or low trait level. This becomes an issue in test-retest scenarios where the examinee may see the same items on the retest as on the initial test. Additionally, test takers are more likely to have friends of similar trait level so shared information about test questions may be more beneficial for increasing scores than the global exposure rate would predict. The threat from high trait level examinee exposure and overlap is especially troublesome since efforts to steal tests are usually focused on the more difficult questions and may employ individuals with good memorization skills who are most likely to be of high trait level. These observations resulted in the development of an extension to the Sympson-Hetter procedure which establishes exposure control parameters conditional on trait level (Parshall, Davey, & Nering, 1998; Chang, Ansley, & Lin, 2000). It is important here to distinguish between conditional selection strategies which select items based on a given condition, such as frequency of administration, and conditional exposure control which provides control of item exposure conditional on trait level. While most attempts to control exposure conditional on trait level have been implemented with conditional selection strategies, it is possible to implement conditional exposure control with a randomization or stratification procedure.

The conditional Sympson-Hetter procedure is conducted in exactly the same fashion as the Sympson-Hetter procedure with the exception that the frequency of item administration is tallied separately for each of 'm' discrete trait levels. The

simulations to develop the exposure control parameters are then conducted separately for each level of theta. The result is a matrix of n(items) X m(theta values) which yields the conditional exposure control parameters. In operation, the examinee's current trait estimate determines which column of parameters will be used.

The advantages of this procedure are that a test developer may exercise direct exposure control at different trait levels and may even select different target exposure rates for different levels within the trait distribution In addition, the exposure control parameters are not dependent on the trait distribution used in the simulation. However, any addition or deletion of items from the pool will continue to require the exposure control parameters to be recalculated and the simulations for the conditional Sympson-Hetter procedure are more time consuming and difficult than those for the Sympson-Hetter procedure.

Parshall, Davey, and Nering (1998) compared the performance of the conditional Sympson-Hetter to the unconditional version of the procedure with a 600 item pool using a variable length stopping rule. They found that while, the conditional Sympson-Hetter provided better pool utilization, it overexposed a much larger portion of the items than the unconditional Sympson-Hetter. The authors attributed this to the lack of global (across trait level) exposure control with the conditional Sympson-Hetter. In addition, the conditional Sympson-Hetter yielded longer test lengths to examinees whose trait levels were in the tails of the distribution, but provided lower rates of test overlap for these same examinees.

Stocking & Lewis Multinomial Procedures. Stocking (1992) observed some difficulty in establishing a smooth convergence of the Sympson-Hetter parameters due to the "fix-up" of setting the 'n' largest parameters equal to 1.0, where 'n' is equal to the test length. While this corrective measure seems to work well when the 'n' parameters all have values reasonably close to 1.0, if the parameters to be corrected are too far from 1.0, the correction can cause fluctuations in the iteration process, and, consequently, the parameters will not converge (Stocking & Lewis, 1995). Stocking and Lewis (1995) proposed a new procedure based on the same technique as the Sympson-Hetter, but using a multinomial model to deal with these convergence problems. The logic behind the procedure remains the same as that behind the Sympson-Hetter, but the new procedure has fewer problems with convergence during the exposure control parameter iterations (Chang, Ansley, & Lin, 2000). Like the Sympson-Hetter, the Stocking and Lewis Multinomial procedure develops a K_i for each item, but it uses a multinomial model for item selection (Chang, Ansley, & Lin, 2000).

Item selection is based on a distribution of multinomial probabilities—that is, the probability that a given item is selected and administered and that all previous items have been rejected. The multinomial probabilities are calculated such that

$$K_1 = P_1(A|S), \quad (17)$$

$$K_2 = (1-P_1(A|S))*P_2(A|S), \quad (18)$$

$$K_3 = (1-P_1(A|S))*(1-P_2(A|S))*P_3(A|S), \text{ etc.}, \quad (19)$$

where K_1 is the probability that the first item is selected and administered, K_2 is the probability that the first item is rejected and the second item is selected and administered, and K_3 is the probability that both the first and second items are rejected and that the third item is selected and administered. Note that the probability of an item being administered given that it is selected, $P_i(A|S)$, is the Sympson-Hetter exposure control parameter. These probabilities are then summed together to create a cumulative multinomial distribution. A random number is drawn from a uniform distribution and compared to this cumulative distribution of probabilities. The location in the cumulative distribution which corresponds to the random number then determines which item will be administered. All items occurring in the list prior to the administered item are blocked from further administration (Stocking & Lewis, 1995; 2000).

Stocking and Lewis (1995) investigated the usefulness of this procedure with the 3PL model using a 35 item fixed length CAT. Item exposure was examined at four different target levels—0.10, 0.20, 0.30, and 0.40. They concluded that the multinomial model provided smoother convergence than the Sympson-Hetter at all levels of target exposure rate and that it was still able to provide a guaranteed level of exposure control.

Stocking and Lewis (1998) evaluated the performance of a conditional version of the multinomial procedure against that of the unconditional multinomial procedure. Like the conditional Sympson-Hetter, the conditional multinomial procedure develops an n(item) X m(trait level) matrix of conditional exposure parameters which

allow exposure to be controlled for examinees of similar trait level. Item selection for the simulations to set the exposure control parameters is conducted, however, according to the multinomial model. They ran a series of simulated 28 item fixed length CATs using the 3PL. The unconditional procedure was evaluated at a target global exposure rate of 0.20 and the conditional procedure was evaluated at a target exposure rate of 0.20 for all trait levels. They found that the conditional procedure yielded slightly poorer reliability and slightly higher conditional standard errors of measurement than the unconditional procedure, but that the conditional procedure increased pool utilization over the unconditional procedure and greatly evened out and reduced test overlap, especially for extreme trait levels.

Chang (1998) compared the conditional multinomial procedure to the unconditional multinomial, the Sympson-Hetter, the 5-4-3-2-1 procedure, and the Davey-Parshall procedure (which will be described in the next section). He determined that the randomization procedure produced the worst results, with the Sympson-Hetter and the unconditional multinomial procedures performing better, but similarly, and the Davey-Parshall and the conditional multinomial performing best. He concluded that the conditional multinomial performed the best when all factors were considered.

Chang, Ansley, & Lin (2000) compared the conditional multinomial, the conditional Sympson-Hetter, and the Davey-Parshall procedures using four pool sizes (360 ,480, 600, and 720 items) and two different target exposure rates (.10 and .20) in a fixed length 30 item CAT. They found that while the conditional multinomial

procedure performed well, the conditional Sympson-Hetter actually outperformed it. The conditional multinomial had higher observed exposure rates than did the conditional Sympson-Hetter especially when the pool size was small or when exposure control was most restrictive and that the variability in exposure rates across ability levels was somewhat less for the conditional Sympson-Hetter than for the conditional multinomial. The authors concluded that the conditional Sympson-Hetter yielded better performance than the conditional multinomial and provided an easier method for obtaining the exposure control parameters.

Stocking and Lewis (2000) have proposed a modification to the conditional multinomial procedure which would condition on estimated trait level rather than on true trait level. They conduct two experiments (one conditioning on true trait level, the other on estimated trait level) where they attempt to set different rates of target exposure for different trait levels. The results demonstrate that due to discrepancies between true and estimated trait level, setting different target exposure rates for different trait levels is ineffective when conditioning on true trait level is used. However, when exposure control parameters are conditioned on estimated trait level, the correspondence between observed and target conditional exposure rates is greatly increased, at the expense, however, of slightly lower reliability.

Davey-Parshall procedures. The Davey-Parshall procedure (Davey & Parshall, 1995) was developed as a modification to the Sympson-Hetter which allows for exposure control parameters to be established not just for each individual item, but also conditionally for pairs or groups of items. The purpose of this procedure is not

just to control overall item usage, but also test overlap. Parshall, Davey, & Nering (1998) observe that in an unconstrained CAT, groups of items will often cluster together—appearing together repeatedly across test administrations. The Davey-Parshall focuses on situations where substantial test overlap could mean that examinees could share large portions of the test with one another or where test-retest of the same examinee is likely. By conditioning on items previously administered (rather than on trait level as in the conditional Sympson-Hetter or the Stocking and Lewis conditional multinomial procedures), the extent that tests overlap across examinees or across testing occasions can be controlled. Once one item of a set appears in a test administration, the Davey-Parshall restricts the administration of other items in that set (Parshall, Davey, & Nering, 1998).

The procedure for the Davey-Parshall is similar to the Sympson-Hetter and other conditional procedures in that it establishes exposure control parameters through a set of simulations. The Davey-Parshall establishes both an individual item exposure control parameter (unconditional) and an item exposure control parameter for each pair of items. This results in an nXn matrix of parameters where n is equal to the number of items. The unconditional exposure control parameters make up the diagonal of the matrix, while the off diagonal elements are composed by the conditional exposure control parameter for each pair of items. The Sympson-Hetter can be viewed as a special case of the Davey-Parshall where all the off diagonal elements are set to 1.0.

All values in the parameter matrix are initialized to 1.0 indicating no constraints on item selection. A series of simulated CATs is administered and the number of times each item appears alone and as part of a pairing is counted. The frequency of administration for individual items is then compared to the desired target value. If the probability of administration is greater than the target value, the exposure control parameter is adjusted downward, and if the probability of administration is less than the target value, the exposure control parameter is adjusted upward (Davey & Parshall, 1995). For item pairs, the adjustment procedure is slightly more complicated. A modified chi-square test of association is performed to determine if a given pair of items is occurring together more frequently than would be expected by chance. If the chi-square value exceeds a predetermined threshold, then the item pair exposure parameter is adjusted up or down to compensate. To use the parameters operationally, multiply the item exposure control parameter by the mean of the set of all item pair exposure control parameters for those items previously administered. This result then becomes the probability with which an item will be administered if selected.

Davey and Parshall (1995) compared this procedure to the Sympson-Hetter and a no exposure control condition with two item pool sizes (100 and 200 items) using a variable length stopping rule. They set unconditional target exposure control rates of .25 and .15 respectively for the 100 and 200 item pools. While test length was increased to about the same degree with both the Sympson-Hetter and the Davey-Parshall, the Davey-Parshall reduced overlap more substantially than did the

Sympson-Hetter, especially for test-retest overlap. This latter finding was not surprising as it is in test-retest scenarios where item clusters are most likely to form and therefore the Davey-Parshall would be most effective. They concluded that the Davey-Parshall reduces overlap at no additional cost to measurement precision above the Sympson-Hetter procedure.

Parshall, Davey, and Nering (1998) followed this initial study with an examination of the Davey-Parshall in comparison to both the Sympson-Hetter and the conditional Sympson-Hetter procedures with a 600 item pool using a variable length stopping rule. Results indicated that the Davey-Parshall and Sympson-Hetter both performed similarly with no items in the highest exposure category and a comparable number of items unused. The conditional Sympson-Hetter had the best pool usage, but also produced the most overexposed items. This was attributed to the lack of global exposure control (exposure is only controlled at specific trait levels) in the procedure. The conditional Sympson-Hetter also produced longer test lengths in the tails of the distribution and shorter test lengths in the middle of the distribution when compared to either the Davey-Parshall or the Sympson-Hetter, again as a result of the trait level specific exposure control. The authors conclude that all three procedures performed well. The conditional Sympson-Hetter produced the most uniform test overlap rates, but at the cost of longer test lengths in the tails of the distribution. The Davey-Parshall's test overlap rates were only slightly worse, but required no increase in test length. Both the Davey-Parshall and the conditional Sympson-Hetter generally outperformed the Sympson-Hetter.

The Tri-Conditional Procedure, also called the Hybrid approach (Nering, Davey, & Thompson, 1998) or the Fully Conditional Method (Fan, Thompson, & Davey, 1999), combines elements of the Sympson-Hetter, the conditional Sympson-Hetter, and the Davey-Parshall procedure to condition exposure control on trait level and previously administered items to provide improved exposure control over any single method alone. The procedure works by constructing two separate tables of exposure control parameters through pre-operational simulations. One table provides the n (item) X m (trait level) parameters, while the other is a nXn pairwise matrix (Parshall, Hogarty, & Kromrey, 1999). There is one set of off diagonal parameters for item pairs, but m sets of diagonal parameter values—one for each level of trait (Stocking & Lewis, 2000).

The value of the combined item exposure control parameter is based on the frequency with which an item is selected for a given trait level and the frequency with which it occurs with items that have been previously administered. The average of all pairwise parameters for items previously administered is multiplied by the individual item parameter at the current theta estimate. The Sympson-Hetter, conditional Sympson-Hetter, and Davey-Parshall are all special cases of the Tri-Conditional procedure in which different portions of the matrix are collapsed or set to unity (Parshall, Hogarty, & Kromrey, 1999).

Nering, Davey, and Thompson (1998) {as cited in Stocking & Lewis, 2000} tested the Tri-Conditional procedure against its component procedures and concluded that it gave better results for overlap, pool usage, and exposure, than any of the

component procedures in isolation or when no conditioning was used. However, they observed that as conditioning became more complex there appeared to be diminishing returns on these results. Parshall, Hogarty, and Kromrey (1999) examined the performance of the Tri-Conditional against that of the Sympson-Hetter procedure and the a-Stratified design (discussed in the next section) using a fixed length content constrained CAT with a target exposure rate of 0.20 for the Sympson-Hetter and Tri-conditional procedures. They determined that while the a-Stratified design was more straightforward in its implementation, it performed less well overall in terms of exposure control and that the Tri-Conditional procedure provided the best results when taking reliability and exposure rates into consideration.

Stratification Procedures

While item selection in CAT has traditionally focused on selecting the most informative item at the examinee's current trait estimate, Chang and Ying (1999) challenged this idea by pointing out several inherent problems with maximum information item selection. The degree to which information is maximized by item selection depends on the extent to which estimated trait level approximates true trait level. When estimated and true trait level are close, information about an examinee will be maximized. However, when estimated and true trait level differ greatly, as is usually the case early in the test, an item selected to maximize information at the estimated trait level may provide much less information at the true trait level, and, therefore, inappropriate items for the examinee's true trait level may be administered. Further, Chang and Ying (1999) point directly to the correspondence between item

information and the a-parameter, illustrating that highly informative items are those with higher discrimination parameters and that these items are most likely to be overexposed, while items with lower discrimination parameters may go completely unused. This imbalance in item usage causes concerns for test security (when items are overexposed), economic considerations in item pool development (when items go unused), and efficiency in trait level estimation (when highly discriminating items are "misapplied" early in the testing process). This section will detail the a-Stratified design along with variants and modifications to the procedure.

a-Stratified Design. The a-Stratified design (Chang & Ying, 1996) requires test developers to partition an item pool into k different strata based on the value of the item discrimination parameter. Strata are then arranged in ascending order of discrimination. The test is divided into stages to match the number of strata such that a given number of items are chosen from each stratum, with the lowest discriminating stratum used at the beginning of the test and the highest discriminating stratum used at the end of the test. Within a stratum, the item to be administered is randomly selected from the two items whose difficulty values most closely match the examinee's current trait estimate.

The particular number of strata to be used depends on several factors (Chang & Ying, 1999). The greater the variation in discrimination parameters among items in the item pool, the more strata can be implemented. If however, the *a* parameters are more similar, fewer strata should be used. In addition, item pools with a larger range of difficulties at different levels of discrimination can be partitioned into more

strata than item pools whose range of item difficulties is more restricted across discrimination values. Finally, test length and pool size must be considered. In larger item pools, the number of strata can be set close to the length of the desired test, such that a single item will be drawn from each stratum. The procedure evens out exposure rates, because items with low and high discrimination values have an equal likelihood of being selected (Chang & Ying, 1999; Tang, Jiang, & Chang, 1998).

Hau and Chang (1998) challenge the assumption that in order to obtain balanced pool usage, we must sacrifice measurement precision, stating that the a-Stratified design may even improve our measurement abilities because of the more targeted use of highly discriminating items. They believe that using the a-Stratified design, can improve measurement precision, reduce the overexposure of highly discriminating items, and increase pool utilization. Further, they suggest that the procedure is more effective overall than the Sympson-Hetter at obtaining these goals and is much easier to implement.

Tang, Jiang, and Chang (1998) compared the a-Stratified procedure to the Sympson-Hetter when a target exposure rate of 0.20 was used. They used 300 TOEFL (Test of English as a Foreign Language) items as their pool and looked at two different test lengths (20 and 40 items) and three different levels of strata (4, 8, and 10 strata). They found that the a-Stratified design performed better than the Sympson-Hetter in terms of both pool utilization and exposure rate. The Sympson-Hetter did exhibit better measurement precision and less bias, but the differences between the two procedures were small. Chang and Ying (1999) again compared the a-Stratified

design to the Sympson-Hetter procedure under both Bayesian and maximum information item selection. They concluded that the a-Stratified design exhibited an overlap rate half that of the rate observed under either Sympson-Hetter condition and that the a-Stratified design greatly improved pool utilization.

Hau and Chang (1998) demonstrated that unlike the a-Stratified design, the Sympson-Hetter used high discrimination items early in the test, medium discrimination items late in the test, and, in general, very few low discrimination items and then only late in the test. To further evaluate these contradictory stances in item selection, they compared the a-Stratified design (ascending stratification) to two contrived conditions in which items were grouped into strata based on discrimination, but the order of presentation of the strata were altered. In the descending stratification design, strata were administered in order from most to least discriminating, which would mimic the ordering seen in the Sympson-Hetter procedure. In the nonsystematic stratification design, medium discriminating strata were administered first, followed by the least and then highest discriminating strata. The results showed that the ascending discrimination design produced consistently smaller mean squared errors than either the descending or the non-systematic stratification designs, with the non-systematic design outperforming the descending design. Further, when they examined the performance of the Sympson-Hetter and the a-Stratified design in more realistic contexts of item and pool rotation and retirement, they found that over time, the Sympson-Hetter resulted in more item retirements and a degraded item pool with few highly discriminating items because the items retired

were disproportionately of high discrimination which are difficult and costly to replace.

While these results speak well for the utility of the a-Stratified design in controlling item exposure, the procedure has suffered from some major criticisms. Stocking (1998) points out that correlations between the difficulty and discrimination parameters as well as correlations between the errors of estimation in the item parameters which often occur in IRT measurement, may cause problems for the a-Stratified design. Since item difficulty and discrimination are often positively correlated, the result of pool stratification based on discrimination alone may result in items in the lowest stratum having a wide range of difficulties, while those in the highest stratum have a much more narrow range of difficulties tending toward the more difficult. Tang, Jiang, and Chang (1998) agree that the procedure does not work well when item difficulties are not well distributed with respect to the strata because item selection is based on matching difficulty to the current trait estimation. When the items in a stratum have an insufficient range of difficulty, some items may become overly exposed. Specifically, the lack of available easy items in the more discriminating strata may cause the easiest items in those strata to be overexposed. Stocking (1998) further points out that there is often a correlation between content area and item difficulty, such that one content may be more difficult than another. This would result in contents being poorly distributed across strata as well. Additionally, Stocking criticizes the procedure for the lack of specific guidelines for

determining the number of strata, the number of items to select per stratum, and the number of items to consider in the random item selection within strata, stating that

> the discovery of an optimum design under this approach is accidental and the search for an optimum may be even more time consuming and less certain of success than the development and application of test designs that the Chang and Ying approach was designed to replace. (p. 25)

After evaluating fifteen different stratification options with a 28-item fixed length CAT using a 397 quantitative item pool, Stocking (1998) concluded that that the method was "impractical" due to unacceptably low reliabilities and failure to meet content constraints. Parshall, Hogarty, and Kromrey (1999) also found that in their evaluation, the a-Stratified procedure did not provide effective exposure control even though it had a lower standard error for certain trait levels than the Sympson-Hetter or the Triconditional procedures.

To address some of these criticisms, Chang and his colleagues (Chang, Qian, & Ying, 1999; Leung, Chang, & Hau, 1999; Leung, Chang, & Hau, 2000) have developed several modifications to the procedure . Chang, Qian, and Ying (1999) added b-blocking to the a-Stratified design. The a-Stratified design assumes that the difficulty and discrimination parameters of an item are uncorrelated. As Stocking (1998) points out, this is rarely true and may cause problems for the procedure. The b-blocking method addresses this problem by dividing the item pool into blocks according to difficulty. These blocks are then arranged in ascending order and each block is individually partitioned into the desired number of strata according to discrimination. The items corresponding to each stratum are then pulled from each

block and reassembled by strata. For example, if 4 strata are desired, the item pool is grouped into four item blocks of ascending difficulty. The four easiest items make up one block; the next four easiest items make up the second block; and so on. Within each block, items are ordered according to discrimination. The least discriminating item from each block will then be used to construct the first stratum. The most discriminating item from each block will make up the fourth stratum. This procedure guarantees that there is an equivalent distribution of item difficulties across strata, however, it may result in the distribution of discrimination values within a stratum being more heterogeneous than when b-blocking is not used. While this may make for some overlap in discrimination between strata, on average, the strata should increase in discrimination as the test proceeds. Chang, Qian, and Ying (1999) evaluated this modification in a 360 item GRE Quantitative item pool which had an average correlation between difficulty and discrimination parameters of 0.44. The a-Stratified design with b-blocking demonstrated better measurement precision for low examinees, had lower bias and RMSEs, and had improved pool utilization and overlap rates than did the a-Stratified design without the modification. In addition, the b-blocking method appeared to reduce the overexposure of the low difficulty items in later strata.

Chang and Ying (1999) observed that while the a-Stratified design was successful in reducing exposure rates, it did not provide a mechanism to set a guaranteed maximum exposure rate. They suggested that incorporating the Sympson-Hetter algorithm into the a-Stratified design might serve this purpose while still

allowing for the benefits of improved pool utilization and overlap. Leung, Chang, and Hau (1999) implemented this recommendation in a modified procedure they termed the Enhanced a-Stratified design or a-Stratified design with Sympson-Hetter. Sympson-Hetter item exposure control parameters are set through pre-operational simulation with the item pool partitioned into the desired strata. Observed probabilities of item administration are, therefore, based on the stratified structure of the pool and exposure control parameters are set accordingly.

The results of the study in which Leung, Chang, and Hau (1999) evaluated the enhanced a-Stratified design in a simulated 400 item pool and a real data 252 item pool, indicated that the enhanced a-Stratified design had similar pool utilization to the original a-Stratified method, but was superior in terms of exposure and overlap rates. They also found that the simulations to generate the exposure control parameters took less time with the enhanced a-Stratified than with the Sympson-Hetter. Leung, Chang, and Hau (2000) extended the enhanced procedure even further by incorporating it into the a-Stratified design with b-blocking. In a two study comparison of the a-Stratified, a-Stratified with b-blocking, enhanced a-Stratified, and enhanced a-Stratified with b-blocking to the Sympson-Hetter and maximum information item selection, the authors determined that the enhanced a-Stratified with b-blocking outperformed all other methods in terms of exposure control, pool utilization, and item overlap. However, it also had slightly higher RMSEs and lower reliability than the other procedures.

As computerized adaptive testing moves out of its adolescence, there is a push to incorporate more performance oriented items into large scale CAT programs—mirroring the trend seen in paper and pencil tests. Many new item types are being developed which make use of the visual and auditory features of the computerized testing environment such as the graphical modeling problem type developed by Bennett, Morely, and Quardt (1998) which requires examinees to model mathematical scenarios by plotting points on a graph or the architectural site-design problems on the NCARB (National Council of Architectural Registration Boards) exam which ask examinees to arrange landscape and structures within a space to meet certain criterion (Bejar, 1991). In addition, the advent of complex scoring algorithms allow for online automated scoring of these and other partial credit items such as essay questions, testlets, and constructed response math items (Bejar, 1991; Burstein, Kukich, Wolff, Lu, & Chodorow, 1998; Clauser, Margolis, Clyman, & Ross, 1997).

Polytomous items differ from dichotomous items in that they yield a higher modal level of information across a larger span of the theta scale than dichotomously scored items and tend to have smaller item pools than dichotomously scored items (Koch & Dodd, 1989). The potential size of the item pool for polytomous items will vary from testing program to testing program depending on the item type, the difficulty and cost of writing the items, and the frequency with which item pools may be rotated in or out of use. Koch and Dodd (1989) concluded that it was possible for a polytomous CAT to perform well in terms of measurement precision with item pools as small as 30 items, however these results did not take into account the need

for content balancing or exposure control. The following section presents a review of exposure control research conducted with polytomous IRT models.

Exposure Control Research with Polytomous IRT models. Pastor, Chiang, Dodd, and Yockey (1999), examined the performance of the Sympson-Hetter exposure control algorithm in fairly small (60 and 120) item pools using the partial credit model and concluded that it provided some protection against item exposure with minimal reduction in measurement precision. Davis, Pastor, Dodd, Chiang, and Fitzpatrick (2000) replicated these results with regard to measurement precision, again examining the performance of the Sympson-Hetter with the partial credit model in a more realistic range of conditions which included content balancing, an additional item pool size (240 items), and two levels of test dimensionality. However, the Davis, et al. (2000) study found that the Sympson-Hetter was relatively ineffective in constraining item exposure rates to the desired target value and that observed improvements in exposure rate, item overlap, and pool utilization when the Sympson-Hetter was implemented were modest. The authors concluded that difficulties incurred in the implementation of the Sympson-Hetter mechanism, especially problems with convergence of the exposure control parameters in small item pools, were not outweighed by the observed gains in test security.

Pastor, Dodd, and Chang (in press) examined a broader range of exposure control mechanisms with two pool sizes (60 and 100 items) using the generalized partial credit model. The authors sought to evaluate the a-Stratified design as a more simplistic alternative to the Sympson-Hetter and modified the procedure to work in

the polytomous case by selecting items within strata using maximum information. The study compared the a-Stratified design to both the Sympson-Hetter and the enhanced stratified design. In addition, conditional exposure control was examined using both the conditional Sympson-Hetter and the newly proposed conditional enhanced stratified design, which incorporated the conditional Sympson-Hetter algorithm into the a-Stratified design. In contrast to the previous two studies, the results demonstrated a noticeable decrease in measurement precision as exposure control became more restrictive. The best results for measurement precision were seen with the a-Stratified design with the worst measurement precision occurring in the conditional enhanced stratified design. While the a-Stratified design was able to decrease item exposure and overlap and increase pool utilization over the no exposure control condition, the other exposure control methods were generally superior in terms of the test security variables. As in previous research, though, convergence problems were observed when establishing exposure control parameters for the conditions incorporating the Sympson-Hetter or conditional Sympson-Hetter. The authors concluded that a more simplistic approach to exposure control such as the a-Stratified design would be most appropriate with low or medium stakes tests or when the item pool to test length ratio was small. However, when high stakes testing necessitates tighter control of item exposure, the more restrictive conditional selection procedures should be considered.

Davis and Dodd (2001) evaluated the performance of the Within .10 logits randomization procedure with the partial credit model in a seven item fixed length

CAT using a 149 item pool calibrated with data from administrations of the Medical College Admissions Test (MCAT). The procedure was modified to work with the polytomous case. Rather than forming an item group of all items within .10 logits of the needed item difficulty, maximum information item selection was used to select the two most informative items at each of three points along the theta scale: estimated theta, estimated theta minus 0.10, and estimated theta plus 0.10. This resulted in a group of six passages from which one was randomly selected for administration. Results indicated that while measurement precision was somewhat reduced over the optimal, the procedure was successful in reducing exposure and overlap rates and increasing pool utilization. Somewhat promising was the finding that maximum exposure rates were reasonably controlled even without the use of exposure control parameters.

Statement of Problem

As the demand for performance items grows and partial credit scoring becomes more feasible on the computer, it becomes necessary to evaluate exposure control mechanisms with polytomous item pools. While the research investigating the extent that measurement precision is affected when using such constraints in dichotomous item pools is extensive, only recently have researchers begun to address the effects of exposure control when using polytomous items. The results of research on dichotomous items is not necessarily generalizable to the polytomous case because polytomously scored items yield a higher modal level of information across a larger span of the theta scale than dichotomously scored items (Koch & Dodd, 1989).

Therefore, the negative impact on measurement precision observed in the dichotomous case of administering suboptimal items may not occur. In fact, the Pastor et al. (1999), Davis et al. (2000), and Davis and Dodd (2001) studies all have supported this compensatory view of polytomous items. In addition, polytomous item pools tend to be smaller than dichotomous item pools. Although, Koch and Dodd (1989) concluded that it was possible for a polytomous CAT to perform well with item pools as small as 30 items, this finding did not take into consideration the threat posed to the test's validity if the item pool were to be compromised. The degree to which exposure control mechanisms can be successfully implemented under these more restrictive pool conditions must also be evaluated.

Exposure control research conducted with polytomous IRT models has yielded mixed results both in terms of the effectiveness of various procedures for improving test security and the impact of exposure control on measurement precision. The Pastor et al. (1999) and the Davis et al. (2000) studies both found that the Sympson-Hetter could be used without negative impact to measurement precision, but the Pastor, et al. (in press) study did find a decrease in measurement precision when the Sympson-Hetter was used. The Pastor, et al. (in press) study found that the Sympson-Hetter was effective in controlling item exposure, but the Davis, et al. (2000) study found that it was not.

Conclusions as to the cause of these inconsistencies cannot yet be drawn because of differences in the item pools and test structures studied. One possibility is that results may differ for Rasch and non-Rasch models due to the influence, or lack

thereof, of item discrimination on information, and thus, maximum information item selection. Way (1998) discusses the differential impact of the underlying measurement model on CAT performance with dichotomous items, citing research that demonstrates that Rasch based CATs tend to be robust to modifications in the item pool and item selection algorithms which may cause the administration of sub-optimal items (Haynie & Way, 1994; Way, Zara, & Leahy, 1996). In a Rasch model, when all items are assumed to be of equal discrimination, the substitution of one item for another may make less of an impact on trait estimation because they are all equally informative, only differing in their location along the difficulty scale.

Polytomous IRT models differ, however, not only in terms of their parameterizations, but also in how response probabilities are computed. Both the generalized partial credit and the graded response models allow for variation in item discrimination, but the generalized partial credit model is a divide-by-total model, whereas the graded response model is a difference model. It is reasonable to think that differences between these classes of models may influence the optimal method for controlling item exposure. Since no research has examined the use of exposure control procedures with the graded response model, however, this, along with the potential effectiveness of various exposure control procedures for the model, remains sheer speculation.

Polytomous exposure research to date has been limited in the scope of exposure control procedures studied. Clearly, the smaller nature of polytomous item pools impacts the implementation of conditional selection strategies because of

convergence problems observed in establishing the exposure control parameters. It is also questionable as to whether procedures such as the Sympson-Hetter are even effective in controlling item exposure in polytomous pools (Davis, et al., 2000). It is, therefore, highly desirable to investigate the utility of more simplistic procedures such as the a-Stratified design and various randomization procedures. The results of the Davis and Dodd (2001) study indicate that randomization procedures may, in fact, yield sufficient test security with polytomous item pools, but further research is needed.

Certain randomization procedures, however, such as the randomesque and within .10 logits procedure (in the polytomous modification; Davis & Dodd, 2001) evoke questions as to the appropriate item group size from which to randomly select the next item for administration. These procedures present no clear recommendations in terms of item group size to test developers and it is unclear how changes in this property will affect measurement precision and test security variables.

The current study attempted to address some of these issues by evaluating the performance of a series of exposure control procedures with three different polytomous IRT models. Rasch and non-Rasch models as well as difference and divide-by-total models are represented in the three models chosen for study. The exposure control procedures examined include representatives from all three classifications (randomization, conditional, and stratification). Specific questions which are addressed by this dissertation are:

1) Can simpler methods such as the a-Stratified design or various randomization procedures be as effective (or more so) for protecting test security as conditional selection strategies such as the Sympson-Hetter?
2) To what extent will measurement precision be sacrificed under each exposure control procedure?
3) Does the optimal method for controlling item exposure differ depending on the polytomous model selected?
4) What impact does item group size have on the effectiveness of certain randomization procedures?

Chapter III

Methodology

Overview of Techniques

Six different exposure control mechanisms were evaluated in the current study, two each from the randomization, conditional selection, and stratification procedures— the randomesque procedure, the within .10 logits procedure, the Sympson-Hetter, the conditional Sympson-Hetter, the a-Stratified design, and the enhanced a-Stratified design. In addition, for the randomesque and the within .10 logits procedures, two different item group sizes from which the item to be administered is randomly drawn were evaluated. Finally, a maximum information item selection condition was used for baseline no exposure control comparison, for a total of nine exposure control conditions.

Three different polytomous IRT models were used in the current study—the graded response model, the generalized partial credit model, and the partial credit model. Each model was completely crossed with all of the above outlined exposure control methods except for the partial credit model. Since this model assumes equal discrimination across items, the two conditions involving the a-Stratified design was not evaluated. The result is a partially crossed 9 (exposure control) X 3 (polytomous model) within group factorial design yielding 25 evaluated conditions.

Item pool

Data were obtained from 22 forms of the Verbal Reasoning section of the Medical College Admissions Test (MCAT), collected during six separate administrations occurring from April 1996 through April 2001, to assemble an item pool consisting of 157 polytomously scored items. Due to previously observed problems with low category frequencies in the extremes of the response distribution which caused difficulties with item calibration (Davis & Dodd, 2001), item responses were collapsed across categories and recoded. Items which originally had 7 categories were collapsed to 3 categories. Items which originally had 8 categories were collapsed to 4 categories. Items which originally had 9 or 11 categories were collapsed to 5 categories. The result was an item pool which consisted of 63% 3-category items, 18.5% 4-category items, and 18.5% 5-category items. Original content area classifications which indicated an item's membership in one of three content areas (Humanities, Social Sciences, Natural Sciences) were retained to preserve any naturally existing differences in item difficulty by content area. Of the 157 items, 39% represented the content area of Humanities, 37.5% represented the content area of Social Sciences, and 23.5% represented the content area of Natural Sciences.

Data Generation

Response data to the 157 items was independently generated using conventional techniques for each of the three polytomous IRT models, resulting in separate, model specific datasets. For each model, a random number was drawn from a normal distribution (0,1) to represent the known trait level for a simulee. The

probability of responding in each category given a simulee's trait level was then computed for each item according to the appropriate model (see the Item Response Theory with Polytomous Items section of the Literature Review). These probabilities were then summed to create a cumulative probability of response ranging from 0 to 1. A random number was drawn from a uniform distribution and compared to the cumulative response probability. The simulee was assigned the score which corresponded to the location in the cumulative probability distribution that the random number fell at or below. This procedure was repeated for all simulees and all items.

Four separate data sets were generated for each model for use in the current study, yielding a total of twelve data sets. The first data set for each model (N(0,1); N=7500) was used as the calibration sample to obtain item parameter estimates for use in the CATs. The second data set for each model (N=8000) was used to set exposure control parameters for the Sympson-Hetter and enhanced a-Stratified design. For procedures using the Sympson-Hetter, the exposure control parameters are distributionally dependent, and, therefore, must be set using a sample of simulees whose distribution of trait level will approximate that expected in operational testing. For the current study, a normal N(0,1) distribution was assumed. The third data set for each model (N=15000) was used to set exposure control parameters for the conditional Sympson-Hetter. For the conditional Sympson-Hetter, exposure control parameters are set independent of the trait level distribution, and, therefore, to attain the same level of precision for the exposure control parameters at all levels of theta,

the sample of simulees should be uniformly distributed across the theta levels. For the current study, 1000 simulees were drawn at each of fifteen discrete theta levels in order to set the exposure control parameters for the conditional Sympson-Hetter. The fourth data set for each model (N(0,1); N=1000) was generated for use in each of the CAT conditions. This last data set is generated separately from the calibration data set so as to avoid any possibility of capitalizing on chance by using the same sample in the CATs that was used to estimate the item parameters.

Parameter Estimation

Responses from the three N=7500 calibration samples to the 157 items were separately submitted to PARSCALE (Muraki & Bock, 1993) for calibration according the graded response, generalized partial credit, and partial credit models. PARSCALE employs a marginal maximum likelihood EM algorithm for parameter estimation that consists of two steps: first, the provisional expected frequency and sample size are calculated, and second, the marginal maximum likelihood is estimated. These steps continue through a series of iterations until item parameter estimates stabilize.

Stratification of the Item Pool

For the a-Stratified and enhanced a-Stratified conditions the pool was divided into five strata with 32 items in the first two strata and 31 items in each of the remaining three strata. Yi and Chang (2000) developed a modification to the a-Stratified design in which items are stratified according to multiple factors which might include not only item discrimination, but also item difficulty, and item content

associations. This multiple stratification modification, which they deemed the CBASTR, was designed to control for any variation in item difficulty across content areas which might result in a poor distribution of item contents across the strata.

In the current study, it was found that a moderately strong negative correlation (-0.54) existed between the estimated item discrimination parameter and the number of categories, which resulted in the more discriminating strata containing items with fewer categories and the less discriminating strata containing items with more categories. This caused problems due to the use of a content balancing mechanism which insured that each test administered reflected an appropriate distribution of items both in terms of content and number of categories. It was determined that the CBASTR method of multiple stratification could be adapted to control for this phenomenon. Therefore, before being stratified according to discrimination, items were first stratified according to content area and number of categories. A triple stratification of the pool was implemented so that each stratum would contain a sufficient number of items from each content area and of each number of categories.

Combining the three levels of content with the three levels of categories resulted in nine unique sets of item characteristics. The item pool was separated into nine different subpools representing each of the item types. Each of these nine subpools was then sorted independently by discrimination. The five strata were then formed by pulling items from each of the sorted subpools into the appropriate stratum—the 3 or 4 least discriminating items from each subpool were used to make the 1st stratum; the 3 or 4 highest discriminating of each type to were used to

construct the 5th stratum. This process was implemented independently for the generalized partial credit and graded response models, utilizing the estimated item discrimination values obtained under each model.

Setting the exposure control parameters

Sympson-Hetter and Enhanced a-Stratified Conditions

Responses from the N=8000 data set were used with the estimated item parameters in setting the exposure control (K_i) parameters for the Sympson-Hetter and enhanced a-Stratified conditions. Hetter and Sympson (1997) report that the maximum probability of administration will approach a value slightly above the target exposure rate. While several operational CAT testing programs using dichotomous items have selected a target exposure rate of .20 (Stocking, 1992), the smaller nature of the polytomous item pool in the current study necessitates a more liberal criterion. For the current research, a target exposure rate, r, of 0.39 was therefore set for each of the conditions to ensure that the maximum probability of administration would converge to the 0.40 level. It should be noted that this target exposure rate was even higher than the 0.30 level used in previous research exploring the Sympson-Hetter in polytomous CAT (Pastor, et al., 1999; Davis, et al., 2000; Pastor, et al., in press), but was chosen after initial experimentation with more restrictive levels of exposure control failed to produce convergence of the exposure control parameters.

For the first iteration, all K_i's were set equal to 1.0 so that every item which was selected through maximum information item selection, would actually be

administered. This provided a baseline exposure rate for all items in the pool. After each iteration the probability of selecting each item, P(S), was computed by dividing the frequency of selection by the number of simulees (in this study, 8000). Based on the probability of selection, new K_i's were computed such that if an item's P(S) was less than or equal to the target exposure rate (r=.39), K_i will be set equal to 1.0. Otherwise, if an item's P(S) was greater than our target exposure rate (r=.39), K_i was set equal to r/P(S). To ensure that there would always be at least as many items available for administration as the maximum test length, the K_i for 20 items was automatically set equal to 1.0 after each iteration. These 20 items were identified according to two criteria. First, the 20 items were selected to meet the necessary distribution of content and number of categories stipulated by the content balancing mechanism. Second, within each set of item and content characteristics, items were chosen which had the lowest probability of selection.

The iterations to set the exposure control parameters were conducted in exactly the same fashion for the enhanced a-Stratified design as for the Sympson-Hetter procedure, with the exception that they were computed within the constraints of stratification. In other words, the probability of selection was computed and compared to the target exposure rate, but only those items in the first stratum were available for selection in the first stage of testing, only those items in the second stratum were available for the second stage of testing, etc.

Conditional Sympson-Hetter

Responses from the N=15000 data set were used with the estimated item parameters in setting the exposure control (K_i) parameters for the conditional Sympson-Hetter. This data set was designed to have 1000 simulees at each of 15 levels of theta—from –3.5 to 3.5 logits in 0.5 logit increments. Iterations to set the exposure control parameters for the conditional Sympson-Hetter were conducted in the same fashion as those for the Sympson-Hetter, with the exception that the probability of selection was computed separately for each trait level and a separate exposure control parameter was computed for each item at each trait level. The target exposure rate for each theta level was set to 0.39. All K_is were set equal to 1.0 for the first iteration and new K_is computed at each iteration by comparing the probability of selection to the target exposure rate.

To ensure that there would always be at least as many items available for administration as the maximum test length, it was again necessary to set the K_i values for certain items equal to 1.0. Unlike the procedures used for the Sympson-Hetter and Enhanced a-Stratified conditions, with the conditional Sympson-Hetter, there were 15 separate sets of K_i values—one for each level of theta. It was, therefore, necessary to perform this operation separately for each level of theta. Despite best attempts to incorporate this procedure, exposure control parameters for the conditional Sympson-Hetter would not converge with it in place. Therefore, in the current study, this step was omitted in order to allow for convergence of the exposure control parameters. This is not a recommended course of action for operational

implementation and implications of the decision to proceed without the procedure in place are covered in the Discussion section.

CAT simulations

SAS computer programs originally developed by Hou, Chen, Dodd, and Fitzpatrick (1996) and Chen, Hou, and Dodd (1998) were modified to meet the specifications of each CAT condition in the current research. The initial theta estimate for each simulee was zero in all administrations with the use of variable stepsize to estimate ability until responses were made into two different categories and MLE thereafter. A 20 item fixed length stopping rule was used.

Content and item type were balanced using the Kingsbury and Zara (1989) constrained CAT (CCAT) method for all conditions. Two content factors, unrelated to the psychometric properties of the items, were jointly balanced with this method—content area affiliation and number of categories per item. By combining the three levels of content area affiliation (Humanities, Social Sciences, Natural Sciences) with the three levels of numbers of categories (3,4, or, 5), nine unique sets of item characteristics were produced (i.e. Humanities with 3 categories, Social Sciences with 5 categories, etc). After each item administration, the proportion of each of the nine item types given was computed and compared to the target desired proportion. The next item administered was constrained to be chosen from the area with the largest discrepancy. Target proportions for each of the nine item types were defined to match the observed percentages of each characteristic in the item pool. Item selection is described for each of condition below.

No Exposure Control (Maximum Information)

In the maximum information item selection condition, items were chosen to maximize the information at the current trait estimate.

Randomesque

An item group of the pool's most informative items for a given trait level was assembled and the next item to be administered was randomly chosen from within this item group. In the current study, two item group sizes were evaluated—three and six. So, the next item to be administered was randomly chosen from among the three most informative items or the six most informative items respectively.

Within .10 logits

This study made use of the modified within .10 logits procedure developed by Davis and Dodd (2001) for use in the polytomous case. The procedure was implemented by using maximum information item selection to select the most informative items at each of three points along the trait metric: estimated theta, estimated theta minus 0.10, and estimated theta plus 0.10. The next item to be administered was randomly selected from among this item group. In the current study, two group sizes were evaluated. In the first case, one item was selected at each of these three theta points, resulting in an item group size of three. In the second case, two items was selected at each of these theta points, resulting in an item group size of six.

Sympson-Hetter

For the Sympson-Hetter condition, the appropriate K_i parameters were read in from an external file. An item was selected for administration based both on maximum information and on the comparison of that item's K_i parameter to a random number drawn from a uniform distribution. If K_i was greater than the random number, then the item was administered, otherwise, the item was blocked from further selection for that simulee and the next most informative item was evaluated for administration.

Conditional Sympson-Hetter

For the conditional Sympson-Hetter condition, the appropriate matrix of K_i parameters was read in from an external file. An item was selected for administration based both on maximum information and on the comparison of the K_i parameter corresponding to the theta level closest to the simulee's estimated trait level to a random number drawn from a uniform distribution. If K_i was greater than the random number, then the item was administered, otherwise, the item was blocked from further selection for that simulee and the next most informative item was evaluated for administration.

a-Stratified design

The a-Stratified design was implemented in this study according to the modifications made by Pastor, Dodd, and Chang (in press) for use in the polytomous case. There were five stages of testing, with four items administered from each of the five increasingly discriminating strata. Within a stratum, items were selected by maximum information item selection.

Enhanced a-Stratified design

The enhanced a-Stratified design condition had the same stratification structure (five stages, four items administered per stratum) as the a-Stratified design. However, selection within a stratum was based both on maximum information item selection and the comparison of the item's exposure control parameter to a random uniform number. If K_i was greater than the random number, then the item was administered, otherwise, the item was blocked from further selection for that simulee and the next most informative item was evaluated for administration.

While the triple stratification of the pool by content, number of categories, and discrimination parameter described above in the Stratification of the Item Pool section, ensured that enough items were present within each stratum to meet the content balancing constraints, the inclusion of the Sympson-Hetter exposure control parameters in the enhanced a-Stratified design could result in those items being unavailable for administration. In the event that no item within the current stratum met the desired content and item type constraints and whose comparison between K_i and the random number allowed it to be administered, a method of backward and forward searching through the strata was implemented (Leung, Chang, & Hau, 2000; Leung, 2001). This method allowed items from other strata to be considered for administration when no item from the current stratum could be administered. Items from previous strata were first considered in a step down fashion from the current stratum with items from successive strata considered only if no item from either the current or all previous strata could be administered. For example, if the current

stratum was stratum 3, the order in which items from other strata would be considered would be stratum 2, stratum 1, stratum 4, and finally stratum 5.

Data Analyses

In order to evaluate the recovery of known theta in each condition, several variables were used. In addition to descriptive statistics, the Pearson product-moment (PPM) correlation coefficients were calculated between the known and estimated theta values. Bias, root mean squared error (RMSE), standardized difference between means (SDM), standardized root mean squared difference (SRMSD), and average absolute difference (AAD) statistics were also calculated. The equations to compute these statistics are as follows:

$$Bias = \frac{\sum_{k=1}^{n}(\hat{\theta}_k - \theta_k)}{n}, \tag{20}$$

$$RMSE = \left[\frac{\sum_{k=1}^{n}(\hat{\theta}_k - \theta_k)^2}{n}\right]^{1/2}, \tag{21}$$

$$SDM = \frac{\bar{\hat{\theta}} - \bar{\theta}}{\sqrt{\frac{s^2\hat{\theta} + s^2\theta}{2}}}, \tag{22}$$

$$SRMSD = \sqrt{\frac{\frac{1}{n}\sum_{k=1}^{n}(\hat{\theta}_k - \theta_k)^2}{\frac{s^2\hat{\theta} + s^2\theta}{2}}}, \text{ and} \tag{23}$$

$$AAD = \frac{\sum_{i=1}^{n}\left|\hat{\theta}_k - \theta_k\right|}{n}, \tag{24}$$

where $\hat{\theta}_k$ is the estimate of trait level for simulee k, θ_k is the known trait level for simulee k, $\bar{\theta}$ is the mean of the known trait levels , $\bar{\hat{\theta}}$ is the mean of the estimated trait levels, $s^2\theta$ was the variance of known trait levels, $s^2\hat{\theta}$ is the variance of estimated trait levels, and n is the total number of simulees.

Item exposure rates (the probability of administering an item) were computed by dividing the number of times an item was administered by the total number of simulees. Frequency distributions of the exposure rates, along with average and maximum exposure rates were examined across conditions. The percent of items that were never administered was used as an index of pool utilization.

In order to measure test overlap, the audit trails of each simulee were compared to the audit trails of every other simulee. A data file containing the number of items shared among the simulees as well as the difference between their known theta values was created to obtain an index of item overlap conditional on theta. Simulees were defined to have "similar" trait levels when their known thetas differed

by two logits or fewer and "different" trait levels when their known thetas differed by more than two logits (Pastor, Chiang, Dodd, & Yockey, 1999; Davis, Pastor, Dodd, Chiang, & Fitzpatrick, 2000; Pastor, Dodd, & Chang, in press).

Chapter IV

Results

This chapter first presents the results for those conditions employing the partial credit model. There were only seven conditions using this model because the assumption of equal discrimination across items prohibited the use of the two exposure control procedures based on the a-Stratified design. Results for the generalized partial credit and graded response models are discussed in separate sections following those for the partial credit model.

Partial Credit Model

This section presents the results for the seven conditions using the partial credit model. Descriptive statistics for the item pool, nonconvergent cases, average theta estimate, and average standard error for each exposure control condition are presented first. In addition, values are provided for the correlation between known and estimated theta, bias, SDM, RMSE, SRMSD, and AAD. These values provide an indication of how well each exposure control condition performed in terms of measurement precision. The results of attempts to set exposure control parameters for the Sympson-Hetter and conditional Sympson-Hetter conditions are discussed next. This provides an indication of the ease or difficulty of implementation for those procedures utilizing individually assigned exposure control parameters. Finally, the degree of pool utilization, item exposure, and item overlap under each condition are presented. These values represent the test security variables and indicate to what

degree an exposure control mechanism was able to control exposure and balance item usage.

Descriptive Statistics

Table 1 lists the mean, standard deviation, minimum, and maximum values of the item parameter estimates for all 157 items in the pool. The majority of items (99 out of 157) had only two step values, therefore the information for the third and fourth step values is only presented for those items for which it is relevant. A plot of test information for the item pool calibrated according to the partial credit model is presented in Figure 1. The item pool information peaks at a theta value of –0.5.

After all conditions had been run, a listwise deletion of 53 nonconvergent cases was performed. A case was defined as nonconvergent if, once the end of the test had been reached, the trait estimate was greater than or equal to 4.0 or less than or equal to –4.0, or if maximum likelihood estimation had never been reached. For the partial credit model, all 53 cases were nonconvergent due to the trait estimates being too extreme ($\hat{\theta} \leq -4.0; \hat{\theta} \geq 4.0$). The number of nonconvergent cases for each condition is listed in Table 2. The number of nonconvergent cases was fairly consistent across conditions (7 to 9 cases) with the exception of the randomesque-6 and within .10 logits-6 conditions where the number of nonconvergent cases was 20 and 19 respectively. The remainder of the results for the partial credit model conditions are reported on the sample (N=947) of observations which remained after the nonconvergent cases had been deleted.

Table 1

Mean, Standard Deviation, Minimum, and Maximum for Item Parameter Estimates Obtained Under the Partial Credit Model

Partial Credit Model					
Item Parameter	***N***	***Mean***	***St. Dev.***	***Min.***	***Max.***
Step Value 1	157	-0.9639	0.7752	-2.3917	0.9522
Step Value 2	157	0.1703	0.8953	-1.8446	2.7836
Step Value 3	58	-1.0906	0.6738	-1.3649	1.4015
Step Value 4	29	0.0053	0.7242	-1.3548	1.8538

Figure 1: Test Information Function for N=157 Items Under the Partial Credit Model

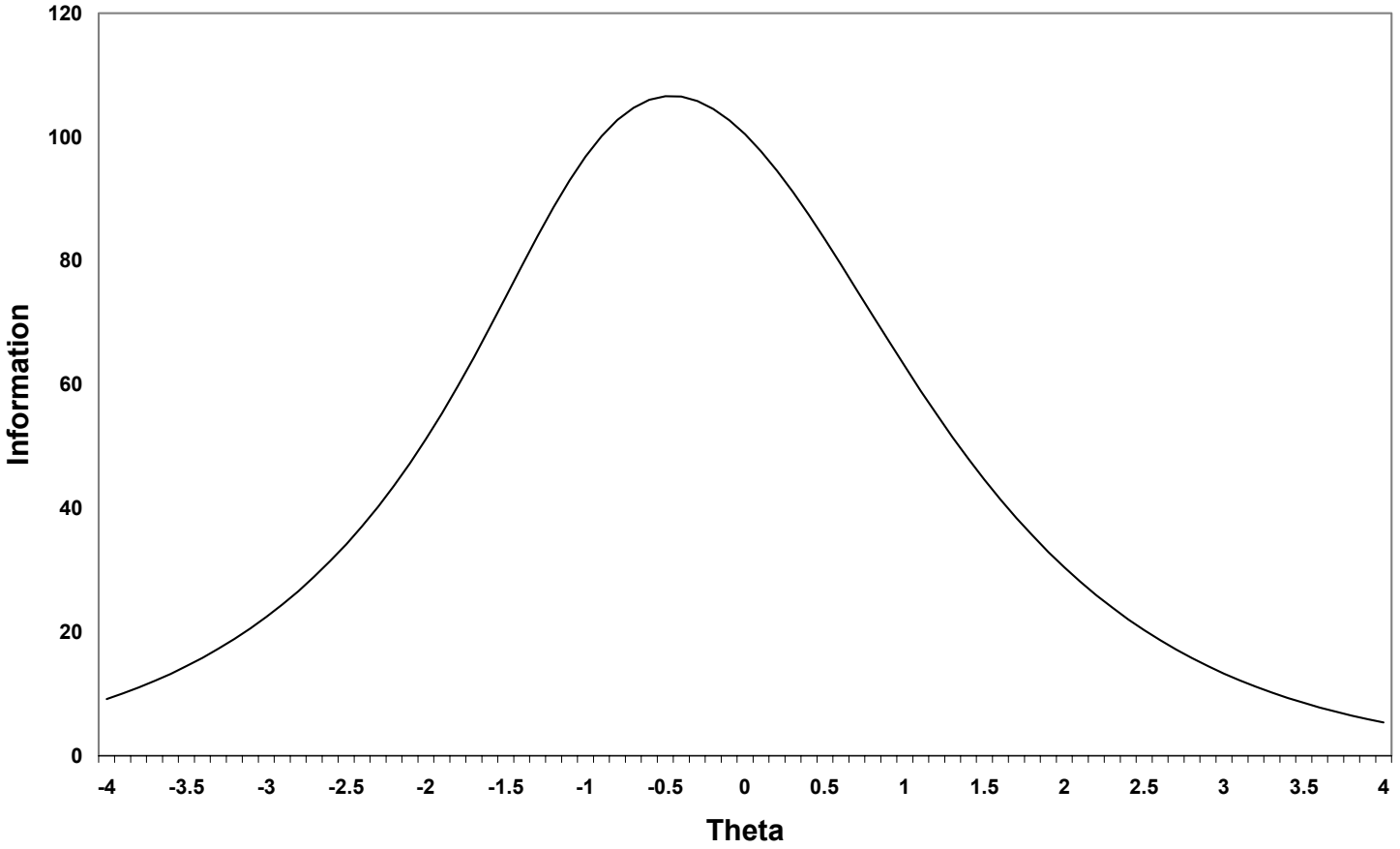


Table 2

Means (and Standard Deviations) for Estimated Theta and Standard Error; Number of Nonconvergent Cases for the Exposure Control Conditions Using the Partial Credit Model

Partial Credit Model			
Exposure Control Condition	***Theta* Estimate***	***Standard Error***	***Nonconvergent Cases***
No Exposure Control	0.01 (1.04)	0.27 (0.04)	9
Randomesque-3	0.01 (1.05)	0.28 (0.05)	8
Randomesque-6	-0.01 (1.03)	0.29 (0.05)	20
Within .10 Logits-3	0.01 (1.05)	0.28 (0.05)	7
Within .10 Logits-6	-0.01 (1.06)	0.30 (0.05)	19
Sympson-Hetter	0.01 (1.04)	0.27 (0.04)	9
Conditional Sympson-Hetter	0.00 (1.04)	0.28 (0.05)	8

*Note: Mean and SD for Known Thetas (N=947) were Mean=-0.04; SD=1.00

Table 2 also contains the average theta estimate and standard error for each condition. The mean of the known thetas was –0.04 with a standard deviation of 1.00. All conditions reported relatively similar results in terms of the theta estimate, with values for the average theta estimate ranging from –0.01 to 0.01 and the standard deviation of the theta estimate ranging from 1.03 to 1.06. As expected, the no exposure control condition yielded the lowest average standard error, 0.27 ,with the randomesque-6 and the within .10 logits-6 yielding the highest average standard errors of 0.29 and 0.30 respectively. Surprisingly, the Sympson-Hetter produced an average standard error, 0.27, comparable to that of the no exposure control condition.

Table 3 presents the correlations between known and estimated theta for each condition as well as statistics for bias, SDM, RMSE, SRMSD, and AAD. All conditions resulted in an identical correlation coefficient between known and estimated theta, 0.96, and were also very similar for the other statistics. Bias and SDM were both functionally zero for all conditions. RMSE was 0.29 to 0.30 for all conditions. SRMSD ranged from 0.52 to 0.53 for all conditions. Finally, AAD ranged from 0.22 to 0.24 for all conditions. Any variation in the value of these statistics across conditions was judged too small to be reliably interpreted.

Exposure Control Parameters

For the Sympson-Hetter, a target exposure rate of 0.39 was chosen after initial attempts to set the exposure rate to more restrictive levels (0.19 and 0.29) were unsuccessful. Thirty iterations were run to ensure converge to the desired level. The maximum probability of administration after the 30th iteration was 0.398. Conditional

Table 3

Correlation Coefficients between Known and Estimated Theta, Bias, RMSE, SDM, SRMSD, and AAD for the Exposure Control Conditions Using the Partial Credit Model

Partial Credit Model						
Exposure Control Condition	***Correlation***	***Bias***	***SDM***	***RMSE***	***SRMSD***	***AAD***
No Exposure Control	0.96	-0.05	0.05	0.29	0.53	0.23
Randomesque-3	0.96	-0.05	0.05	0.30	0.53	0.23
Randomesque-6	0.96	-0.03	0.03	0.29	0.53	0.23
Within .10 Logits-3	0.96	-0.05	0.05	0.30	0.53	0.23
Within .10 Logits-6	0.96	-0.03	0.03	0.30	0.53	0.24
Sympson-Hetter	0.96	-0.06	0.06	0.29	0.52	0.22
Conditional Sympson-Hetter	0.96	-0.04	0.04	0.29	0.53	0.23

maximum exposure rates of 0.39 were chosen for use with the conditional Sympson-Hetter and 30 iterations were conducted to ensure convergence of the parameters to this level. Previous research (Stocking & Lewis, 1998) has demonstrated that the conditional maximum probability of administrations consistently converged to values between 0.05 and 0.10 above the target value. Consistent with this finding, the maximum probability of administration at each theta level in the current study converged to values slightly greater than the target (0.426-0.518).

Pool Utilization and Exposure Rates

Table 4 contains the frequency of observed exposure rates along with the average, maximum, and standard deviation of exposure rates, and the percent of pool not administered for each condition. Chen, Ankenmann, & Spray (1999) state that the average exposure rate for any fixed length test will always be constant and mathematically equal to the ratio of test length to pool size. Since test length was the same for all conditions studied, the observed average exposure rates, therefore, did not differ and were equal to 0.127 across all conditions. The standard deviation of exposure rates was highest for the no exposure control condition (0.167), indicating the most uneven item exposure rates, and lowest for the randomesque-6 and within .10 logits-6 conditions (0.095 and 0.096 respectively), indicating the most even item exposure rates. The maximum exposure rate of any item was largest for the no exposure control condition with a value of 0.655 and lowest for the randomesque-6, within .10 logits-6, and conditional Sympson-Hetter conditions with values of 0.396, 0.398, and 0.395 respectively. The percent of pool not administered

Table 4

Pool Utilization and Exposure Rates for the Exposure Control Conditions Using the Partial Credit Model

Partial Credit Model							
Exposure Control Condition	No Exposure Control	Randomesque 3	Randomesque 6	Within .10 Logits 3	Within .10 Logits 6	Sympson-Hetter	Conditional Sympson-Hetter
Number of Items	157	157	157	157	157	157	157
Exposure Rate							
1	0	0	0	0	0	0	0
.91-.99	0	0	0	0	0	0	0
.81-.90	0	0	0	0	0	0	0
.71-.80	0	0	0	0	0	0	0
.61-.70	2	0	0	0	0	0	0
.51-.60	4	2	0	2	0	0	0
.41-.50	12	5	0	7	0	8	0
.36-.40	3	4	2	3	1	13	3
.31-.35	5	4	8	5	8	9	10
.26-.30	5	13	7	11	10	9	14
.21-.25	13	10	18	13	19	8	18
.16-.20	9	17	27	15	25	10	18
.11-.15	11	24	29	25	28	10	17
.06-.10	7	26	27	22	27	12	26
.01-.05	28	20	26	22	26	30	28
Not Administered	58	32	13	32	13	48	23
Exposure Rate AVG	0.127	0.127	0.127	0.127	0.127	0.127	0.127
Exposure Rate SD	0.167	0.123	0.095	0.126	0.096	0.147	0.11
Exposure Rate MAX	0.655	0.503	0.396	0.535	0.398	0.434	0.395
% of Pool Not Administered	37%	20%	8%	20%	8%	31%	15%

was greatest for the no exposure control condition where 37% of the pool was never used. The percent of pool not administered was lowest for the randomesque-6 and within .10 logits-6 conditions where only 8% of the pool was never used.

For all exposure rate and pool utilization variables, the randomesque-3 and within .10 logits-3 procedures demonstrated improvement over the no exposure control condition, but did not perform as well as their 6 item group counterparts. While the Sympson-Hetter and conditional Sympson-Hetter controlled maximum exposure rates to levels close to the target value (0.434 and 0.395 respectively), the percent of pool not administered under each procedure (31% and 15% respectively) was well above the randomesque-6 and within .10 logits-6 procedures.

Item Overlap

Audit trails for each simulee were compared to the audit trails of every other simulee resulting in 447,931 pairwise comparisons per condition. Table 5 contains the average item overlap for all simulees, those of different trait levels (known thetas differed by more than two logits), and those of similar trait levels (known thetas differed by two logits or fewer) for each condition. Item overlap information in Table 5 is presented both in terms of average number of items shared across a 20 item test and average percent of items shared across a 20 item test.

No exposure control produced the highest overall overlap rates with an average of 34% overlap and the randomesque-6 and within .10 logits-6 procedures result in the lowest overall overlap rates with an average of 20% overlap for both. Results for examinees of similar trait level demonstrate the same pattern with highest

Table 5

Item Overlap Rates for the Exposure Control Conditions Using the Partial Credit Model

Partial Credit Model			
Exposure Control Condition	***Overall Average Overlap*** (N=447,931)	***Different Abilities Average Overlap*** (N=69,084)	***Similar Abilities Average Overlap*** (N=378,847)
No Exposure Control	6.87 34%	1.24 6%	7.89 39%
Randomesque-3	4.89 24%	1.21 6%	5.56 28%
Randomesque-6	3.93 20%	1.65 8%	4.35 22%
Within .10 Logits-3	4.99 25%	1.30 6%	5.66 28%
Within .10 Logits-6	3.96 20%	1.65 8%	4.38 22%
Sympson-Hetter	5.89 29%	1.04 5%	6.77 34%
Conditional Sympson-Hetter	4.43 22%	1.01 5%	5.05 25%

overlap rates (39%) occurring with the no exposure control condition and lowest overlap rates (22%) occurring with the randomesque-6 and within .10 logits-6 conditions. Results for examinees of different trait levels show a different pattern with the randomesque-6 and within .10 logits-6 actually increasing overlap slightly (8%) from the no exposure control condition (6%). However, these overlap rates are uniformly so small that too much emphasis should not be placed on them.

Again, the randomesque-3 and within .10 logits-3 conditions do reduce overlap when compared to the no exposure control condition (24% and 25% overall overlap respectively), but the reduction is not as great as that seen with their 6 item group counterparts. The conditional Sympson-Hetter does a reasonable job of reducing overlap (22% overall), but the Sympson-Hetter maintains the second highest overlap rate of all conditions with 29% overall overlap.

Generalized Partial Credit Model

This section discusses the results for the nine conditions using the generalized partial credit model. Results with regard to measurement precision are presented first. Descriptive statistics are given for the item pool, number of nonconvergent cases, average theta estimate, and average standard error. In addition, values are provided for the correlation between known and estimated theta, bias, SDM, RMSE, SRMSD, and AAD. This is followed by a discussion of attempts to set exposure control parameters for the Sympson-Hetter and conditional Sympson-Hetter conditions. This speaks to the ease or difficulty of implementation of the conditional selection strategies. Finally, results with regard to test security are presented in terms

of the degree of pool utilization, item exposure, and item overlap observed under each condition.

Descriptive Statistics

Table 6 lists the mean, standard deviation, minimum, and maximum values of the item discrimination and step values for all 157 items in the pool. Most of the items (99 out of 157) had only two step values, therefore, the information for the third, and fourth step values is only given for the remaining 58 and 29 items respectively. The test information function for the item pool calibrated according to the generalized partial credit model is plotted in Figure 2. The item pool information peaked at a theta value of –0.6.

Table 7 gives the mean, standard deviation, minimum, and maximum values for the item discrimination parameters broken out by strata. Due to implementation of the Yi and Chang (2000) CBASTR modifications to the a-Stratified design, which stratified the item pool by content and number of categories as well as discrimination, item discrimination values may overlap across strata. Yi and Chang (2000) state that this overlap is acceptable so long as the average item discrimination increases across strata. As can be seen from the table, while there is some degree of overlap across the strata, the average 'a'-value does increase as the strata increases.

A listwise deletion of 191 nonconvergent cases was conducted after all conditions had been run. A case was defined as nonconvergent if, once the end of the test had been reached, the trait estimate was greater than or equal to 4.0 or less than or equal to –4.0, or if maximum likelihood estimation had never been reached. For the

Table 6

Mean, Standard Deviation, Minimum, and Maximum for Item Parameter Estimates Obtained Under the Generalized Partial Credit Model

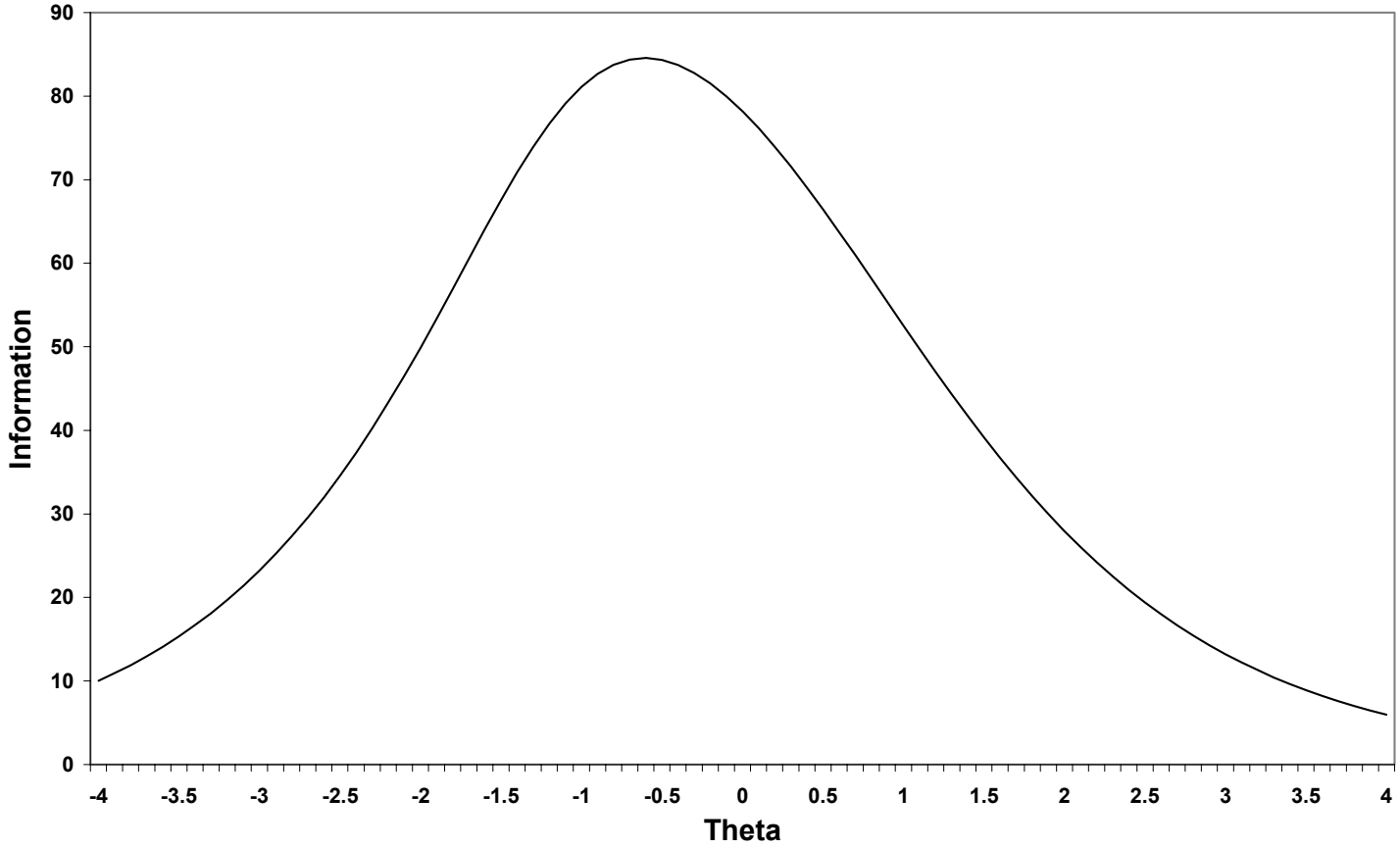
Generalized Partial Credit Model					
Item Parameter	***N***	***Mean***	***St. Dev.***	***Min.***	***Max.***
Item Discrimination	157	0.9159	0.1919	0.5377	1.519
Step Value 1	157	-0.9940	0.9028	-3.1316	1.5021
Step Value 2	157	0.1790	0.9906	-1.8144	3.5687
Step Value 3	58	-0.1949	0.7613	-1.4769	1.5110
Step Value 4	29	-0.1169	0.8973	-2.3620	2.3416

Table 7

Mean, Standard Deviation, Minimum, and Maximum of Item Discrimination Parameters Across Strata for the Generalized Partial Credit Model

Generalized Partial Credit Model					
Strata	***N***	***Mean***	***St. Dev.***	***Min.***	***Max.***
Stratum 1	32	0.73	0.09	0.54	0.87
Stratum 2	32	0.82	0.11	0.63	0.96
Stratum 3	31	0.91	0.13	0.68	1.05
Stratum 4	31	0.98	0.13	0.69	1.17
Stratum 5	31	1.14	0.19	0.80	1.52

Figure 2: Test Information Function for N=157 items Under the Generalized Partial Credit Model



generalized partial credit model, all 191 cases were nonconvergent due to the trait estimates being too extreme ($\hat{\theta} \leq -4.0; \hat{\theta} \geq 4.0$). Table 8 lists the number of nonconvergent cases for each condition. The number of nonconvergent cases was somewhat high, ranging from 34 to 47 across conditions, with the exception of the a-Stratified and enhanced a-Stratified conditions where the number of nonconvergent cases was 2 and 10 respectively. The remainder of the results for the generalized partial credit conditions are reported on the sample (N=809) of observations which remained after the nonconvergent cases had been deleted.

Table 8 also provides the average theta estimate and standard error for each exposure control condition. The mean of the known thetas was 0.03 with a standard deviation of 1.03. All conditions reported a slightly increased average theta estimate, with no exposure control giving the average theta estimate closest to the known (0.05) and randomesque-6 and within .10 logits-6 producing average theta estimates furthest from the known (0.08 and 0.09 respectively). Standard deviations of the theta estimates were also slightly inflated relative to the standard deviation of the known thetas and ranged from 1.09 to 1.12. As expected, no exposure control yielded the lowest average standard error, 0.28, with the a-Stratified and enhanced a-Stratified yielding the highest average standard errors (0.33 for both procedures). The randomesque-3, within .10 logits-3, and Sympson-Hetter all yielded average standard errors of 0.29, while the randomesque-6, within .10 logits-6, and conditional

Table 8

Means (and Standard Deviations) for Estimated Theta and Standard Error; Number of Nonconvergent Cases for the Exposure Control Conditions Using the Generalized Partial Credit Model

Generalized Partial Credit Model			
Exposure Control Condition	***Theta* Estimate***	***Standard Error***	***Nonconvergent Cases***
No Exposure Control	0.05 (1.10)	0.28 (0.05)	34
Randomesque-3	0.07 (1.10)	0.29 (0.05)	41
Randomesque-6	0.08 (1.09)	0.31 (0.06)	44
Within .10 Logits-3	0.07 (1.09)	0.29 (0.05)	39
Within .10 Logits-6	0.09 (1.10)	0.31 (0.06)	46
Sympson-Hetter	0.06 (1.10)	0.29 (0.05)	34
Conditional Sympson-Hetter	0.07 (1.12)	0.31 (0.06)	47
A-Stratified	0.06 (1.11)	0.33 (0.06)	2
Enhanced A-Stratified	0.07 (1.10)	0.33 (0.06)	10

*Note: Mean and SD for Known Thetas (N=809) were Mean=0.03; SD=1.03

Sympson-Hetter yielded average standard errors of 0.31. Table 9 states the correlations between known and estimated theta for each exposure control condition as well as values for the bias, SDM, RMSE, SRMSD, and AAD statistics. The correlation between known and estimated theta was highest for no exposure control (0.96) and lowest for the conditional Sympson-Hetter (0.93). Both the within .10 logits-3 and the Sympson-Hetter conditions had correlation coefficients comparable to the no exposure control condition. Bias and SDM were functionally zero across conditions. No exposure control shared the lowest RMSE, SRMSD, and AAD values with the Sympson-Hetter (RMSE=.29; SRMSD=0.51; AAD=0.23). Within .10 logits-6 and the conditional Sympson-Hetter conditions had the highest RMSEs and SRMSDs (RMSE=0.39, 0.40 respectively; SRMSD=0.59 for both). While both procedures also had high AAD values (0.26), the a-Stratified design had the highest value for AAD at 0.27.

Exposure Control Parameters

Initial attempts were made to set the target exposure rate for the Sympson-Hetter to more restrictive levels (0.19 and 0.29), but after 30 iterations the observed maximum probability of administration did not approach the desired level. A target exposure rate of 0.39 was, therefore, chosen to ensure converge to the 0.40 level. The maximum probability of administration after the 30th iteration with a target exposure rate of 0.39 was 0.404. A target exposure rate of 0.39 was also set for the enhanced a-Stratified condition which yielded a maximum observed probability of administration of 0.399 after the 30th iteration. For the conditional Sympson-Hetter,

Table 9

Correlation Coefficients between Known and Estimated Theta, Bias, RMSE, SDM, SRMSD, and AAD for the Exposure Control Conditions Using the Generalized Partial Credit Model

Generalized Partial Credit Model						
Exposure Control Condition	***Correlation***	***Bias***	***SDM***	***RMSE***	***SRMSD***	***AAD***
No Exposure Control	0.96	-0.03	0.03	0.29	0.51	0.23
Randomesque-3	0.95	-0.05	0.04	0.36	0.56	0.25
Randomesque-6	0.95	-0.06	0.05	0.35	0.56	0.26
Within .10 Logits-3	0.96	-0.04	0.04	0.30	0.52	0.24
Within .10 Logits-6	0.94	-0.06	0.06	0.39	0.59	0.26
Sympson-Hetter	0.96	-0.03	0.03	0.29	0.51	0.23
Conditional Sympson-Hetter	0.93	-0.05	0.04	0.40	0.59	0.26
A-Stratified	0.95	-0.03	0.03	0.34	0.55	0.27
Enhanced A-Stratified	0.95	-0.04	0.04	0.34	0.55	0.26

conditional maximum exposure rates of 0.39 were chosen and 30 iterations were conducted to ensure convergence of the parameters to this level. Stocking and Lewis (1998) reported that the conditional maximum probability of administrations consistently converged to between 0.05 and 0.10 above the target value. In the current study, the maximum probability of administration at each theta level converged to values slightly greater than the target (0.420-0.498).

Pool Utilization and Exposure Rates

The frequency of observed exposure rates along with the average, maximum, and standard deviation of exposure rates, and the percent of pool not administered for each condition are presented in Table 10. Given that test length was the same for all conditions studied, the observed average exposure rates did not differ was equal to the test length (20) divided by the item pool size (157), or 0.127 across all conditions (Chen, Ankenmann, & Spray, 1999). The no exposure control condition demonstrated the highest standard deviation of exposure rates (0.206), signifying the most uneven item usage. The lowest values for standard deviation of exposure rates were those for the randomesque-6, within .10 logits-6, and conditional Sympson-Hetter conditions (0.12 for all three conditions), signifying the most even item usage. The maximum exposure rate of any item was largest for the no exposure control condition with a value of 0.878 and lowest for conditional Sympson-Hetter with a value of 0.407. The Sympson-Hetter, enhanced a-Stratified design, randomesque-6, and within .10 logits-6 also had relatively low maximum exposure rates (0.419, 0.424, 0.477, and 0.482 respectively). The randomesque-3, within .10 logits-3, and

Table 10

Pool Utilization and Exposure Rates for the Exposure Control Conditions Using the Generalized Partial Credit Model

Generalized Partial Credit Model									
Exposure Control Conditions	No Exposure Control	Randomesque 3	Randomesque 6	Within .10 Logits-3	Within .10 Logits-6	Sympson Hetter	Conditional Sympson-Hetter	A-Stratified	Enhanced A-Stratified
Number of Items	157	157	157	157	157	157	157	157	157
Exposure Rate									
1	0	0	0	0	0	0	0	0	0
.91-.99	0	0	0	0	0	0	0	0	0
.81-.90	3	0	0	0	0	0	0	0	0
.71-.80	0	1	0	1	0	0	0	1	0
.61-.70	5	3	0	3	0	0	0	3	0
.51-.60	6	3	0	3	0	0	0	5	0
.41-.50	4	6	3	8	4	8	2	9	8
.36-.40	5	3	9	1	8	24	5	8	21
.31-.35	3	2	8	4	6	5	14	7	9
.26-.30	6	10	6	8	9	8	11	4	4
.21-.25	4	13	8	12	7	3	13	6	6
.16-.20	7	11	19	12	19	6	15	5	7
.11-.15	12	15	29	16	31	8	15	9	9
.06-.10	4	17	25	14	23	9	22	6	8
.01-.05	32	32	29	34	28	35	38	9	15
Not Administered	66	41	21	41	22	51	22	85	70
Exposure Rate AVG	0.127	0.127	0.127	0.127	0.127	0.127	0.127	0.127	0.127
Exposure Rate SD	0.206	0.161	0.12	0.162	0.12	0.158	0.12	0.188	0.158
Exposure Rate MAX	0.878	0.705	0.477	0.713	0.482	0.419	0.407	0.735	0.424
% of Pool Not Administered	42%	26%	13%	26%	14%	32%	14%	54%	45%

a-Stratified design, on the other hand, resulted in maximum exposure rates above acceptable levels (0.705, 0.713, and 0.735 respectively).

Surprisingly, the percent of pool not administered was greater for the a-Stratified design (54%) and enhanced a-Stratified design (45%) than for the no exposure control condition (42%). This is especially troubling given that one of the stated strengths of the a-Stratified design is to increase pool utilization. Possible reasons for this unexpected finding are presented in the Discussion. The percent of pool not administered was lowest for the randomesque-6, within .10 logits-6, and conditional Sympson-Hetter conditions where only 13-14% of the pool was never used.

While the randomesque-3 and within .10 logits-3 procedures did show improvement over the no exposure control condition, they did not perform as well as their 6 item group counterparts when all exposure rate and pool utilization variables were evaluated. The Sympson-Hetter controlled the maximum exposure rate to a level close to the target value, but the percent of pool not administered under this procedure (32%) was well above that observed with other procedures.

Item Overlap

Pairwise comparisons of examinee audit trails were made for each simulee with every other simulee, resulting in 326,836 pairwise comparisons per condition. Table 11 contains the average item overlap for all simulees, those of different trait levels (known thetas differed by more than two logits), and those of similar trait levels (known thetas differed by two logits or fewer) for each condition. Information

Table 11

Item Overlap Rates for the Exposure Control Conditions Using the Generalized Partial Credit Model

Generalized Partial Credit Model			
Exposure Control Conditions	***Overall Average Overlap*** (N=326,836)	***Different Abilities Average Overlap*** (N=56,306)	***Similar Abilities Average Overlap*** (N=270,530)
No Exposure Control	9.18 46%	2.43 12%	10.58 53%
Randomesque-3	6.56 33%	2.27 11%	7.46 37%
Randomesque-6	4.78 24%	2.39 12%	5.27 26%
Within .10 Logits-3	6.61 33%	2.26 11%	7.51 38%
Within .10 Logits-6	4.79 24%	2.48 12%	5.27 26%
Sympson-Hetter	6.43 32%	1.44 7%	7.47 37%
Conditional Sympson-Hetter	4.77 24%	1.97 10%	5.36 27%
A-Stratified	8.04 40%	3.01 15%	9.09 45%
Enhanced A-Stratified	6.40 32%	3.21 16%	7.07 35%

regarding item overlap information is presented both in terms of average number of items shared across a 20 item test and average percent of items shared across a 20 item test.

The highest overall overlap rate was seen with the no exposure control condition produced which yielded an average of 46% overlap. The randomesque-6, within .10 logits-6, and conditional Sympson-Hetter procedures resulted in the lowest overall overlap rates, with an average of 24% overlap for all three procedures. The a-Stratified design yielded an unsatisfactory 40% overall overlap, with the randomesque-3, within .10 logits-3, Sympson-Hetter, and enhanced a-Stratified performing only slightly better with overall overlap rates between 32-33%. Results for examinees of similar trait level demonstrated the same pattern with highest overlap rates (53%) occurring with the no exposure control condition and lowest overlap rates (26-27%) occurring with the randomesque-6, within .10 logits-6, and conditional Sympson-Hetter conditions. Results for examinees of different trait levels show a different pattern with no exposure control, and both randomesque and within .10 logits procedures resulting in about 11-12% overlap. The Sympson-Hetter and conditional Sympson-Hetter produce slightly lower overlap values (7% and 10% respectively) and the a-Stratified and enhanced a-Stratified designs yield slightly higher values (15% and 16% respectively). It is unclear, however, exactly how much emphasis should be placed on the findings for examinees of different trait levels as the overlap rates are fairly small and the interest in reducing overlap lies mainly with examinees of similar trait level.

Graded Response Model

This section presents the results for the nine conditions employing the graded response model model. Results with regard to measurement precision, ease of implementation, and test security are provided. First, descriptive statistics are given for the item pool, nonconvergent cases, average theta estimate, and average standard error. Second, the correlation between known and estimated theta, bias, SDM, RMSE, SRMSD, and AAD are presented. Third is a discussion of attempts to set exposure control parameters for the Sympson-Hetter, enhanced a-Stratified, and conditional Sympson-Hetter conditions. Finally, the degree of pool utilization, item exposure, and item overlap under each condition are discussed.

Descriptive Statistics

Table 12 lists the mean, standard deviation, minimum, and maximum values of the item discrimination and category boundaries for all 157 items in the pool. The bulk of items (99 out of 157) had only two category boundaries, therefore the information for the third and fourth category boundaries is only presented on those items for which it is relevant. Figure 3 shows a plot of the test information function for the item pool calibrated according to the graded response model. The item pool information peaked at a theta value of –0.5. The mean, standard deviation, minimum, and maximum values for the item discrimination parameters for each of the five strata are presented in Table 13. The implementation of the CBASTR multiple stratification process to the a-Stratified design (Yi & Chang, 2000), might be expected to cause values across strata to overlap. As can be seen from the table,

Table 12

Mean, Standard Deviation, Minimum, and Maximum for Item Parameter Estimates Obtained Under the Graded Response Model

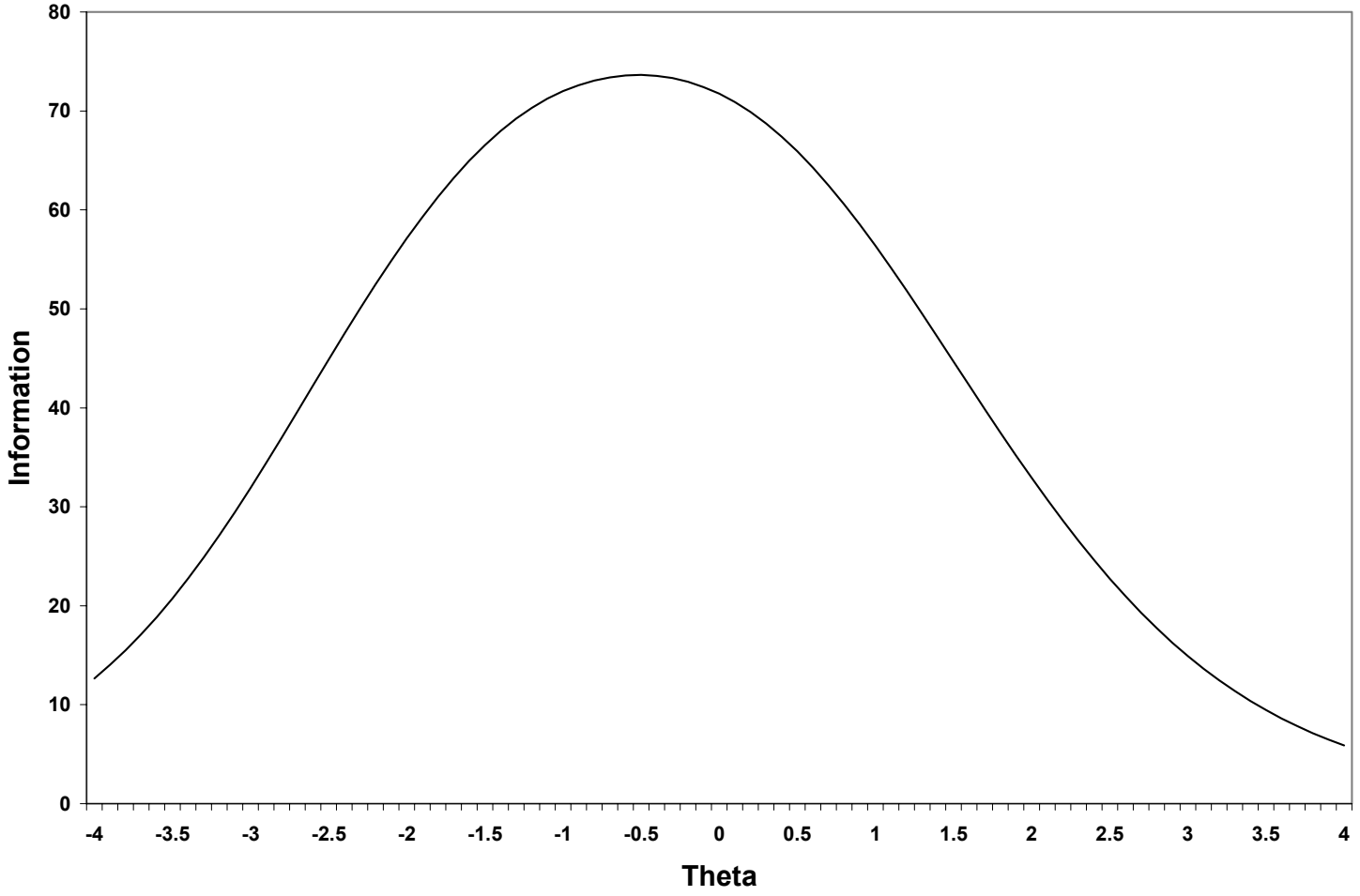
Graded Response Model					
Item Parameter	***N***	***Mean***	***St. Dev.***	***Min.***	***Max.***
Item Discrimination	157	1.2749	0.2281	0.7373	1.9168
Category Boundary 1	157	-1.4924	0.8127	-3.3712	0.8071
Category Boundary 2	157	0.2470	1.1211	-1.8629	4.5490
Category Boundary 3	58	0.2517	0.8815	-1.1926	2.0422
Category Boundary 4	29	0.6688	0.8262	-0.4889	3.3881

Table 13

Mean, Standard Deviation, Minimum, and Maximum of Item Discrimination Parameters Across Strata for the Graded Response Model

Graded Response Model					
Strata	***N***	***Mean***	***St. Dev.***	***Min.***	***Max.***
Stratum 1	32	1.03	0.15	0.74	1.34
Stratum 2	32	1.17	0.13	1.03	1.46
Stratum 3	31	1.27	0.13	1.13	1.60
Stratum 4	31	1.36	0.14	1.22	1.72
Stratum 5	31	1.56	0.17	1.33	1.92

Figure 3: Test Information Function for N=157 items Under the Graded Response Model



while there is some degree of overlap across the strata, the average 'a'-value does increase as the strata increases.

Twenty-nine nonconvergent cases were deleted in a listwise fashion after all conditions had been completed. As with the previous two models, a case was defined as nonconvergent if, once the end of the test had been reached, the trait estimate was greater than or equal to 4.0 or less than or equal to –4.0, or if maximum likelihood estimation had never been reached. All 29 cases deleted from the graded response model sample were nonconvergent due to the trait estimates being too extreme ($\hat{\theta} \leq -4.0; \hat{\theta} \geq 4.0$). The number of nonconvergent cases for each condition is listed in Table 14. The number of nonconvergent cases was uniformly low with between 1 and 6 nonconvergent cases reported across conditions. The remainder of the results for the graded response model conditions are reported on the sample (N=971) of observations which remained after the nonconvergent cases had been deleted.

The average theta estimate and standard error for each condition are also reported in Table 14. The mean of the known thetas was -0.01 with a standard deviation of 0.95. All conditions reported a slightly increased mean of the theta estimate relative to the standard deviation of the known thetas (between 0.00 and 0.04), except for the conditional Sympson-Hetter which reported an average theta value of -0.01. Standard deviations of the theta estimates were elevated across conditions and ranged from 1.10 to 1.13. As expected, no exposure control yielded

Table 14

Means (and Standard Deviations) for Estimated Theta and Standard Error; Number of Nonconvergent Cases for the Exposure Control Conditions Using the Graded Response Model

Graded Response Model			
Exposure Control Condition	***Theta* Estimate***	***Standard Error***	***Nonconvergent Cases***
No Exposure Control	0.00 (1.10)	0.30 (0.06)	6
Randomesque-3	0.03 (1.11)	0.31 (0.07)	2
Randomesque-6	0.02 (1.12)	0.32 (0.06)	5
Within .10 Logits-3	0.03 (1.13)	0.31 (0.06)	2
Within .10 Logits-6	0.04 (1.12)	0.33 (0.06)	4
Sympson-Hetter	0.01 (1.13)	0.31 (0.07)	4
Conditional Sympson-Hetter	-0.01 (1.12)	0.32 (0.06)	1
A-Stratified	0.02 (1.10)	0.34 (0.04)	5
Enhanced A-Stratified	0.00 (1.13)	0.35 (0.06)	3

*Note: Mean and SD for Known Thetas (N=971) were Mean=-0.01; SD=0.95

the lowest average standard error, 0.30, with the a-Stratified and enhanced a-Stratified yielding the highest average standard errors (0.34 and 0.35 respectively). The randomesque-3, within .10 logits-3, and Sympson-Hetter all yielded average standard errors of 0.31, while the randomesque-6 and conditional Sympson-Hetter yielded average standard errors of 0.32. Finally, within .10 logits-6 produced an average standard error of 0.33.

Table 15 displays values for the correlation between known and estimated theta, bias, SDM, RMSE, SRMSD, and AAD for each condition. The correlation between known and estimated theta was highest for no exposure control (0.89) and lowest for the within .10 logits-3 and enhanced a-Stratified conditions (0.85). Correlation coefficients were most comparable to the no exposure control condition for the randomesque-3, randomesque-6, and within .10 logits-6 conditions, where values ranged from 0.88-0.89. The Sympson-Hetter, conditional Sympson-Hetter, and a-Stratified conditions had correlations between 0.86 and 0.87. Bias and SDM were functionally zero across conditions with values ranging from 0.00 to –0.05 for bias and 0.00 to 0.04 for SDM. No exposure control yielded the lowest values for RMSE (0.50), SRMSD (0.69), and AAD (0.31). Within .10 logits-3, the Sympson-Hetter, and the enhanced a-Stratified conditions gave the highest RMSEs and SRMSDs (RMSE=0.59, 0.58, 0.58 respectively; SRMSD=0.74, 0.73, 0.73). The AAD values were highest for the Sympson-Hetter (0.35), a-Stratified design (0.35), and the enhanced a-Stratified design (0.36).

Exposure Control Parameters

Table 15

Correlation Coefficients between Known and Estimated Theta, Bias, RMSE, SDM, SRMSD, and AAD for the Exposure Control Conditions Using the Graded Response Model

Graded Response Model						
Exposure Control Condition	***Correlation***	***Bias***	***SDM***	***RMSE***	***SRMSD***	***AAD***
No Exposure Control	0.89	-0.01	0.01	0.50	0.69	0.31
Randomesque-3	0.88	-0.04	0.04	0.52	0.70	0.33
Randomesque-6	0.89	-0.02	0.02	0.51	0.69	0.34
Within .10 Logits-3	0.85	-0.04	0.04	0.59	0.74	0.34
Within .10 Logits-6	0.88	-0.05	0.04	0.53	0.70	0.33
Sympson-Hetter	0.86	-0.02	0.02	0.58	0.73	0.35
Conditional Sympson-Hetter	0.87	0.00	0.00	0.55	0.72	0.34
A-Stratified	0.87	-0.02	0.02	0.53	0.71	0.35
Enhanced A-Stratified	0.85	0.00	0.00	0.58	0.73	0.36

Thirty iterations were run using target exposure rates of 0.19 and 0.29 to set the exposure control parameters for the Sympson-Hetter and enhanced a-Stratified conditions. However, the observed maximum probability of administration did not approach the desired level. A target exposure rate of 0.39 was, therefore, chosen to ensure converge to the 0.40 level. The maximum probability of administration after the 30th iteration with a target exposure rate of 0.39 was 0.399 for the Sympson-Hetter and 0.406 for the enhanced a-Stratified design. For the conditional Sympson-Hetter, conditional maximum exposure rates of 0.39 were chosen and 30 iterations were conducted to ensure convergence of the parameters to this level. Concurrent with the findings of the Stocking and Lewis (1998) study, the maximum probability of administration at each theta level converged to values slightly greater than the target (0.416-0.469).

Pool Utilization and Exposure Rates

Table 16 contains the frequency of observed exposure rates along with the average, maximum, and standard deviation of exposure rates, and the percent of pool not administered for each condition. The observed average exposure rate for all conditions was 0.127 as dictated by the relationship between test length and pool size (Chen, Ankenmann, & Spray, 1999). The standard deviation of exposure rates was highest for the no exposure control condition (0.232), indicating the most uneven item exposure rates, and lowest for the randomesque-6, within .10 logits-6, and conditional Sympson-Hetter conditions (0.13 for all three conditions), indicating the most even item exposure rates. The maximum exposure rate of any item was largest for the no

Table 16

Pool Utilization and Exposure Rates for the Exposure Control Conditions Using the Graded Response Model

Graded Response Model									
Exposure Control Condition	No Exposure Control	Randomesque 3	Randomesque 6	Within .10 Logits-3	Within .10 Logits-6	Sympson Hetter	Conditional Sympson-Hetter	A-Stratified	Enhanced A-Stratified
Number of Items	157	157	157	157	157	157	157	157	157
Exposure Rate									
1	0	0	0	0	0	0	0	0	0
.91-.99	2	0	0	0	0	0	0	0	0
.81-.90	2	0	0	0	0	0	0	2	0
.71-.80	5	1	0	1	0	0	0	2	0
.61-.70	5	5	0	4	0	0	0	4	0
.51-.60	4	4	0	5	0	0	0	6	0
.41-.50	4	5	9	6	9	10	0	8	7
.36-.40	2	2	8	3	7	23	14	5	24
.31-.35	2	5	5	2	5	4	13	2	3
.26-.30	0	7	4	6	4	6	12	1	3
.21-.25	4	11	6	13	6	2	6	4	8
.16-.20	7	7	19	7	21	5	10	8	10
.11-.15	7	11	24	12	26	11	14	12	11
.06-.10	12	20	22	18	19	15	24	9	10
.01-.05	28	43	42	45	43	35	40	7	17
Not Administered	73	36	18	35	17	46	24	87	64
Exposure Rate AVG	0.127	0.127	0.127	0.127	0.127	0.127	0.127	0.127	0.127
Exposure Rate SD	0.232	0.173	0.13	0.173	0.129	0.158	0.13	0.208	0.154
Exposure Rate MAX	0.935	0.733	0.494	0.735	0.496	0.422	0.395	0.896	0.421
% of Pool Not Administered	46%	23%	11%	22%	11%	29%	15%	55%	41%

exposure control condition with a value of 0.935 and lowest for conditional Sympson-Hetter with a value of 0.395. The Sympson-Hetter, enhanced a-Stratified design, randomesque-6, and within .10 logits-6 also had relatively low maximum exposure rates (0.422, 0.421, 0.494, and 0.496 respectively). The randomesque-3, within .10 logits-3, and a-Stratified design, on the other hand, resulted in maximum exposure rates above acceptable levels (0.733, 0.735, and 0.896 respectively).

As observed with the generalized partial credit model, the percent of pool not administered was actually greater for the a-Stratified design (55%) than for the no exposure control condition (46%), and only slightly better for the enhanced a-Stratified design (41%). This finding further underscores the concern expressed previously with regard to the postulated versus observed impact of the a-Stratified design on pool utilization. The percent of pool not administered was lowest for the randomesque-6, within .10 logits-6, and conditional Sympson-Hetter conditions with 11% of the pool not used in the first two conditions and 15% remaining unused in the third condition.

In terms of test security, the randomesque-3 and within .10 logits-3 procedures demonstrated improvement over the no exposure control condition, but did not perform as well as their 6 item group variants. While the Sympson-Hetter controlled the maximum exposure rate to a level close to the target value, the percent of pool not administered under this procedure (29%) was well above that observed with other procedures.

Item Overlap

Audit trails for each simulee were compared to the audit trails of every other simulee resulting in a pairwise fashion resulting in 470,935 comparisons per condition. Table 17 contains the average item overlap for all simulees, those of different trait levels (known thetas differed by more than two logits), and those of similar trait levels (known thetas differed by two logits or fewer) for each condition. Item overlap information is presented both in terms of the average percent of items shared across a 20 item test and the average number of items shared across a 20 item test. No exposure control produced the highest overall overlap rates with an average of 55% overlap and the randomesque-6, within .10 logits-6, and conditional Sympson-Hetter procedures resulted in the lowest overall overlap rates with an average of 26% overlap. The a-Stratified design yielded an unsatisfactory 46% overall overlap, with the randomesque-3 and within .10 logits-3 performing only slightly better with overall overlap rates between 36%. The Sympson-Hetter and enhanced a-Stratified conditions resulted in overlap rates between 31 and 32%. Results both for examinees of similar and different trait levels demonstrate the same pattern with highest overlap rates (59% similar; 30% different) occurring with the no exposure control condition and lowest overlap rates (27% similar; 17% different) occurring with the randomesque-6, within .10 logits-6, and conditional Sympson-Hetter conditions.

Additional Analyses

Inspection of the results for the graded response model revealed some peculiarities that merited additional investigation. Primary to these observations were

Table 17

Item Overlap Rates for the Exposure Control Conditions Using the Graded Response Model

Graded Response Model			
Exposure Control Conditions	***Overall Average Overlap*** (N=470,935)	***Different Abilities Average Overlap*** (N=64,788)	***Similar Abilities Average Overlap*** (N=406,147)
No Exposure Control	10.93 55%	5.90 30%	11.74 59%
Randomesque-3	7.22 36%	4.18 21%	7.70 39%
Randomesque-6	5.19 26%	3.48 17%	5.46 27%
Within .10 Logits-3	7.22 36%	4.12 21%	7.71 39%
Within .10 Logits-6	5.13 26%	3.40 17%	5.41 27%
Sympson-Hetter	6.43 32%	3.63 18%	6.88 34%
Conditional Sympson-Hetter	5.18 26%	3.34 17%	5.48 27%
A-Stratified	9.30 46%	5.45 27%	9.91 50%
Enhanced A-Stratified	6.24 31%	4.30 21%	6.55 33%

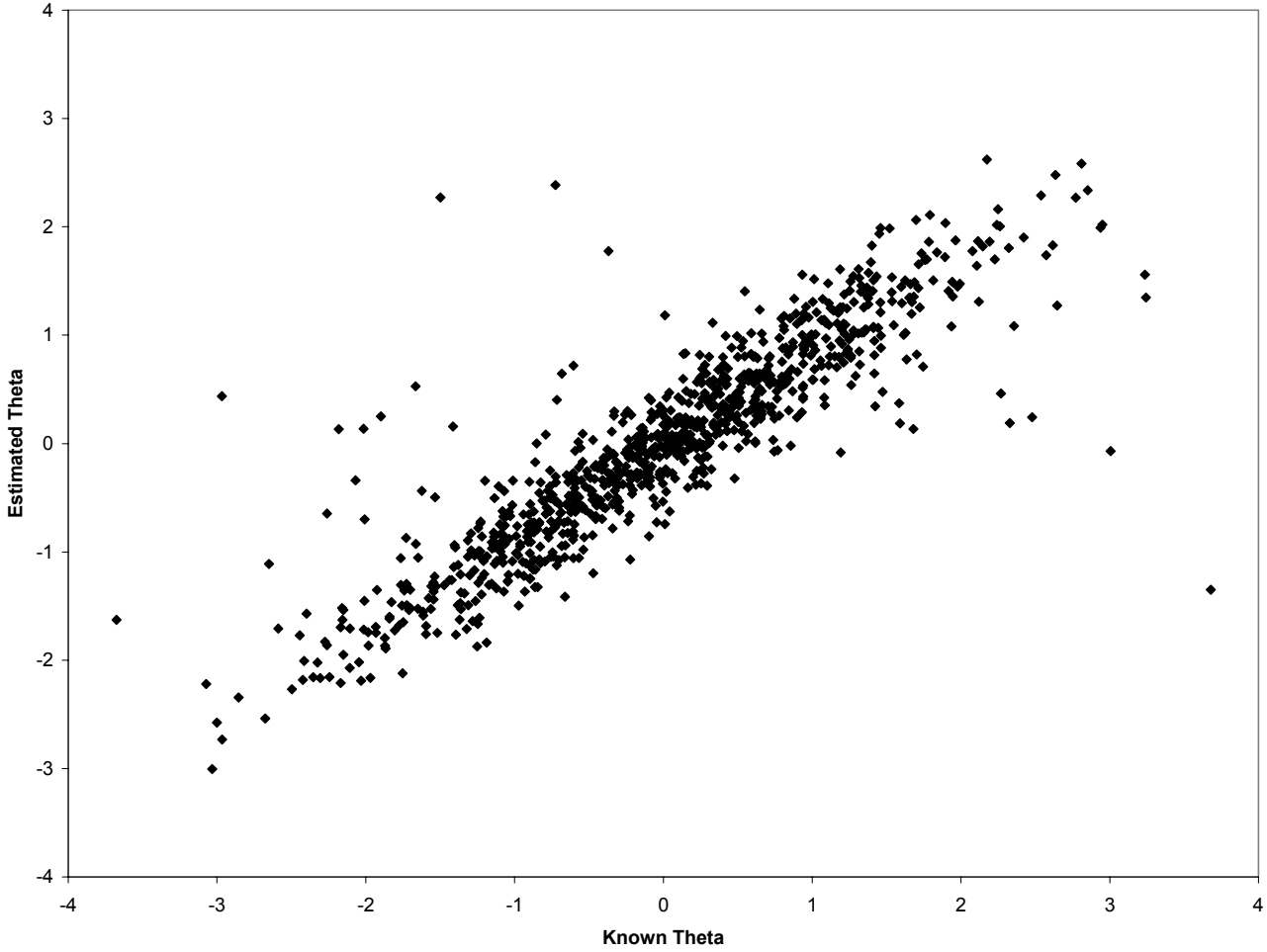
the worse than expected results for all measurement precision variables. Correlations between known and estimated theta were lower than expected (0.85-0.89) and values for RMSE (0.50-0.58), SRMSD (0.69-0.74), and AAD (0.31-0.36) were higher than expected across the board. In addition, the ordering of certain procedures in terms of measurement precision was counterintuitive. The within .10 logits-3 procedure yielded worse results than the within .10 logits-6 procedure. This was not consistent with the findings in the partial credit and generalized partial credit models. Further, it simply did not make sense as we would expect the smaller item group size to provide results closer to the no exposure control condition.

A scatterplot of the relationship between known and estimated thetas for the no exposure control condition (r=0.89) is provided in Figure 4. It can be clearly seen that there are a handful of cases which fall far from the main cluster of points. Examination of these cases revealed unacceptably high standard errors associated with the simulees' final theta estimate. In addition, the theta estimates for these cases tended to be extreme. It was, therefore, surmised that these cases represented aberrant response strings which were not well estimated by the model, and were suppressing the correlation value.

Previous research has also demonstrated that the graded response model does have difficulty in estimating ability when using MLE, producing in one study, correlations between known and estimated theta in the range of 0.91 to 0.93 (Hou, Chen, Dodd, & Fitzpatrick, 1996). Further, Chen (1996) concluded that "…MLE yields less satisfactory results than does EAP in the CAT based on a difference

Figure 4: Known vs. Estimated Theta for Graded Response Model No Exposure Control Condition

Number of Items Administered 20



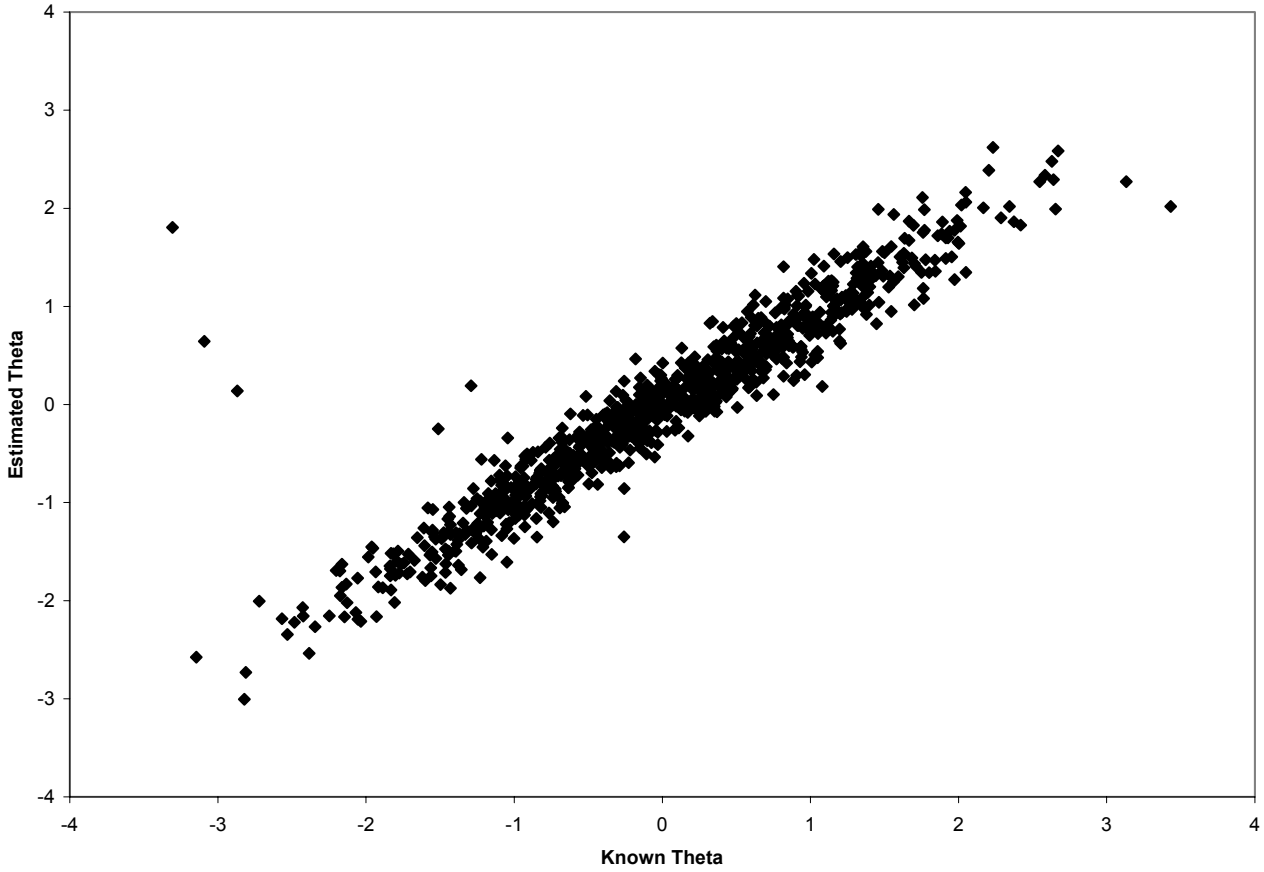
model..." The fact that the current study records correlation values even smaller than those previously observed can be explained by the implementation of a content balancing mechanism in all conditions which decreased measurement precision over the optimal. Therefore it appears that findings with the graded response model in the current study are consistent with previous research with regard to the expected level of measurement precision under MLE.

To further verify this explanation, it was decided to run the graded response model conditions with a longer test length to see if better measurement precision could be obtained if more items were administered. Figure 5 presents a scatterplot of the relationship between known and estimated thetas for the no exposure control condition with a test length of 35 items. The correlation at 35 items increased substantially to 0.95 and visual inspection of the plot demonstrates that the aberrant cases seen with the 20 item test length have greatly improved. The precision of five cases which exhibited the largest standard error was compared between the 20 item and 35 item tests. With the 20 item test, these five cases had an average standard error of 0.78. With the 35 item test, the average standard error dropped to 0.22. Clearly, then, the graded response model requires more items to reliably estimate ability when MLE is used.

With regard to the counterintuitive ordering of the within .10 logits procedures on the measurement precision variables, it is believed that the presence of the aberrant cases in the sample from the 20 item test created noise which caused these unexpected results. Using a 35 item test length eliminated this phenomenon, resulting

Figure 5: Known vs. Estimated Theta for Graded Response Model No Exposure Control Condition

Number of Items Administered 35



in a correlation coefficient of 0.95 for the within .10 logits-3 and 0.94 for the within .10 logits-6.

Chapter V

Discussion

The results of this study will be grouped and discussed by the type of exposure control procedure employed in the current study. The three categories of methods are randomization procedures, conditional selection procedures, and stratification procedures.

Randomization Procedures

The randomization procedures, represented in this study by the randomesque and within .10 logits methods were easy to implement and required only a minimum of additional programming over that needed to run the CAT. One of the driving questions of this study was whether or not these randomization procedures could be as effective (or more so) for protecting test security as the conditional selection strategies such as the Sympson-Hetter. The answer to this question seems to be overwhelmingly positive. In all three models studied, the randomesque-6 and within .10 logits-6 procedures provided relatively good control over exposure and overlap rates and the highest degree of pool utilization of any procedures examined in the current study.

A second question posed by this study was to address what impact item group size would have on the effectiveness of the randomization procedures for controlling exposure. It is evident from the results that item group size does have strong implications for the ability of these procedures to control exposure, with an item group size of 6 providing superior test security to an item group size of 3 for both the

randomesque and within .10 logits procedures. Maximum exposure rates for the 3 item group variations were on the order of 0.10-0.15 larger than for the 6 item group variations for the partial credit model and on the order of 0.20-0.25 larger for the generalized partial credit and graded response models. Overlap rates for the 3 item group variations were 4-5% larger for the partial credit model and 9-10% larger for the generalized partial credit and graded response models than for the 6 item group variations. Finally, the percent of pool not administered was twice as large for the 3 item group variations as for the 6 item group variations.

While these procedures were quite effective for improving test security, a corresponding decrease in measurement precision was observed. The average standard error increased between 0.02 and 0.03 with the use of either the randomesque-6 or within .10 logits-6 procedures across all three models. Further, the correlation between known and estimated theta decreased and the RMSE increased for both the generalized partial credit and graded response models.

In distinguishing between the two randomization procedures presented in this study, it appears that the randomesque procedure may have a very slight edge over the within .10 logits procedure both in terms of its effectiveness in controlling exposure and in terms of its impact on measurement precision with the generalized partial credit and graded response models. However, these differences are extremely small and both procedures performed very well.

Conditional Selection Procedures

The Sympson-Hetter and conditional Sympson-Hetter both proved to be challenging and time-consuming to implement in the current study. The simulations to set the exposure control parameters took between 4 and 6 hours to run for the Sympson-Hetter and between 8 and 12 hours to run for the conditional Sympson-Hetter depending on the exact processor speed of the computer (ranged from Pentium II 266Mhz to Athelon T-Bird 1Ghz). The simulations had to be run multiple times when convergence of the parameters to the initial target exposure rates was not achieved. In addition, the programming requirements for these procedures were substantial.

The performance of the Sympson-Hetter with regard to test security was disappointing given the amount of effort required. In all three models, the maximum observed exposure rate for the Sympson-Hetter condition was slightly above, but close to the target value (0.434 partial credit, 0.419 generalized partial credit, 0.422 graded response). Overlap rates and percent of pool not administered, while reduced over the no exposure control condition, were still high relative to other options. The Sympson-Hetter appears to be good at one thing—holding down the maximum exposure rate. Reductions in overlap and percent of pool not administered are almost byproducts of this main effort and are not explicit goals of the procedure.

Measurement precision for the Sympson-Hetter remained relatively high, however, with no measurable differences over the no exposure control condition with the partial credit model and only slight decreases in measurement precision with the generalized partial credit and graded response models. The rather lenient target

exposure rates used in the current study might provide at least some explanation for this. Had more restrictive levels of exposure control been possible, it is anticipated that measurement precision would have suffered more of an impact.

The conditional Sympson-Hetter faired somewhat better than its unconditional counterpart. The global maximum exposure rate was controlled to levels below those seen with the Sympson-Hetter in spite of the fact that the procedure only guarantees conditional target exposure rates. The conditional Sympson-Hetter has been criticized for the fact that its lack of global exposure control can allow global exposure rates to exceed acceptable levels, however, that did not appear to be the case in the current study. Further, the conditional Sympson-Hetter did a much better job of controlling overlap rates than the Sympson-Hetter, providing overlap rates equivalent to or just slightly above those seen with the randomesque-6 and within .10 logits-6 procedures. The effect of controlling exposure conditional on trait level is especially visible when looking at overlap rates among simulees of similar trait level. This is where the strength of the conditional Sympson-Hetter lies. Finally, pool utilization is much improved with the use of the conditional Sympson-Hetter yielding rates equivalent to or just slightly above those seen with the randomesque-6 and within .10 logits-6 procedures.

The tradeoff for this level of test security came in the form of lower measurement precision. This impact was most noticeable with the generalized partial credit and graded response models, where the average standard error increased from 0.28 to 0.31 for the generalized partial credit model and from 0.30 to 0.32 for the

graded response model, the correlation between known and estimated theta dropped from 0.96 to 0.93 (generalized partial credit model) and from 0.89 to 0.87 (graded response model), and the RMSE increased from 0.29 to 0.40 (generalized partial credit model) and from 0.50 to 0.55 (graded response model), over the no exposure control conditions. The partial credit model showed a negligible impact with an increase in average standard error from 0.27 to 0.28 relative to the no exposure control condition and no changes in other values.

Despite the favorable results for controlling exposure, the conditional Sympson-Hetter cannot be fully endorsed due to the inability to implement the procedure in its complete and intended fashion. Attempts were made to set the K_i values of 20 items at each theta level equal to 1.0 to ensure that there would always be at least as many items available for administration as the maximum test length. However, exposure control parameters for the conditional Sympson-Hetter would not converge with this operation in place. Therefore, this step was omitted in order to allow for convergence of the exposure control parameters. Because this step was omitted, it is possible that for a given simulee there might have been an insufficient number of items available to administer a full test and testing would have stopped prematurely. In the particular sample of 1000 simulees used for this study, this possibility did not occur, therefore, results have been presented to allow comparison of the conditional Sympson-Hetter to other procedures. However, operational implementation of this procedure would most likely involve a much larger volume of examinees which increases the opportunity for such an occurrence.

Stratification Procedures

The poor performance of the a-Stratified and enhanced a-Stratified designs in this study is surprising and merits further consideration. The concept of the a-Stratified design is essentially very simple—stratify the item pool according to item discrimination parameters and limit item selection at various stages of testing to certain strata. This should, in theory, boost pool utilization by ensuring that those items with lower discrimination values which would otherwise be overlooked have a better chance to be administered. However, when there are content and item type constraints to consider, such as the content area affiliation and number of categories used in the current study, modification to the procedure is necessary which both increases its complexity and may nullify the benefits of a-stratification.

The maximum exposure rate with the a-Stratified design was high for both the generalized partial credit and graded response models. This, by itself, is not unexpected, as other researchers (Chang & Ying, 1999; Parshall, Hogarty, & Kromrey, 1999) have found similar results and there is no mechanism in the procedure by which to directly control the maximum exposure rate. However, what is surprising are the extremely high values for percent of pool not administered. This variable should be the strength of the a-Stratified design, but, in fact, the current study shows that pool utilization actually goes down in comparison to the maximum exposure condition when the a-Stratified design is used. It is hypothesized that this results from overlap in discrimination values across strata due to the multiple stratification of the item pool necessary to meet content and item type constraints.

Tables 7 and 13 show that while the average a-values do increase across strata, the overlap of item discrimination values across strata is substantial. With the generalized partial credit model (Table 7), the minimum discrimination value of an item in stratum 4 (0.69) is not substantially different from discrimination values seen in stratum 1 (0.54 to 0.87). In contrast, with the unmodified a-Stratified design, there would be no overlap across strata. With such extensive overlap amongst strata, the effect of stratification is largely nullified, and therefore the advantages of the a-Stratified design are lost. These findings suggest that, it is not only important that the average a-value increase across strata as suggested by Yi and Chang (2000) and Leung (2001), but also that the overlap across strata be minimized to maintain the benefits of stratification. Further research is necessary to make specific recommendations as to the allowable amount of overlap across strata.

Differences Across Models

While the design of the current study does not allow direct variable by variable comparisons to be made across the three models, certain general trends do emerge. First it is interesting to note that while there were some differences in the performance of the procedures across the models, the randomesque-6 and within .10 logits-6 stood out as the apparent best options in all three cases. Secondly, item group size of the randomization procedures did have a strong impact on test security with the larger item group size providing better exposure control in all three cases. Third, the relative rankings of the other procedures was maintained across the models.

Areas of differential performance did appear in the results which seemed to support the Rasch vs. non-Rasch dichotomy discussed in the Statement of Problem. Exposure rates, overlap rates, and percent of pool not administered were, on the whole, lowest when the partial credit model was used. Additionally, the implementation of the various exposure control strategies had the least (almost negligible) impact on measurement precision with the partial credit model. Conversely, the impact on measurement precision was much more pronounced with both the generalized partial credit and graded response models. This concurs with the findings of previous studies of exposure control with polytomous models (Pastor et al., 1999; Davis et al., 2000, Pastor, Dodd, & Chang, in press; Davis & Dodd, 2001) and seems to support the extension of Way's (1998) Rasch vs. non-Rasch dichotomy to the polytomous case.

No obvious differences were observed between the graded response and generalized partial credit models with regard to the performance of the exposure control mechanisms or their impact on measurement precision except for the difficulties with trait level estimation using the graded response model previously discussed in the Additional Analyses section of the Results. Therefore, nothing in the current study yields evidence that there might be differences between the difference and divide-by-total models with regard to exposure control.

Conclusions and Directions for Future Research

The results of this study for all three models provide strong support for the use of randomization procedures with sufficient item group sizes to control test security

with polytomous item pools. The randomesque-6 and within .10 logits-6 procedures proved themselves to be simple to implement and provided substantial reductions in exposure rates and item overlap as well as substantial increases in pool utilization over the no exposure control condition. Further, when compared head to head with other options for controlling exposure, they demonstrated comparable or better performance on almost all measures of test security. On those measures where performance was somewhat less than other procedures, this small cost in performance has to be weighed against the expense of added complexity that would be necessary to adopt the competitive option. Only the conditional Sympson-Hetter was competitive to the randomization procedures. While it did produce slightly lower maximum exposure rates and some comparable pool utilization and overlap rates, the conditional Sympson-Hetter is the single most complex procedure examined in the current study and was not able to be implemented in its complete and intended form. Further, at least for the generalized partial credit and graded response models, the negative impact on measurement precision seemed to be somewhat greater for the conditional Sympson-Hetter than for the randomesque-6 or within .10 logits-6 procedures. The results of the current study, therefore, indicate that the randomesque-6 and within .10 logits-6 procedures provide the best all around options for controlling test security when ease of implementation and impact to measurement precision are considered.

Having made these statements, however, it is important to point out that none of the procedures examined was able to control exposure rates to levels traditionally

acceptable with dichotomous item pools (0.20) or to levels previously used with polytomous models (0.30) with the test structure used in the current study. A maximum observed exposure rate of approximately 0.40 was the lowest rate which could be obtained with any of the procedures across all three models. This suggests that the size and structure of the item pool with regard to content and item characteristics plays a large role in determining the ability of test developers to control item exposure. Also, it is important to recognize that in order to make these gains in test security, some measurement precision was sacrificed in the generalized partial credit and graded response models. As Parshall, Davey, and Nering (1998) have suggested, there seems to be an unavoidable tradeoff which occurs between exposure control and measurement precision. None of the procedures examined in the current study demonstrated the ability to defy this relationship, and, therefore, a balance between these goals is the best which could be obtained.

Future research needs to be conducted to determine how different pool sizes and characteristics would affect the utility of the various exposure control mechanisms. The item pool size and characteristics examined in the current study were selected because they represented a realistic test structure that might be used by a large scale testing program. However, it is quite clear that the content and item type constraints imposed by this structure severely affected the ability of certain procedures to control test security because there were very few items available for certain combinations of content and item type. While having an item pool of sufficient size to estimate ability and minimize item exposure is important, Stocking

and Lewis (2000) emphasized the need for the available item pool to adequately reflect test specifications for content, item type, and other nonstatistical properties. Their research has demonstrated that item pools which do not have a sufficient number of items to match these specifications make the use of conditional exposure control strategies difficult, if not impossible, to implement because of problems in obtaining the convergence of the exposure control parameters. The small nature of polytomous item pools can amplify this issue. However, since practical and economic constraints often make expansion of an item pool infeasible, it is important to identify appropriate options for controlling exposure with a less than optimal item pool.

This research has identified two strategies which seem to be useful for controlling exposure in small polytomous item pools—the randomesque and within .10 logits procedures. However, the results have also shown that the item group size from which the next item to be administered is randomly chosen has a significant impact on their effectiveness. Future research should provide a more detailed study of the item group size variable both by itself and in relation to the size of the pool to attempt to develop a set of recommendations for implementing the procedures. Finally, other randomization procedures such as Revuelta and Ponsoda's (1998) progressive and restricted no exposure control procedures should be examined with polytomous item pools to determine whether they can offer similar benefits for exposure control.

References

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

Bejar, I.I. (1991). A methodology for scoring open-ended architectural design problems. Journal of Applied Psychology, 76, 522-532.

Bennett, R.E., Morley, M., & Quardt, D. (1998, April). Three response types for broadening the conception of mathematical problem solving in computerized-adaptive tests. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Bergstrom, B.A., & Lunz, M.E. (1999). CAT for certification and licensure. In F. Drasgow & J. Olson-Buchanan (Eds.), Innovations in computerized assessment (pp. 67-91). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, Statistical theories of mental test scores (chapters 17-20). 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, 433-459.

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

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998, April). Computer analysis of essays. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Chang, H.H., Qian, J., & Ying, Z. (1999). A-stratified multistage CAT with b-blocking. Manuscript accepted for publication.

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

Chang, H.H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. Applied Measurement in Education, 23(3), 211-222.

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

Chang, S., Ansley, T.N., & Lin, S. (2000). Performance of item exposure control methods in computerized adaptive testing: Further explorations. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Chen, S. (1996). A comparison of maximum likelihood estimation and expected a posteriori estimation in computerized adaptive testing using the generalized partial credit model. Unpublished doctoral dissertation, University of Texas, Austin.

Chen, S. Ankenmann, R.D., & Spray, J.A. (1999, April). Exploring the relationship between item exposure rate and test overlap rate in computerized

adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Chen, S., Hou, L., & Dodd, B.G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation on CAT using the partial credit model. Educational and Psychological Measurement, 53, 61-77.

Clauser, B.E., Margolis, M.J., Clyman, S.G., & Ross, L.P. (1997). Development of automated scoring algorithms for complex performance assessments: A comparison of two approaches. Journal of Educational Measurement, 34, 141-161.

Davey, T., & Parshall, C.B. (1995, April). New algorithms for item selection and exposure control with computerized adaptive testing. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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

Davis, L.L., Pastor, D.A., Dodd, B.G., Chiang, C., & Fitzpatrick, S. (2000). An examination of exposure control and content balancing restrictions on item selection in CATs using the partial credit model. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Eignor, D.R., Stocking, M.L., Way, W.D., & Steffan, M. (1993). Case studies in computer adaptive test design through simulation (RR-93-56). Princeton, NJ: Educational Testing Service.

Fan, M., Thompson, T., & Davey, T. (1999). Constructing adaptive tests to parallel conventional programs. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

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

Hau, K.T., & Chang, H.H. (1998). Item selection in computerized adaptive testing: Should more discriminating items be used first? Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Haynie, K.A., & Way, W.D. (1994, April). The effects of item pool depth on the accuracy of pass/fail decisions for NCLEX using CAT. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

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

Hou, L., Chen, S., Dodd, B.G., & Fitzpatrick, S. J. (1996, April). The effect of methods of theta estimation, prior distribution, and number of quadrature points on CAT using the graded response model. Paper presented at the annual meeting of the American Educational Research Association, New York.

Kalohn, J.C. & Spray, J.A. (April, 1998). Effect of item selection on item exposure rates within a computerized classification test. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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

Koch, W.R., & Dodd, B.G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. Applied Measurement in Education, 2(4), 335-337.

Luecht, R.M. (1998). A framework for exploring and controlling risks associated with test item exposure over time. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Lunz, M.E., & Bergstrom, B.A. (1994). An empirical study of computer adaptive test administration formats. Journal of Educational Measurement, 31, 251-263.

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

Leung, C., Chang, H.H., & Hau, K. (1999). An enhanced stratified computerized adaptive testing design. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Leung, C., Chang, H.H., & Hau, K. (2000). Content balancing in stratified computerized adaptive testing designs. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Leung, C. (2001). Stratified computerized adaptive testing: Further control on item exposure and extension to constrained situations (Doctoral dissertation, The Chinese University of Hong Kong, 2001).

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

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

McBride, J.R. (1997). Research antecedents of applied adaptive testing. In Sands, W.A., Waters, B.K., & McBride, J.R. (Eds.), Computerized adaptive testing: From inquiry to operation (pp.47-58). Washington, DC: American Psychological Association.

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.223-226). New York, Academic Press.

McKinley, R.L., & Mills, C.N. (1985). A comparison of several goodness-of-fit statistics. Applied Psychological Measurement, 1985, 49-57.

Mills, C.N. (1999). Development and introduction of a computer adaptive graduate records examinations general test. In F. Drasgow & J. Olson-Buchanan

(Eds.), Innovations in computerized assessment (pp. 117-135). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Morrison, C.A., Subhiyah, R.G., & Nungester, R.J. (1995, April). Item exposure rates for unconstrained and content-balanced computerized adaptive tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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

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

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

Nering, M.L., Davey, T., & Thompson, T. (1998). A hybrid method for controlling item exposure in computerized adaptive testing. Paper presented at the annual meeting of the Psychometric Society, Urbana, IL.

O'Neill, T., Lunz, M.E., & Thiede, K. (April, 1998). The impact of item exposure on repeat examinee performance for computerized adaptive tests. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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

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

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

Pastor, D.A., Chiang, C., Dodd, B.G., & Yockey, R., (1999, April). Performance of the Sympson-Hetter exposure control algorithm with a polytomous item bank. Paper presented at the annual meeting of American Educational Research Association, Montreal, Canada.

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

Patsula, L.N., & Steffan, M. (1997). Maintaining item and test security in a CAT environment: A simulation study. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9, 401-412.

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

Reckase, M.D., & McKinley, R.L. (1991). The discriminating power of items that measure more than one dimension. Applied Psychological Measurement, 15, 361-373.

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

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

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

Segall, D.O., Moreno, K.E., & Hetter, R.D. (1997). Item pool development and evaluation. In Sands, W.A., Waters, B.K., & McBride, J.R. (Eds.), Computerized adaptive testing: From inquiry to operation (pp.117-130). Washington, DC: American Psychological Association.

Stocking, M.L. (1992). Controlling item exposure rates in a realistic adaptive testing paradigm. (Research Report 93-2). Princeton, NJ: Educational Testing Service.

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

Stocking, M.L., & Lewis, C. (1995). A new method for controlling item exposure in computer adaptive testing (Research Report 95-25). Princeton, NJ: Educational Testing Service.

Stocking, M.L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing, Journal of Educational and Behavioral Statistics, 23(1), 57-75.

Stocking, M.L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W.J. van der Linden & C.A.W. Glas (Eds.), Computerized adaptive testing: Theory and practice (pp. 163-182). Netherlands: Kluwer Academic Publishers.

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

Stocking, M.L., & Swanson, L. (1998). Optimal design of item banks for computerized adaptive tests. Applied Psychological Measurement, 22, 271-279.

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

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. San Diego, CA: Navy Personnel Research and Development Center.

Tang, K.L., Jiang, H., & Chang, H.H. (1998). A comparison of two methods of controlling item exposure in computerized adaptive testing. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Thomasson, G.L. (1998). CAT Item exposure control: New evaluation tools, alternate methods and integration into a total CAT program. Paper presented at the

annual meeting of the National Council of Measurement in Education, San Diego, CA.

Traub, R.E. (1983). A priori considerations in choosing an item response model. In R.K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

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

Wainer, Howard. (2000). Rescuing computerized testing by breaking Zipf's law. Journal of Educational and Behavioral Statistics, 25, 203-224.

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

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

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

Way, W., Zara, A., & Leahy, J. (1996, April). Strategies for managing item pools to maximize item security. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Wright, B.D. (1968). Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.

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

Zipf, G.K. (1949). Human behavior and the principle of least effort. Cambridge, MA: Addison-Wesley.

VITA

Laurie Laughlin Davis was born Laurie Michele Laughlin in Athens, Georgia on August 27, 1975, the daughter of Roseann Laughlin and Donald Lewis Laughlin. After completing her work at S.H. Rider High School, Wichita Falls, Texas, in 1993, she entered the University of Texas at Dallas in Richardson, Texas and majored in Psychology. She received the degree of Bachelor of Arts from the University of Texas at Dallas in May 1997 with *summa cum laude* honors. In September 1997, she entered the Graduate School of The University of Texas and studied quantitative methods and psychometrics in the Department of Educational Psychology.

Permanent Address: 4831 Angelina, Wichita Falls, Texas 76308

This dissertation was typed by the author.