The Dissertation Committee for Aimee Michelle Boyd certifies that this is the approved version of the following dissertation:

Strategies for Controlling Testlet Exposure Rates in

Computerized Adaptive Testing Systems

COMMITTEE:

Barbara Dodd, Supervisor

Gary Borich

Hua Hua Chang

Steven Fitzpatrick

George Kozmetsky

Strategies for Controlling Testlet Exposure Rates in Computerized Adaptive Testing Systems

by

Aimee Michelle Boyd, B.S., M.A.

Dissertation

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

The University of Texas at Austin
May 2003

Dedicated to my parents

Don and Lori Summers

Acknowledgements

There are several individuals I wish to thank for supporting me throughout graduate school and during the dissertation process. I am most grateful for the intellectual and emotional support provided by my dissertation supervisor, Barbara Dodd. I was fortunate to attend several of her courses in which I not only gained a theoretical understanding of measurement and statistics, but I also learned the practical applications. As my dissertation supervisor, Dr. Dodd, encouraged me to stretch beyond my limits. Her continuous support and guidance during the dissertation process was phenomenal. In addition, I am very thankful for the friendship that has emerged.

Throughout most of my graduate education I was a graduate research assistant at the IC^2 Institute where I had the privilege to work with and learn from George Kozmetsky. Dr. Kozmetsky provided insight into blending academic research with the business world through his valuable guidance and advice. In addition, I am very thankful for the support, understanding, and friendship of my coworkers, Melinda Jackson and Gale McDonnell.

I wish to thank Gary Borich for his guidance and mentoring while evaluating the EnterTech Project, completing my master's thesis, and throughout the dissertation process. I am grateful to Steven Fitzpatrick for his valuable edits to my dissertation and for his excellent SAS programming skills. I am thankful for the support and encouragement provided by Hua-Hua Chang.

I am grateful to Ellen Julian and the research section of the Medical College Admission Test for the Association of American Medical Colleges for supporting the initial research and providing the data for my dissertation.

I have dedicated my dissertation to my father and mother, Don and Lori Summers, to whom I owe everything. My parents provided a loving home during my childhood, instilling in me a drive and an enthusiasm for learning. As I embraced adulthood, they gave me wisdom and friendship. Their struggles and accomplishments have allowed me to glimpse the power of love and friendship that stems from marriage. Thank you, thank you, thank you, Mom and Dad.

This journey would not have been completed without the support, friendship, and love of my husband, Brett. I am very grateful to Brett for his constant encouragement. Thank you for the sacrifices that were made that allowed me to complete my dissertation. I look forward to the next journey that I will take with you.

Strategies for Controlling Testlet Exposure Rates in

Computerized Adaptive Testing Systems

Publication No. _____

Aimee Michelle Boyd, Ph.D.

The University of Texas at Austin, 2003

Supervisor: Barbara G. Dodd

Exposure control procedures in computerized adaptive testing (CAT) systems protect item pools from being compromised, however, this impacts measurement precision. Previous research indicates that exposure control procedures perform differently for dichotomously scored versus polytomously scored CAT systems. For dichotomously scored CATs, conditional selection procedures are often the optimal choice, while randomization procedures perform best for polytomously scored CATs. CAT systems modeled with testlet response theory have not been examined to determine optimal exposure control procedures.

This dissertation examined various exposure control procedures in testlet-based CAT systems using the three-parameter logistic testlet response theory model and the partial credit model. The exposure control procedures were the randomesque

procedure, the modified within .10 logits procedure, two levels of the progressive restricted procedure, and two levels of the Sympson-Hetter procedure. Each of these was compared to a baseline no exposure control procedure, maximum information. The testlets were reading passages with six to ten multiple-choice items.

The CAT systems consisted of maximum information testlet selection contingent on an exposure control procedure and content balancing for passage type and the number of items per passage; expected a posteriori ability estimation; and a fixed length stopping rule of seven testlets totaling fifty multiple-choice items. Measurement precision and exposure rates were examined to evaluate the effectiveness of the exposure control procedures for each measurement model.

The exposure control procedures yielded similar results for measurement precision within the models. The exposure rates distinguished which exposure control procedures were most effective. The Sympson-Hetter conditions, which are conditional procedures, maintained the pre-specified maximum exposure rate, but performed very poorly in terms of pool utilization. The randomization procedures, randomesque and modified within .10 logits, yielded low maximum exposure rates, but used only about 70% of the testlet pool. Surprisingly, the progressive restricted procedure, which is a combination of both a conditional and randomization procedure, yielded the best results in its ability to maintain and control the maximum exposure rate and it used the entire testlet pool. The progressive restricted conditions were the optimal procedures for both the partial credit CAT systems and the three-parameter logistic testlet response theory CAT systems.

Table of Contents

List of Figures

List of Tables

CHAPTER ONE: INTRODUCTION

While examinees experience seamless tests on computers, they are unaware that their items are being selected and administered by a computer algorithm based on a measurement model, item content balancing, and an exposure control procedure that take into account the examinees' previous responses to items. The procedures that take place "behind the scenes" are required to meet the needs and goals of examinees, testing companies, and test developers.

For examinees, performance on a computerized adaptive test (CAT) represents admission to a favored college or university, scholarship or grant funding, or possibly employment or a job promotion. CATs offer examinees flexible testing schedules and the opportunity to obtain their scores immediately following administration of the test. For testing companies, a CAT represents the ability to provide continuous testing such that examinees may take tests on-demand. The testing company's focus is on providing accurate ability estimates and maintaining test reliability and validity. For test developers, CATs represent the ability to administer the most appropriate items to examinees, thereby reducing examinees' anxiety and frustration in dealing with longer tests. Administering tests on computers also enables test developers to incorporate new item formats.

Compared to traditional paper-and-pencil tests, CAT systems provide test developers with new challenges and in some cases new twists on old challenges. New challenges include accurately modeling item formats in the context of CATs where examinees' performances are based on different items, compared to other examinees,

for the same test. Computers allow for the development and administration of new item formats, such as incorporating graphics or sound. For CAT systems, test security provides a new twist to an old challenge. The frequency of CAT administrations makes CATs more susceptible to cheating by examinees. Extensive research with simulated and live CATs is needed to provide solutions to these challenges.

As test developers transform well established, reliable paper-and-pencil tests to CAT formats, various benefits are gained, including enhanced measurement precision, better test security, and shorter test lengths due to administration of more informative items (Wainer, 2000). In order to take advantage of these benefits, the psychometric properties of the test are based on item response theory (IRT), rather than traditional true score theory (Crocker & Algina, 1986). IRT enables two examinees, one with high ability and one with low ability, to encounter different subsets of items that are matched to the respective examinees' ability and reports the examinees' performance on the same scale (Embretson & Reise, 2000). Through IRT, CAT tailors a test for each individual examinee by taking into account the examinee's responses to previous items and selecting additional items that will most accurately discern and measure the examinee's ability.

Multiple-choice items are the most frequently used item format in CATs to date. This is due to the relative ease of developing and scoring multiple-choice items compared to other item formats (Haladyna, 1997). In addition, multiple-choice items tend to meet the assumptions of IRT, such as local independence and unidimensional latent trait (Hambleton & Swaminathan, 1985). However, a set of multiple-choice

items centered on a single stimulus, often referred to as a testlet, violates the assumption of local independence. This occurs because an examinee's response to one item within the testlet is impacted by an examinee's response to another item within the same testlet (Wainer & Kiely, 1987). The practice of using one stimulus for a group of items creates local dependence among the items.

Various ways have been proposed to handle testlet data within a CAT system. One commonly used approach is to ignore the dependency problem and use one of the unidimensional dichotomous IRT models. The problem with this approach is that the ability levels will be incorrectly estimated due to the inflation of item information (Wainer & Lewis, 1990). Another approach is to use a measurement model that takes the dependency into account. Polytomous IRT models handle the dependency problem by defining the testlet rather than the item within the testlet as the unit of measurement. This creates a polytomous item with a score ranging from 0 to the total number of items associated with the stimulus and eliminates the dependency problem (Wainer & Lewis, 1990).

Alternatively, one of the measurement models based on testlet response theory (TRT; Wainer, Bradlow, & Du, 2000) can be employed. In TRT, the item associated with a given testlet remains the unit of measurement. With TRT the most frequently used dichotomous IRT models, one-parameter, two-parameter, and three-parameter logistic models, have been modified to include a random effect parameter to account for the shared variance among items within a testlet, called the testlet effect parameter (Wainer, Bradlow, & Du, 2000). The *b*-, *a*-, and *c*-parameters of the

TRT models retain the same interpretations and meanings as with the dichotomous IRT models. By incorporating local dependence of items within a testlet into the model, the issue is no longer being ignored or sidestepped.

The precision of measurement of a CAT system is dependent not only on the measurement model on which it is based, but also the method of item exposure control that is selected. Exposure controls must balance the need for test security with precision of measurement. In unconstrained CATs, the most informative items are overexposed and threaten test security. Optimal utilization of the item pool for test security, however, means less informative items are given and the accuracy of the ability estimates is decreased. A number of exposure control procedures have been proposed to accommodate these two conflicting goals.

Previous research indicates that exposure control procedures seem to perform differently for dichotomously scored CAT systems versus polytomously scored CAT systems. For dichotomously scored CATs, conditional selection procedures appear to be the optimal choice (Chang, 1998), while randomization procedures perform best for polytomously scored CATs (Davis, 2002). Testlet scored CAT systems modeled with testlet response theory have not been examined to determine optimal exposure control procedures.

This dissertation investigates various exposure control procedures in CAT systems based on the three-parameter logistic testlet response theory (TRT) model and the partial credit (PC) model. The exposure control procedures are the randomesque procedure (Kingsbury & Zara, 1989), two levels of the progressive

restricted procedure (Revuelta & Ponsoda, 1998), two levels of the Sympson-Hetter procedure (Sympson & Hetter, 1985), the modified within .10 logits procedure (Davis & Dodd, 2001), and a maximum information procedure. Through realistic CAT simulations that include content balancing, this dissertation examines the viability of these exposure control procedures for testlet-based CATs.

CHAPTER TWO: LITERATURE REVIEW

This literature review provides background information pertaining to the current study. First, assumptions and characteristics of item response theory are explored. This leads to a discussion of common item response theory models for scoring dichotomous and polytomous items. Testlet response theory is then presented as a viable method for modeling items that are locally dependent. Next is a description of the dichotomous testlet response theory models. The next section discusses the components of a computerized adaptive testing (CAT) system. This includes a description and examination of current research for common exposure control procedures for both dichotomous and polytomous item response theory models. The final section is the statement of problem.

Item Response Theory

Item response theory (IRT) depicts the relationship between examinees and items through mathematical models (Wainer & Mislevy, 2000). IRT models the probability of a given response to an item conditional on ability (trait) level. Two common classes of IRT models are determined by the way item responses are scored. Items with only two response options (correct or incorrect) are modeled with the dichotomous IRT models. Multiple-choice items and true-false items are examples of items that can be scored dichotomously. Items with more than two response options can be modeled with polytomous IRT models. Examples of polytomously scored items are items that allow for partial credit scoring, such as a math problem or an essay item where partially correct solutions receive more points than incorrect

answers but fewer points than correct answers. Thus the response categories are ordered from low to high to represent varying amounts of the ability measured.

Assumptions

There are three main assumptions for unidimensional item response theory (IRT) models. One assumption for IRT models is that a mathematical function can be derived to model the probability of a given response to an item conditional on ability level (Hambleton & Swaminathan, 1985). The mathematical function contains item parameters (characteristics) that model the probability of a given response for each ability level.

The second assumption of unidimensional IRT models is a single ability underlies the difference in person responses to items (Embretson & Reise, 2000). For example, a math exam modeled with unidimensional IRT assumes that the ability, math, "accounts for the statistical dependence among the items" (Crocker & Algina, 1986). A test is unidimensional if the distribution of test scores, conditioned on ability, is identical (Hambleton & Swaminathan, 1985). There are models that assume more than one ability underlies examinees' responses to items. With multidimensional item response theory (MIRT) models, examinees' responses to items are explained by a weighted combination of the underlying abilities (Embretson & Reise, 2000; Reckase, 1997). Only unidimensional models are examined in this dissertation, therefore the MIRT models will not be discussed further.

The third assumption is that the items within the test have local independence such that the probability of responding to an item is statistically independent of the

probability of responding to any other item while conditioned on ability (Hambleton & Swaminathan, 1985). For example, the content provided in one item should not aide an examinee in answering any other items. Stated differently, local independence is present when the probability of an examinee's response pattern equals the product of each item's probability given the examinee's response (Hambleton & Swaminathan, 1985). The presence of local independence specifies that together, the IRT model and the parameters, fully explain the relationship between items (Embretson & Reise, 2000). Through local independence, item parameter invariance allows examinees' ability to be estimated from any subset of items and yields examinees' performance levels on the same ability scale. This is critical to the adaptive nature of computerized adaptive tests, whereby examinees may be given different items from a group of items that were calibrated together and the examinee's estimated abilities will be on the same scale (Embretson & Reise, 2000).

Measurement Models for Item Response Theory

The following sections discuss the most commonly used dichotomous and polytomous item response theory (IRT) models. The dichotomous IRT models selected for presentation are the three-parameter logistic model, the two-parameter logistic model and the one-parameter logistic model. The polytomous IRT models include the graded response model, the partial credit model, and the generalized partial credit model.

Dichotomous Item Response Theory Models

Often the most commonly employed IRT models in practical applications are the dichotomous models due to current trends toward multiple-choice item test formats. Multiple-choice items are relatively easy to develop and score compared to other item formats (Haladyna, 1997). For example, the Armed Services Vocational Aptitude Battery (ASVAB) CAT was developed for the Department of Defense using the three-parameter logistic IRT model due to its "superior accuracy in modeling response probabilities of multiple choice test questions" (Segall & Moreno, 1999).

The dichotomous IRT models can be characterized based on the number of item parameters included in the model, which is reflected in the names of the models. The three common item parameters are difficulty (*b)*, discrimination (*a)*, and the psuedo-guessing parameter (*c)*. The three-parameter logistic IRT model is the most general of the dichotomous IRT models. The two-parameter logistic IRT model is a mathematical simplification of the three-parameter model and the one-parameter logistic IRT model is a mathematical simplification of the two-parameter IRT model. The following sections describe the three-parameter, two-parameter, and one-parameter logistic IRT models.

The three-parameter logistic IRT model has three item parameters: the difficulty parameter, b, the discrimination parameter, a, and the psuedo-guessing parameter, c, (Birnbaum, 1968). For the three-parameter logistic IRT model, the probability of success ($x = 1$) for person j with an ability level, θ, on item i is denoted

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

where b_i is the difficulty parameter for item *i*, a_i is the discrimination parameter for item *i*, and c_i is the psuedo-guessing parameter for item *i*.

The two-parameter logistic IRT model has two item parameters: the difficulty parameter, *b*, and the discrimination parameter, *a* (Birnbaum, 1968). For the two-parameter logistic IRT model, the probability of success (*x* =1) for person *j* with an ability level, θ, on item *i* is denoted

$$P_{ij}(x_i = 1 \mid \theta_j) = \frac{\exp(a_i(\theta_j - b_i))}{1 + \exp(a_i(\theta_j - b_i))} \quad , \qquad (2)$$

where b_i is the difficulty parameter for item *i* and a_i is the discrimination parameter for item *i*. The two-parameter logistic IRT model assumes that guessing does not exist.

The one-parameter logistic IRT model, also known as the Rasch model, estimates a person's ability based on the person's responses to items that have been calibrated for one item parameter (Rasch, 1960; Wright, 1968). It is the most parsimonious of the IRT models. The difficulty parameter, *b*, is the item parameter included in the model. For the one-parameter logistic IRT model, the probability of success (*x* = 1) for person *j* with an ability level, θ, on item *i* is denoted

$$P_{ij}(x_i = 1 \mid \theta_j) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} \quad , \qquad (3)$$

where b_i is the difficulty parameter for item *i*. The one-parameter logistic IRT model assumes that the items in the test discriminate equally well and that guessing is nonexistent.

The item parameters can be defined in relation to the item characteristic curve (ICC) for the dichotomous IRT models. The ICC is a monotonically increasing curve such that as ability increases the probability of obtaining a correct response to an item also increases. The form of the ICC is dependent on the measurement model. The ICC provides a graphical representation of the probability of a correct response to an item conditional on the ability level of the examinee (Embretson & Reise, 2000). Figure 1 shows an example of an ICC plot for the three-parameter logistic model (Birnbaum, 1968) with ability on the abscissa and the probability of a correct response on the ordinate. For the ICC in Figure 1, the difficulty parameter, *b*, equals 0.50. The discrimination parameter, *a*, equals 1.5. And the psuedo-guessing parameter, *c*, equals 0.15.

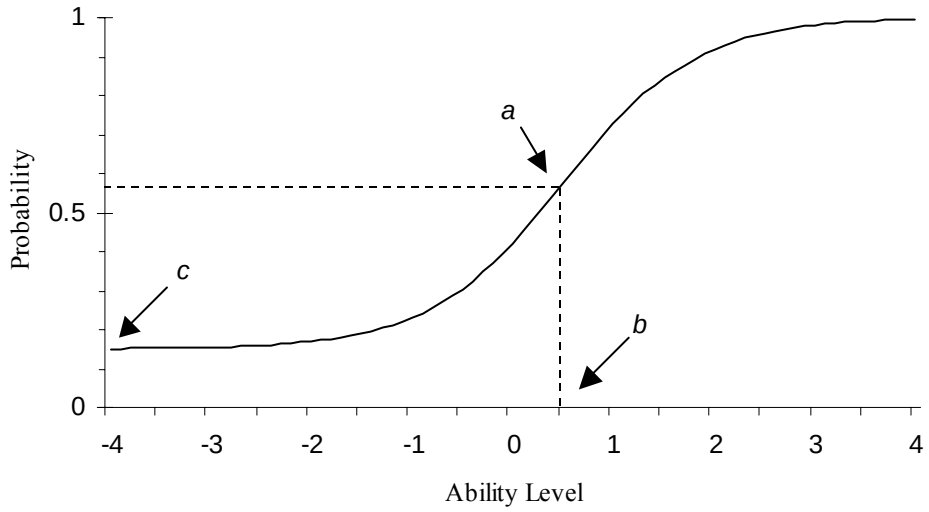

FIGURE 1: Item Characteristic Curve for the Three-Parameter Logistic Model

The difficulty parameter identifies the relative easiness of an item and places it on the same scale as ability (θ), typically ranging from -4 to +4. In terms of the ICC, the difficulty parameter is the point on the ability scale that reflects the maximum for the slope of the ICC (Hambleton & Swaminathan, 1985). For the three-parameter logistic IRT model, the maximum of the slope is where $p = (1 + c)/2$. For the two-parameter and one-parameter logistic IRT models, the maximum of the slope is at $p = 0.50$, since c equals zero.

The discrimination parameter indicates how well an item distinguishes low ability examinees from high ability examinees. The discrimination parameter ranges from zero to infinity, but in practical terms, the discrimination value ranges from zero to about four. The slope of the ICC in Figure 1 is related to the discrimination

parameter such that at *b* the slope equals 0.425**a*(1 - *c*). Therefore, the steeper the slope and the higher the discrimination parameter, the better the item is at distinguishing between changes in ability level around the difficulty parameter (Hambleton & Swaminathan, 1985).

The psuedo-guessing parameter may be included for items in which examinees may obtain a correct response due to chance rather than skill. This parameter is on the same scale as the probability. In Figure 1, the guessing parameter is the location of the lower asymptote. When guessing is not included in the model, the lower asymptote is at zero on the probability scale (Hambleton & Swaminathan, 1985).

An item information function specifies the precision of measurement that an item provides for each ability level (Embretson & Reise, 2000). Since items do not measure all ability levels with equal precision, information is not consistent across the ability scale. For dichotomous IRT models, the item information function, $I_i(\theta)$, is denoted

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)(1 - P_i(\theta))} \quad , \qquad (4)$$

where $P_i(\theta)$ is the probability of a correct response to item *i* conditioned on ability, θ, and $P_i'(\theta)$ is the first derivative with respect to ability, θ, (Embretson & Reise, 2000).

Test information, $TI(\theta)$, is the sum of the item informations. This additive property is due to local independence among items. Test information, $TI(\theta)$, is denoted

$$TI(\theta) = \sum_{i=1}^{I} I_i(\theta) \quad . \tag{5}$$

Information can be used in CAT systems to select an item for administration and provide measurement precision through the standard error associated with a given ability, θ.

Measurement precision for a test can be evaluated through the standard error associated with a given ability, θ, which is the square root of the reciprocal of the test information:

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}} \quad . \tag{6}$$

The standard error associated with a given ability, θ, is not necessarily constant across the ability continuum. Often the standard error will be higher at the extremes of the ability continuum, indicating the lack of information provided by the IRT model and its parameters (Embretson & Reise, 2000).

Polytomous Item Response Theory Models

Polytomous item response theory (IRT) models allow examinees to obtain credit for knowing part of an answer, if not all of an answer. These models distinguishes between examinees with no knowledge and examinees with varying degrees of knowledge. For example, an essay is typically scored using a rubric that

awards more points for better responses. Similarly, several multiple-choice items referring to a single stimulus might be scored polytomously to assess the number of items answered correctly for that stimulus. The following sections describe three of the most commonly used polytomous IRT models: the graded response model, the partial credit model and the generalized partial credit model.

Thissen and Steinberg (1986) organized polytomous models into three categories: the difference models, the divide-by-total models, and the left-side added models. The polytomous models discussed in this dissertation come from two of these categories. The graded response model is characterized as a difference model because the probability of an examinee receiving a category score is determined by calculating the difference between two successive category probabilities (Thissen & Steinberg, 1986). The partial credit model and the generalized partial credit model are described as divide-by-total models (Thissen & Steinberg, 1986). The divide-by-total models are calculated by dividing the probability of attaining a specific category score by the sum of all allowable probabilities of attaining a category score for that item (Thissen & Steinberg, 1986; Dodd, De Ayala, & Koch, 1995).

Graded Response Model

Samejima's (1969) graded response (GR) model is appropriate for items whose responses are ordered to indicate an examinee's level of knowledge. The item response, x, ranges from 0 to m_i (the total number of response options) such that lower values reflect less knowledge of the correct response to the item and higher values reflect more knowledge of the correct response to the item. Samejima (1969)

developed a two-stage process to determine the probability of responding in a particular category. The first stage consists of determining the probability that an examinee with an ability level, θ, will obtain a category score of x or higher on item i. This is denoted

$$P_{ix}^{*}(\theta) = \frac{\exp[a_i(\theta - b_{ix})]}{1 + \exp[a_i(\theta - b_{ix})]} \quad , \tag{7}$$

where a_i represents the discrimination parameter for item i and b_{ix} represents the category boundary for item i between category score x and category score x - 1. The category boundaries for item i are located on the ability scale that reflects the point of inflection of the category characteristic curves ($P_{ix}^{*}(\theta)$). Figure 2 shows the category characteristic curves for an example item with four ordered responses, x = 0, 1, 2, or 3, and three category boundaries, b_{i1} to b_{i3}. The category boundaries, b_{ix}, must be sequential on the ability scale from low to high. The category boundary, b_{ix}, is defined as the ability level, θ, that corresponds to $P_{ix}^{*}(\theta) = 0.5$

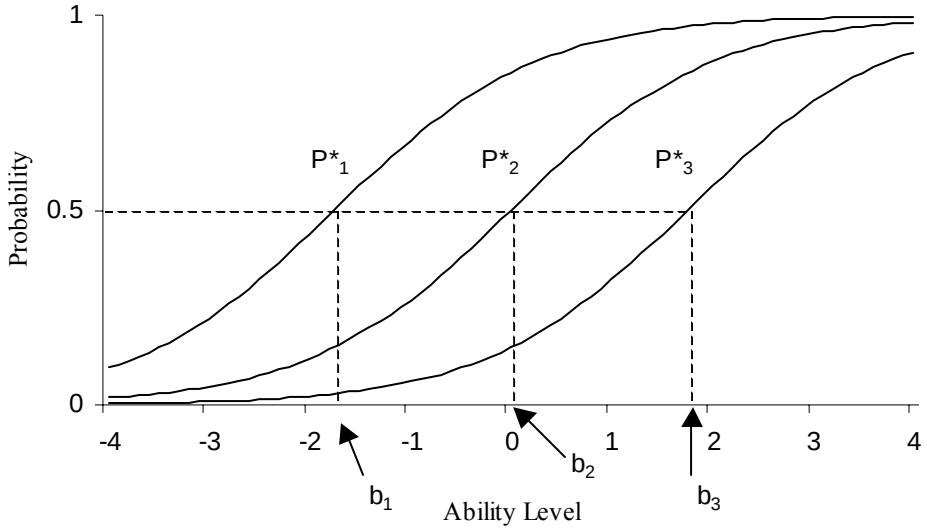

FIGURE 2: Graded Response Model Category Characteristic Curves

The second stage determines the probability of responding in a specific category by subtracting adjacent category characteristic curves. The probability of a specific category score, $P_{ix}(\theta)$, is denoted

$$P_{ix}(\theta) = P^{*}_{ix}(\theta) - P^{*}_{i,x+1}(\theta) \quad . \tag{8}$$

For the extreme categories where $x = 0$ or $x = m$, the probability of responding in category x or higher is $P^{*}_{i0}(\theta) = 1.0$ and $P^{*}_{i,m}(\theta) = 0.0$, respectively. The graded response model simplifies to the two-parameter logistic IRT model when there are only two response categories (0,1). Figure 3 shows the operating characteristic curves, which illustrate the probability of obtaining a specific category score, x, for the same item in Figure 2 that has four response categories and three category boundaries.

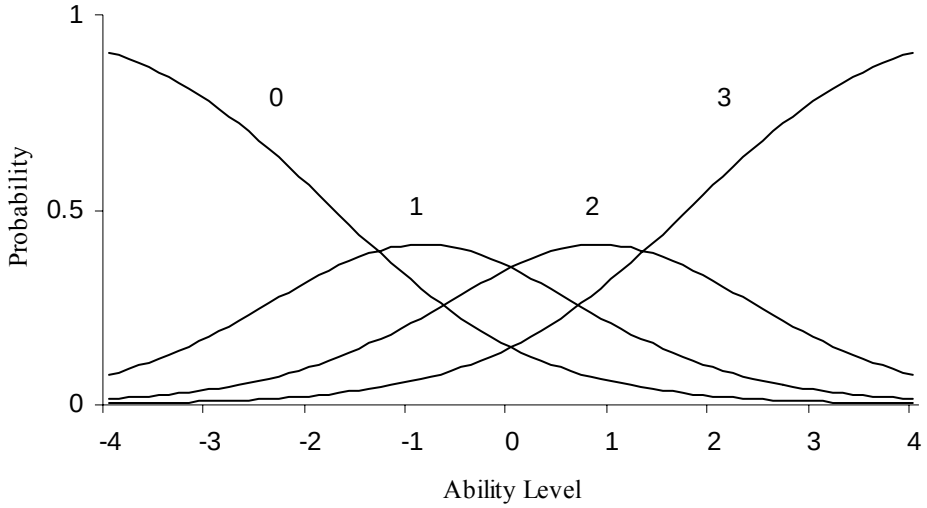

FIGURE 3: Graded Response Model Operating Characteristic Curves

Partial Credit Model

Like the graded response model, Masters' (1982) partial credit (PC) model is appropriate when responses to an item can be scored into more than two categories to represent varying degrees of the ability measured by the item. Thus for each item *i,* a person's item score will be categorized in one of $m_i + 1$ category scores, ranging from 0 to m_i. For the PC model, the probability that a person with an ability level, θ, will obtain a score of x on item i is denoted

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

where b_{ik} represents the step difficulty or threshold of transitioning from one category of m_i to the next category. For notational purposes, $\sum_{k=0}^{0}(\theta - b_{ik}) = 0$. The PC model step difficulties do not have to be in sequential order, as do the category boundaries for the GR model. Earlier step values may be more difficult than later step values. The PC model is an extension of the Rasch model to polytomously scored items and assumes that items within a given test do not differ in their discrimination level and guessing is not a factor (Masters, 1982).

Figure 4 presents operating characteristic curves for a partial credit item with four response options and three step difficulties. The step difficulties correspond to the ability level where the operating characteristic curves of adjacent categories intersect.

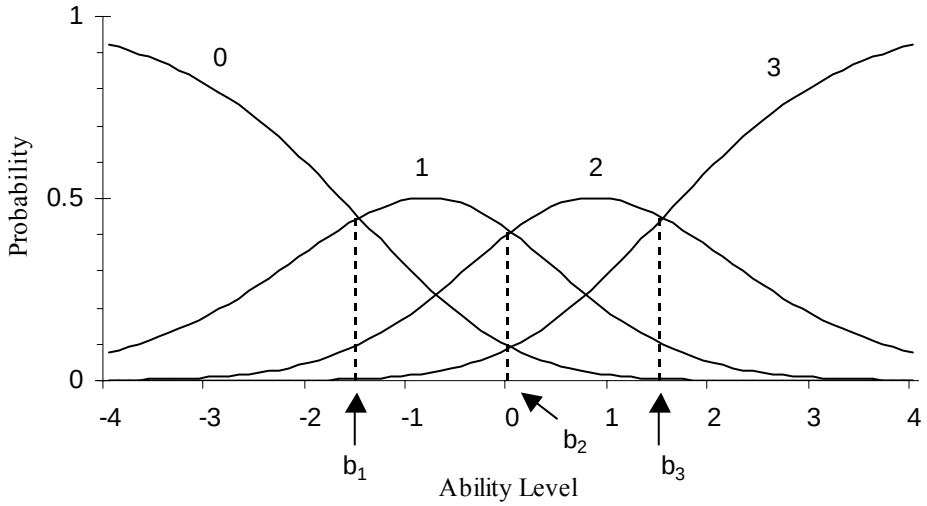

FIGURE 4: Partial Credit Model Operating Characteristic Curves

Generalized Partial Credit Model

Muraki (1992) developed the generalized partial credit (GPC) model by extending the PC model to allow items to vary in their level of discrimination. For the GPC model, the probability that a person with an ability level, θ, will obtain a score of x on item i is denoted

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

where a_i represents item discrimination and b_{ik} represents the step difficulty or threshold of transitioning from one category of m_i to the next category. Similarly to the GR model, the GPC model has one discrimination parameter per item and simplifies to the two-parameter logistic IRT model when there are only two response categories (0,1; Muraki, 1992).

Information Functions

Samejima (1969) developed a general formula for calculating information for polytomous models. Not only can information be determined for an item and a test, information can also be calculated for the response categories. The category information, $I_{ix}(\theta)$, for item i is denoted

$$I_{ix}(\theta) = \frac{[P'_{ix}(\theta)]^2}{P_{ix}(\theta)} - \frac{P''_{ix}(\theta)}{P_{ix}(\theta)}, \qquad (11)$$

where $P_{ix}(\theta)$ is the probability of responding in a given category, *x*, for item *i* conditioned on ability, θ, $P'_{ix}(\theta)$ is the first derivative with respect to ability, θ, and $P''_{ix}(\theta)$ is the second derivative with respect to ability, θ, (Koch & Dodd, 1989).

The item information, $I_i(\theta)$, for item *i* is denoted

$$I_i(\theta) = \sum_{x_i=0}^{m_i} I_{ix}(\theta) P_{ix}(\theta) \quad , \tag{12}$$

By substituting Equation 11 for $I_{ix}(\theta)$ in Equation 12 and simplifying, the item information, $I_i(\theta)$, for item *i* can be written

$$I_i(\theta) = \sum_{x=0}^{m_i} \frac{[P'_{ix}(\theta)]^2}{P_{ix}(\theta)} - \sum_{x=0}^{m_i} P''_{ix}(\theta) \quad , \tag{13}$$

Samejima (1969) demonstrated that in Equation 13 the second term equals zero and can be removed from the equation.

Test information, $TI(\theta)$, is denoted

$$TI(\theta) = \sum_{i=1}^{I} I_i(\theta) \quad . \tag{14}$$

Measurement precision for a test can be evaluated through the standard error associated with a given ability, θ, which is the square root of the reciprocal of the test information:

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}} \quad . \tag{15}$$

The standard error associated with a given ability, θ, is not necessarily constant across the ability continuum.

Testlet Response Theory

Testlets are defined as a group of items that relate to a single stimulus, such as a reading passage or graphic (Wainer & Kiely, 1987). Since the items have a stimulus in common, they are no longer independent of each other. This violates the assumption of local independence for dichotomous item response theory models. Several methods are available to determine if local item dependence (LID) is present among items in a test. The following presents two forms of measuring LID: the Q_3 statistic and the G^2 statistic.

Yen (1984) applied the Q_3 statistic as a measure of LID. The Q_3 statistic is the correlation between performance on two items, after accounting for an examinees' overall performance on a test. Based on the *j*th examinee's responses to *i* items, an ability estimate, $\hat{\theta}_j$, is calculated and used to determine an examinee's expected performance on the *i*th item, E_{ij}. Based on the examinee's observed and expected performance on each item, a deviation, d_{ij}, value is calculated:

$$E_{ij} \equiv E(X_i \mid \hat{\theta}_j) = \sum_{k=1}^{m_i} (k-1) P_{ik}(\hat{\theta}_j) \quad \text{, and} \tag{16}$$

$$d_{ij} = x_{ij} - E_{ij} \quad . \tag{17}$$

The deviations are correlated across examinees to obtain a measure of LID for items *i* and i',

$$Q_{3ii'} = r(d_i, d_{i'}) \quad . \tag{18}$$

When local independence is present, the expected value for Q_3 is approximately

–1/(n-1).

Chen and Thissen (1997) employed the G^2 statistic to measure LID. This statistic has a χ^2 distribution with one degree of freedom. The G^2 statistic is calculated as:

$$G^2 = -2\sum_{i=1}^{2}\sum_{i'=1}^{2} O_{ii'} \ln\left(\frac{E_{ii'}}{O_{ii'}}\right) \quad . \qquad (19)$$

When compared to the Q_3 statistic, the G^2 statistic did not perform as well in identifying LID (Chen & Thissen, 1997).

The dichotomous item response theory models cannot account for the common variance created by locally dependent items. Previous studies have shown that modeling data that have local dependence with dichotomous IRT models yields an overestimate of the precision of measurement (Sireci, Wainer, & Thissen, 1991; Yen 1993). The overestimation is because the probability of two or more dependent events occurring is less than the probability of two or more independent events occurring (Devore, 1995). During computerized adaptive testing, overestimation of the precision of measurement may lead to early termination of the test (Fennessy, 1995).

The polytomous item response theory models consider the testlet as the unit of measurement. The testlet (a polytomous item) is scored from zero to the total number of items associated with the common stimulus. By changing the unit of measurement from the item to the testlet, polytomous models account for the

dependencies across the items in the testlet. This has been shown to be an effective method (Wainer, 1995).

Yet, there are a couple of reasons why it is advantageous to maintain the item as the unit of measurement. One reason pertains to CAT implementations. The nature of CATs is to allow items to be selected adaptively. With polytomously scored testlets, there cannot be an interchange of items within the testlet. If the item is the unit of measurement, then items can be selected adaptively within the testlet. By varying the items, this would increase the number of times the stimulus could be used across examinees. The second reason relates to the information gained from the examinee's responses. The polytomous models provide a number correct score for each testlet, but do not include information about the pattern of responses. Knowing exactly which items the examinee answered correctly and incorrectly could prove beneficial, especially if the items varied by cognitive type or content (Wainer, Bradlow, & Du, 2000).

Alternatively, one of the measurement models based on testlet response theory (TRT; Wainer, Bradlow, & Du, 2000) can be employed. In TRT, the item associated with a given testlet remains the unit of measurement. With TRT the most frequently used dichotomous IRT models have been modified to include a random effect parameter to account for the shared variance among items within a testlet, called the testlet effect parameter. The *b*-, *a*-, and *c*-parameters of the TRT models retain the same interpretations and meanings as with the dichotomous IRT models.

By incorporating local dependence of items within a testlet into the model, the issue is no longer being ignored or sidestepped.

Measurement Models for Testlet Response Theory

The testlet response theory models (Wang, Bradlow, & Wainer, 2002) are dichotomous models with as many as three item parameters, difficulty (*b*), discrimination (*a*), and psuedo-guessing (*c*) parameters; and two person-specific parameters, ability (θ) and the testlet effect ($\gamma_{jd(i)}$). The item parameters are interpreted the same for the TRT models as they are for the corresponding dichotomous IRT models. The additional parameter that is person-specific is the testlet effect parameter which accounts for violations of the local independence assumption for dichotomously scored items modeled with IRT. The testlet effect parameter models the local dependency by including the same random effect for each item within a testlet. This common parameter across items accounts for the communality created by the items' association with the same stimulus. Independent items, those not associated with a common stimulus, included in the TRT models will have a testlet effect equal to zero, thereby defaulting to the dichotomous IRT model counterpart.

The testlet effect parameter can also be used to measure local item dependence (LID). When items are calibrated for the TRT model, the parameter estimate retrieved is the estimated variance of the testlet effect. The variance of the testlet effect is a measure of LID. Testlets that contain independent items will have

variance estimates of zero. The testlets that contain dependent items will have nonzero variance estimates. The testlet effect parameter used in the TRT model is a random variable selected from a normal distribution with mean zero and standard deviation equal to the square root of the variance of the testlet effect for a given testlet.

Dichotomous Testlet Response Theory Models

Starting with the most general model, the three-parameter logistic TRT model, the following sections demonstrate the inclusion of the testlet effect parameter into the dichotomous item response theory models to create the testlet response theory models.

Wainer, Bradlow, and Du (2000) introduced the three-parameter logistic testlet response theory model (3PL-TRT). For the 3PL-TRT model, the probability of success ($x = 1$) on item i for person j with an ability level, θ, on testlet $d(i)$ is denoted

$$P_{ij}(x_i = 1 \mid \theta_j) = c_i + (1 - c_i)\left[\frac{\exp(a_i(\theta_j - b_i - \gamma_{jd(i)}))}{1 + \exp(a_i(\theta_j - b_i - \gamma_{jd(i)}))}\right], \qquad (20)$$

where the testlet effect parameter $\gamma_{jd(i)}$ models the extra dependency for person j responding to item i that is nested in testlet $d(i)$, b_i is the difficulty parameter for item i, a_i is the discrimination parameter for item i, and c_i is the psuedo-guessing parameter for item i.

The performance of the 3PL-TRT model was compared through four simulation conditions: the 3PL-IRT model using marginal maximum likelihood

estimation, the 3PL-IRT model using Markov Chain Monte Carlo estimation, the 3PL-TRT model with the same testlet effect parameter for each testlet using Markov Chain Monte Carlo estimation, and the 3PL-TRT model allowing the testlet effect parameter to vary across testlets using Markov Chain Monte Carlo estimation. The models were fit to three data sets: a no testlet effect, an equal testlet effect, and an unequal testlet effect data set. The data sets were based on tests with 30 independent items and 4 testlets with 10 items each (40 dependent items). The simulations yielded similar ability correlations across the models for the no testlet effect data set. Yet, for the other two conditions with testlet effects in the data sets, the 3PL-TRT model with equal effects and the 3PL-TRT model with varying effects reported higher correlations between the true and estimated parameters: ability, discrimination, difficulty, and psuedo-guessing. Similar results were reported for the models at each level of the conditions in terms of mean absolute error and relative efficiency. Overall, when data exhibited local dependency, the TRT models performed better than the IRT model.

Bradlow, Wainer, and Wang (1999) modified the two-parameter logistic item response theory model (Birnbaum, 1968) to account for violations of local independence as a result of multiple items referring to a similar stimulus. For the two-parameter logistic testlet response theory model (2PL-TRT), the probability of success ($x = 1$) on item i for person j with an ability level, θ, on testlet $d(i)$ is denoted

$$P_{ij}(x_i = 1 \mid \theta_j) = \frac{\exp(a_i(\theta_j - b_i - \gamma_{jd(i)}))}{1 + \exp(a_i(\theta_j - b_i - \gamma_{jd(i)}))} \quad , \tag{21}$$

where the testlet effect parameter $\gamma_{jd(i)}$ models the extra dependency for person *j* responding to item *i* that is nested in testlet *d(i)*, b_i is the difficulty parameter for item *i*, and a_i is the discrimination parameter for item *i*.

Bradlow et al. (1999) examined a 2PL-IRT model analyzed with BILOG (Mislevy & Bock, 1983) and a Data Augmented Gibbs Sampler (DAGS; Tanner & Wong, 1987) with a 2PL-TRT model analyzed with DAGS. The 2PL-TRT model yielded lower mean absolute errors and higher correlation coefficients for ability, discrimination, and difficulty compared to the 2PL-IRT models when the data sets contained testlets with local dependencies.

For the one-parameter logistic testlet response theory model (1PL-TRT), the probability of success ($x = 1$) on item *i* for person *j* with an ability level, θ, on testlet *d(i)* is denoted

$$P_{ij}(x_i = 1 \mid \theta_j) = \frac{\exp(\theta_j - b_i - \gamma_{jd(i)})}{1 + \exp(\theta_j - b_i - \gamma_{jd(i)})} \quad , \qquad (22)$$

where the testlet effect parameter $\gamma_{jd(i)}$ models the extra dependency for person *j* responding to item *i* that is nested in testlet *d(i)* and b_i is the difficulty parameter for item *i*.

The item information, $I_i(\theta)$, for the more general model, the three-parameter logistic testlet response theory model, conditional on theta for a single item response is denoted

$$I_i(\theta) = a_i^2 \left(\frac{\exp(a_i(\theta_j - b_i - \gamma_{jd(i)}))}{1 + \exp(a_i(\theta_j - b_i - \gamma_{jd(i)}))} \right)^2 \frac{1 - c_i}{c_i + \exp(a_i(\theta_j - b_i - \gamma_{jd(i)}))}, \qquad (23)$$

(Wainer, Bradlow, & Du, 2000).

Testlet information, $I_T(\theta)$, is the sum of the item informations within a testlet:

$$I_T(\theta) = \sum_{i=1}^{I} I_i(\theta) \qquad . \qquad (24)$$

Test information, $TI(\theta)$, is the sum of the testlet informations:

$$TI(\theta) = \sum_{t=1}^{T} I_t(\theta) \qquad . \qquad (25)$$

Information can be used in CAT systems to select an item for administration and provide measurement precision through the standard error associated with a given ability, θ. Measurement precision for a test can be evaluated through the standard error associated with a given ability, θ, which is the square root of the reciprocal of the test information:

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}} \qquad . \qquad (26)$$

The standard error associated with a given ability, θ, is not necessarily constant across the ability continuum.

Computerized Adaptive Testing Systems

The goals of a computerized adaptive test (CAT) are to accurately assess examinees with items tailored to each examinee's ability; to evaluate examinees

equivalently on content specifications; and to maintain test security through controlling the exposure of items to examinees (Davey & Parshall, 1995). With traditional paper-and-pencil tests, examinees encounter items of varying difficulty levels. The benefit of item response theory (IRT) and testlet response theory (TRT) models is the administration of fewer items targeted to the examinee's ability. CATs reduce examinees' frustration and/or test anxiety caused by the administration of items that are too easy or too difficult. Wainer (2000) described the concept of a CAT, as "the basic notion of an adaptive test is to mimic automatically what a wise examiner would do." By administering items that match the examinee's ability level, the result is often a more precise measurement of ability and a shorter test (McBride & Martin, 1983; Urry, 1977).

The psychometric properties of IRT and TRT enable accurate estimation of examinee's proficiency, but may not meet content specifications of the test. Therefore, constraints are often used to ensure examinees are assessed fairly and equally over the material covered in the test. These constraints often lead to a reduction in the precision of measurement for estimating examinees' proficiencies because less optimal items are administered.

Additionally, the item pool must be protected to ensure administration of fair and valid tests. Administration of the most optimal items, especially early in the CAT algorithm, often leads to the same items being repeatedly administered across examinees. This overexposure of items may compromise the item pool. To ensure the security of the item pool, constraints restrict the selection and administration of items

to examinees. When developing a CAT system, these competing goals lead to a balancing act in which the testing companies must decide the priority of precision, content, and security (Davey & Parshall, 1995; Davey & Nering, 2002).

There are four main components of a CAT system: the item pool, the item selection procedure, the ability estimation, and the stopping rule (Dodd, De Ayala, & Koch, 1995; Reckase, 1989). Due to practical considerations mentioned previously, two additional components are also included, content balancing and exposure control. The following sections describe the six components of a CAT system.

Item Pool

The item pool or item bank consists of all the items that may be administered during the test and the items' parameters (characteristics). The item parameters included in the pool are dependent upon the IRT/TRT model selected to model the data and to measure the examinees' ability levels. An item pool will have many more items than a single paper-and-pencil test administration. Ideally, there will be enough items to generate multiple test forms for a range of examinee abilities (Davey & Nering, 2002). Similar to the items on paper-and-pencil tests, CAT items within the pool must meet sensitivity and psychometric standards. For sensitivity standards, items must not function differently based on examinee characteristics other than the ability measured by the test, such as gender or ethnicity.

Psychometrically, items are evaluated based on their item parameters and item information across the ability scale. The desired distribution of items across the ability scale varies based on the purpose of the test. For achievement tests, the item

pool generally contains a range of items from very easy to very difficult. Ideally the distribution of difficulty would be uniform, unlike item difficulty distributions for paper-and-pencil tests, which tend to be normally distributed with most of the items grouped around the mean of the distribution. For criterion-referenced tests, most of the items in the item pool have difficulty values that provide the most information around the cut point. In this circumstance, the goal is to discriminate between examinees above the cut point and those below the cut point. Flaugher (2000) noted "the better the quality of the item pool, the better the job the adaptive algorithm can do."

Item Selection Procedure

For CATs, the item selection procedure is the process of selecting an item from the item pool to be administered to the examinee. The item selection procedure may be used to select each item individually or a group of items together, depending on the test format. An examinee's current estimated ability plays a key role in determining which item will be selected next. Once an item has been selected and administered, it is tagged so the item selection procedure does not administer it to that examinee again. The two most frequently used item selection procedures are maximum information selection and Bayesian selection procedures.

Maximum Information Selection

One of the more common item selection procedures is maximum information selection (Birnbaum, 1968; Lord, 1977). An item is selected if it provides the most information based on the examinee's current estimate of ability. The process involves

using the examinee's estimated ability based on their responses to the previously administered items to determine the amount of information that would be provided by each of the items remaining in the item pool. The item that provides the most information is then selected for administration. Usually, this process is repeated after the administration of each item. The initial item for administration is often selected with an assumed examinee ability level at the mean of the distribution.

Bayesian Selection

Bayesian selection (Owens, 1969) evaluates each item in the item pool to determine the expected variance of the posterior distribution. The item that minimizes the expected variance is selected for administration. A new prior distribution is calculated by including the recently administered item in the likelihood. This prior distribution is evaluated for each remaining item in the item pool to determine which item minimizes the expected variance of the new distribution. This procedure is computationally easier than maximum information. A disadvantage of the Bayesian selection is the order in which items are administers impacts the estimation of ability, $\hat{\theta}$ (Thissen & Mislevy, 2000).

Ability Estimation

Examinees' performance on a test is scored based on their responses to items, the items' characteristics, and the IRT/TRT model used to fit the data. When estimating ability, $\hat{\theta}$, the item parameters are assumed to be known values. The estimation of an examinee's ability is performed at two stages during a CAT

administration. After an examinee responds to an item, an interim estimate of ability is calculated and used by the item selection procedure to select the next item for administration. Usually interim ability estimation is calculated after each item administration, although for testlets scored polytomously or by TRT, all the items within the testlet can be administered before the interim ability estimation is calculated. The second stage for ability estimation is at the end of the CAT administration. After all the items have been administered, a final estimate of ability is determined based on the examinee's responses to all the items. The estimation procedure for estimating the interim abilities does not have to be the same as the estimation procedure for the final ability (Chang, Ansley, & Lin, 2000; Parshall, Hogarty, & Kromrey, 1999). The next two sections describe two common procedures for estimating ability, maximum likelihood estimation and expected a posteriori estimation.

Maximum Likelihood Estimation

The most commonly used estimation procedure is maximum likelihood estimation (MLE; Lord 1980). Based on an examinee's responses to items, MLE finds the ability value, θ, that maximizes the likelihood of an item response pattern, ($x_1, x_2, \ldots, x_i$). The likelihood of the ability value, θ, given an item response pattern, ($x_1, x_2, \ldots, x_i$), is denoted

$$L(x_1, x_2, \ldots, x_I | \theta) = \prod_{i=1}^{I} P_{i_{x_i}}(\theta), \qquad (27)$$

where $P_i(\theta)$ represents the probability of a given response to item *i* and item *i* is the number of items administered during a CAT.

Due to the diminutive nature of multiplying numbers between zero and one (Embretson & Reise, 2000), the natural log of the likelihood function, ln $L(\theta)$, is used in further calculations. By taking the log of the likelihood function, the log of the probability of an item response is summed across items. The same ability value, θ, maximizes both the likelihood function, $L(\theta)$, and the log of the likelihood function, ln $L(\theta)$. To determine the ability value, θ, which maximizes the log of the likelihood, the first derivative with respect to θ is set equal to zero and solved for the unknown θ. The maximum likelihood estimate of θ is obtained because the first derivative with respect to θ is the slope of the log of the likelihood function, ln $L(\theta)$. A slope equal to zero represents the highest point of the distribution. Solving the log likelihood function for its first derivative with respect to θ is denoted

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \sum_{i=1}^{I} [x_{si} - P_i(\theta)] \frac{P_i^{'}(\theta)}{P_i(\theta) Q_i(\theta)} = 0 \quad , \qquad (28)$$

(Embretson & Reise, 2000).

The Newton-Raphson procedure is applied due to the inability to solve Equation 24 directly. This is an iterative procedure in which the ratio of the first derivative to the second derivative is subtracted from the previous $\hat{\theta}$ resulting in a new $\hat{\theta}$. This procedure is repeated using the new $\hat{\theta}$ and calculating a new ratio of the

derivatives until the ratio reaches a pre-specified small value, such as 0.001 (Embretson & Reise, 2000).

Due to computation problems with the likelihood function when item responses are either all correct or all incorrect, variable step-size estimation is employed until there is one correct and one incorrect response for dichotomous models or two different response categories for polytomous models. Variable step-size estimation assigns an examinee's interim ability estimate to be half the distance between the current ability estimate and a maximum or minimum item difficulty value depending on whether all the responses are in the upper or lower half of the response scale (Koch & Dodd, 1989). In addition, Bock and Mislevy (1982) noted that MLE can yield extreme outliers when examinee's responses are in an abnormal pattern.

Expected a Posteriori Estimation

Expected a posteriori (EAP) estimation is easier to calculate than MLE estimation and does not require an iterative process (Bock & Mislevy, 1982). Bayesian in nature, EAP estimation represents the mean of the posterior distribution. For the *i*th item in an adaptive test, the EAP estimate of ability, $\hat{\theta}_i$, based on the item response string ($x_1, x_2, \ldots, x_i$), is approximated by

$$\hat{\theta}_i = \frac{\sum_{k=1}^{q} Z_k L_i(Z_k) W(Z_k)}{\sum_{k=1}^{q} L_i(Z_k) W(Z_k)}, \tag{29}$$

where q represents the number of quadrature points, Z_k represents one of the quadrature points, $L_i(Z_k)$ is the likelihood of Z_k given the item response string (x_1, x_2, ..., x_i), and $W(Z_k)$ represents the quadrature weight for that point (Bock & Mislevy, 1982). The weights at the various quadrature points represent a discrete prior distribution. Multiplying the prior distribution by the likelihood function and summing across the quadrature points produces a posterior distribution in which the mean represents the estimated ability, $\hat{\theta}_i$.

Unlike MLE, EAP estimation provides an ability estimate even if the examinee responses are all in the same category (Bock & Mislevy, 1982; Embretson & Reise, 2000). In addition, EAP estimation is not impacted by the order in which items are administered like Owen's (1969) Bayes procedure for estimation or by abnormal response patterns (Bock & Mislevy, 1982). Bock and Mislevy (1982) reported that EAP yields smaller mean square errors than MLE when the population ability distribution matches that of the prior distribution. Glas, Wainer, and Bradlow (2000) applied EAP estimation of the testlet response theory model. The results indicated a loss in measurement precision if the testlet effect parameter was not included in the EAP estimation of data containing local dependencies. EAP does have some shortcomings including the tendency for Bayesian estimates to regress toward the mean of the prior distribution (Kim & Nicewander, 1993; Weiss, 1982). Additionally, an inappropriate prior distribution reduces the accuracy of the EAP estimation (Mislevy & Stocking, 1989; Seong, 1990).

Stopping Rule

Due to the adaptive nature of CATs and the properties of IRT/TRT, termination of the test may be based on the number of items administered (fixed length), the precision of measurement (variable length), or a combination of both. There are advantages and disadvantages to both procedures that are discussed in the following sections.

Fixed Length

A fixed length stopping rule administers a pre-specified number of items. Once these items are administered the examination is terminated and the examinee's proficiency is estimated. The advantages to a fixed length test include that it is easier to determine whether to administer another item or complete the test by a simple count of the number of items previously administered. In addition, fixed length tests are easier for examinee's to understand. If two examinees take the same CAT and receive different scores, the examinee that took fewer items may not feel he/she was properly assessed. A disadvantage of fixed length tests is the inability to obtain the same level of measurement precision for the range of examinees' abilities. Often the examinees at the extremes of the ability distribution are given ability estimations based on a less precise adaptive test than examinees in the middle of the ability distribution (Thissen & Mislevy, 2000).

Variable Length

A variable length stopping rule terminates a test once a pre-specified level of measurement precision has been reached. Measurement precision is usually assessed

based on the standard error associated with a given ability. After each item is completed, the standard error associated with a given ability is calculated to determine if the examinee should be administered another item or if the exam is finished. Most often this leads to CATs with different item lengths, which as mentioned previously, can be difficult to explain to examinees. The advantage of implementing variable length stopping rules is that all examinees' ability estimates have the same measure of precision (Thissen & Mislevy, 2000).

Content Balancing

Often tests are required to meet content specifications in order to ensure examinees are assessed over the same material across tests. During CAT administrations, the item selection procedure may administer items that provide the most information at the examinee's interim ability estimate, but not fulfill content specifications. For example, a math test designed to assess examinees over addition, subtraction, multiplication, and division may result in a CAT containing only subtraction and multiplication items. For an examinee with little knowledge in addition and division, this would overestimate their final ability estimate. On the other hand, an examinee proficient in addition and division may receive an underestimated final ability estimate. Content balancing is a nonpsychometric issue that does not involve the IRT/TRT measurement model selected for the test. In order to provide a fair assessment across content areas, it might be necessary to include a content balancing procedure within the CAT system. The following sections discuss

the Kingsbury and Zara (1989) content balancing procedure and the weighted deviations model (Stocking & Swanson, 1993).

Kingsbury and Zara Procedure

The Kingsbury and Zara (1989) procedure allows test administrators to determine a priori what proportions of the test will assess the various content areas. Once the target proportions are determined, the procedure compares the target proportions for content balancing to the actual proportions during the administration of items. The content area with the largest discrepancy between the target proportion and actual proportion during administration of the items will be the next content area from which an item is selected. The items in the selected content area are evaluated psychometrically through an item selection procedure to determine which item is selected for administration. The content balancing procedure is repeated after each item is administered to determine the next content area. For the initial item administered, the content of the item may be randomly selected from all the contents, may be selected from the content with the largest proportion of items, or may be pre-specified to start with a particular content.

Morrison, Subhiyah, and Nungester (1995) employed the Kingsbury and Zara (1989) procedure to examine item exposure rates for content-balanced and unconstrained CATs. Their findings indicate that content balancing, while enabling tests to meet content specifications, did not significantly inflate the number of items administered for a test in a variable length CAT.

Weighted Deviations Model

Stocking and Swanson (1993) proposed the weighted deviations model as a method for balancing content specifications in CATs. The Stocking and Swanson (1993) procedure differs from the Kingsbury and Zara (1989) procedure in that it combines the nonpsychometric issue of content balancing with the psychometric evaluation of the items by weighting each of the desired properties. Each constraint is assigned an upper and lower bound (which may be equal). For each item in the pool, deviations from the upper and lower bounds are calculated for each constraint and multiplied by the weight of importance assigned to the constraint. The weighted deviations are summed across the constraints (nonpsychometric and psychometric) and the item with the smallest weighted sum of deviations is selected for administration (Stocking & Swanson, 1993; Stocking & Lewis, 1998). This procedure is particularly useful when the number of content constraints is large.

Exposure Control

Parshall, Davey, and Nering (1998) noted "the goal of good exposure control is use as much of the item pool as possible, without overly using any part of it." Exposure control refers to constraining the administration of more popular items that would otherwise become compromised due to repeated administrations. If examinees have prior knowledge of an item due to frequent administrations, the psychometric properties of the item will not accurately estimate the examinees' abilities and the item will no longer be valid. Administering a test through CAT does not in itself cause the overexposure of items. The issue arises with the increased frequency with

which CATs may be administered. The necessity of exposure control is most relevant to high stakes continuous testing. High stakes tests generally are those that affect examinees' admissions or candidacy opportunities and are offered to a large examinee population. An example of a high stakes tests, as distinguished from a low stakes tests, is the Graduate Record Examination that impacts students' chances of admissions to graduate schools. Continuous testing refers to administering tests on an ongoing basis, rather than periodic tests whereby the test is offered two to three times a year (Stocking & Swanson, 1998).

Way (1998) classified exposure control procedures into two categories: randomization and conditional selection procedures. Rather than selecting a single item at the maximum information level, randomization procedures select several items near the optimal level of maximum information from which one item is then randomly selected for administration. Although relatively easy to implement, randomization procedures do not allow specification of a maximum exposure rate. Conversely, conditional selection procedures have preset exposure control parameters that meet a pre-selected maximum exposure rate. Obtaining the exposure control parameter can be an arduous process that may need to be repeated if the ability distribution of the examinee population changes. In addition to the randomization and conditional selection procedures, Chang and Ying (1996) developed stratification procedures in which items with low discrimination are administered first followed by items with high discrimination, as more accurate estimations of examinees' ability levels are determined.

Several methods are usually employed to evaluate the various exposure control procedures: precision of measurement, exposure rate, pool utilization, and test overlap. Precision of measurement refers to how well the CAT system with exposure controls estimates examinees' abilities in comparison to the examinees' known abilities. This is evaluated through the Pearson product-moment correlation, bias, standardized difference between means (SDM), root mean squared error (RMSE), standardized root mean squared difference (SRMSD), and average absolute difference (AAD).

The exposure rate is the proportion of the number of times an item is administered to the total number of CATs administered. The pool utilization represents the percentage of items not administered during any of the CAT administrations. Ideally, all the items in the pool will be used; therefore this number should be very low. Test overlap, or item/testlet overlap, refers to the number of items that two examinees have in common. High levels of test overlap indicate that many examinees are seeing the same items. This is also calculated based on the similarity between the examinees' abilities. Previous research has defined item overlap for examinees with 'similar' abilities as two examinees having ability values within two logits and 'different' abilities are examinees with discrepancy in ability values larger than two logits (Pastor, Chiang, Dodd, & Yockey, 1999; Davis, Pastor, Dodd, Chiang, & Fitzpatrick, 2000; Pastor, Dodd, & Chang, 2001).

Chang and Zhang (2002) expanded the definition of test overlap by comparing the number of items that several examinees may have in common, rather

than comparing the number of items that a pair of examinees have in common. This stems from the likelihood that an examinee would obtain information about a test from several examinees rather than just one examinee. Chang and Zhang (2002) proved that for a fixed length CAT with random item selection, the number of items that overlap for any random sample of α examinees follows the hypergeometric distribution family for $\alpha \geq 1$. This provides, for α examinees, the lower bound for item overlaps, which may be used as a benchmark in comparing the overlap rates from CAT systems with exposure control procedures (Chang & Zhang, 2002).

The following sections discuss randomization, conditional, and stratification exposure control procedures separately for the dichotomous IRT models and the polytomous IRT models. Current research on the application of exposure control procedures in CAT systems will also be reported for the dichotomous and polytomous IRT models. To date, there is no research evaluating exposure control procedures in CAT systems based on the testlet response theory model.

Exposure Control Procedures for Dichotomous Models

The most common exposure control procedures used in CAT systems based on dichotomous IRT models are discussed in this section. The randomization procedures include the 5-4-3-2-1 procedure, randomesque procedure, within .10 logits procedure, and the progressive procedure. The conditional selection procedures consist of the Sympson-Hetter procedure, the conditional Sympson-Hetter procedure, Davey-Parshall procedure, Stocking and Lewis multinomial procedure, restricted

maximum information procedure, and the progressive restricted procedure. The stratification procedures include the a-Stratified procedure and the enhanced stratified procedure.

Randomization Procedures

Rather than selecting the most appropriate item for administration, randomization exposure control procedures select a group of the most appropriate items from which one is randomly selected for administration. Although these procedures do not allow for specification of a maximum exposure rate, randomization procedures are easy to administer in CAT systems. In the following sections, four randomization procedures are described, along with a discussion on current research for each procedure.

5-4-3-2-1 Procedure. An initial randomization procedure, the 5-4-3-2-1 procedure, was proposed by McBride and Martin (1983). This procedure selects the first item for administration randomly from the five most informative items. The second item is randomly selected from the four most informative items. This process is continued such that the third and fourth items are randomly selected from the three and two most informative items, respectively, until the fifth item. From this point, the remaining items administered are selected based on maximum information. The initial selection of five items is arbitrary. A smaller or larger number of items can be selected initially (Stocking, 1992).

The 5-4-3-2-1 procedure focuses on the initial items selected for administration since these tend to be the items most likely to be overexposed because

examinees usually begin with a common ability for the item selection procedure (Stocking & Lewis, 2000). The advantage to this procedure is the simplicity of its implementation. The disadvantages to this procedure include an overexposure of the most popular items due to returning items selected in the groups, but not administered, to the item pool and the inability to constrain the exposure of items to a specific maximum exposure rate (Parshall, Davey, & Nering, 1998; Stocking & Lewis, 2000).

Randomesque Procedure. Kingsbury and Zara's (1989) randomesque procedure is similar to the 5-4-3-2-1 procedure by randomly selecting an item from a group of optimal items for administration. The randomesque procedure differs in that it repeatedly selects the same number of the most informative items (e.g. 2, 3, 4,…, 10) from which one is randomly selected for administration throughout testing and does not switch to maximum information selection at anytime. Kingsbury and Zara (1989) proposed that continuing the randomization technique throughout testing will decrease the overlap in items seen by examinees of similar abilities.

Revuelta and Ponsoda (1998) applied the randomesque procedure, choosing a group of five items from which one was randomly selected for administration. They used a real item pool and evaluated it based on precision and exposure control. Precision was assessed through bias (overall difference between known ability and the estimated ability). The randomesque procedure had high levels of precision for a fixed-length condition (bias = 0.08) and a variable length condition (bias = 0.14). Exposure control was assessed through coefficient of variation, percentage of items

never administered, and minimum and maximum values for the exposure rate. The coefficient of variation was high indicating that some items had high exposure rates and some items were rarely administered. This was also reflected in the percentage of items never administered (19.5%). The minimum exposure rate was zero and the maximum exposure rate was 0.70, indicating that some items were seen by almost 70% of the examinees. Although, the randomesque procedure provided good measurement precision, it performed poorly in protecting the item pool.

Within .10 Logits Procedure. Lunz and Stahl (1998) developed the within .10 logits procedure to examine the number and pattern of items that overlapped across examinees with similar abilities. The within .10 logits procedure switches the focus of item selection from information to item difficulty because the Rasch (1962) model was used and therefore the information and item difficulty yield the same item selection. This procedure randomly selects an item from all items within .10 logits of the desired difficulty level. Therefore all items within the specified range are available for selection rather than an arbitrary number of items. If there are no items available within this range, the item with the closest difficulty level is administered. This procedure is continued throughout testing.

Lunz and Stahl (1998) observed a decrease in common items when examinee abilities were different and a decrease in the mean percent of common items across examinees the larger the item pool. Bergstrom and Lunz (1999) applied this procedure to a certification exam with an item bank of 900 items. The maximum exposure rate did not exceed 30% of the item pool. Typically, items that reported

maximum exposure rates close to 30% were around the cut point for the certification exam.

Progressive Procedure. Revuelta and Ponsoda (1998) developed the progressive procedure, which includes both a randomization component and maximum information selection. At first the randomization component influences item selection more than the maximum information selection. As the test progresses, the roles are reversed resulting in maximum information selection impacting item selection more than the randomization component. For the progressive procedure, the number of items previously administered to a person is denoted h and the total number of items to be administered is denoted m. The influence of the randomization component and maximum information is controlled by the serial position, s, of the item. The item's serial position is defined as $s = h/m$. When selecting an item for administration, the remaining items' information (I_i) is calculated based on the person's current estimated ability level. The highest information value, denoted H, is used to create a uniform distribution ranging from 0 to H from which a random number (R_i) is assigned to each item. Then a weight is calculated for each item, denoted $w_i = (1 - s)R_i + sI_i$, and the item with the greatest weight, w_i, is selected for administration.

In the same study as the randomesque procedure mentioned earlier, Revuelta and Ponsoda (1998) examined the progressive procedure with real item data and additionally, under simulated conditions that varied in terms of test length, item pool size, and discrimination parameter distributions. The results indicated high levels of

precision as measured by bias for both the real data and all the simulated conditions. For exposure control, the progressive procedure reported zero or close to zero for the percentage of items never administered. This indicates that unlike the randomesque procedure, most, if not all, of the items were administered during the CATs. Yet, the maximum exposure rates for the items were still high, ranging from .45 to .65.

Conditional Procedures

The conditional procedures allow a maximum exposure rate of the items in the item pool to be selected. Most of these procedures require extensive simulations to determine the exposure control parameters for the items in the item pool. The simulations can be computationally complex and time consuming. In the following sections, six conditional procedures are described, along with a discussion on current research for each procedure.

Sympson-Hetter Procedure. The most commonly used conditional selection procedure is the Sympson-Hetter procedure (Sympson & Hetter, 1985). The Sympson-Hetter procedure assigns an exposure control parameter, K_i, value ranging from zero to one for each item based on the frequency of item administrations during an iterative CAT simulation program. Items with high administration frequencies will have smaller exposure control parameters to limit their administration in a live CAT test. This ensures a maximum item exposure rate, r.

The procedure for setting the exposure control parameter, K_i, for each item is described as follows. For the first CAT simulation, the exposure control parameter, K_i, is initially set to one for each item. As an item is selected for administration, the

item's exposure control parameter, K_i, is compared to a random number, x, that is selected from a uniform distribution. If $x \leq K_i$, then the item is administered. Whether or not the item is administered, the item will not be selected for the simulee again. Once a CAT has been completed for all simulees, two probability values are computed: the probability of selecting an item, *P(S)*, and the probability of administering an item, *P(A)*.

$$P(S) = \frac{NS}{NE} \quad , \tag{30}$$

$$P(A) = \frac{NA}{NE} \quad , \tag{31}$$

where *NS* represents the number of times an item is selected, *NA* represents the number of times an item is administered, and *NE* represents the total number of examinees. Next, new K_i values are computed for each item. If $P(S) > r$, then the new $K_i = r/P(S)$. If $P(S) \leq r$, then the new $K_i = 1.0$. The new K_i values are examined to ensure there are enough items to administer a complete test. Therefore n items must have K_i values equal to one. If not, the items with K_i values closest to one are assigned K_i values equal to one until n items have K_i values equal to one (Sympson & Hetter, 1985).

The CAT simulations are repeated until the maximum item exposure rate, *r,* converges. Convergence is when the maximum value of *P(A)* for any item nears a limit slightly above *r* and fluctuates around this value. Therefore, to obtain a maximum exposure rate of 0.20, a value of r should be selected that is slightly lower,

such as 0.19. The final exposure control parameters, K_i, are used for the real CAT administrations (Sympson & Hetter, 1985).

Hetter and Sympson (1997) applied the Sympson-Hetter procedure to simulation data based on the Computerized Adaptive Testing – Armed Services Vocational Aptitude Battery (CAT-ASVAB). The results indicate that the exposure of items stayed below the pre-specified maximum exposure rate of 0.33. In addition, measurement precision was not significantly affected by the exposure control parameters.

The dependency of the Sympson-Hetter procedure on an expected examinee ability distribution limits the use of the exposure control parameters. Using exposure control parameters with an inappropriate examinee ability distribution will lead to items with overly relaxed or overly restricted exposure parameters. Therefore, changes in the actual examine ability distribution requires recalibrating exposure control parameters for the Sympson-Hetter based on this new expected distribution (Parshall, Davey, & Nering, 1998). In addition, the Sympson-Hetter exposure control parameters are global in nature, in that an item has one exposure control parameter no matter the examinee's ability level. This may inaccurately control items for ability levels at the tails of the ability distribution (Parshall, Davey, & Nering, 1998).

An extension of the Sympson-Hetter procedure was proposed by Stocking (1992) to reflect a realistic adaptive testing paradigm. In live CAT administrations, practical issues need to be considered, such as items that need to be administered together, as a block, because they have a common stimulus, require the same set of

directions, or refer to the same knowledge content area. Stocking (1992) extended the Sympson-Hetter to account for items requiring block administrations and to control the exposure of stimuli (e.g. reading passages) during a CAT administration. In comparison to a randomization procedure, 8-7-6-…-1, the extension of the Sympson-Hetter, had smaller item exposures. This modified Sympson-Hetter procedure is dependent on both the ability distribution used to create the exposure control parameters and the item pool due to creation of blocks and exposure rates for stimuli. A change in the examinee ability distribution or item pool requires recalibrating the exposure control parameters (Stocking, 1992).

Conditional Sympson-Hetter Procedure. Stocking and Lewis (1995) developed the conditional Sympson-Hetter procedure in which the exposure control parameters are estimated based on ability level. This removes the requirement of and issues pertaining to assuming an expected examinee ability distribution (Parshall, Davey, & Nering, 1998; Stocking, 1992; Stocking & Lewis, 1995). During the simulation stage of setting the exposure control parameters, an *n x m* matrix is created in which *n* rows reflect the number of items in the pool and *m* columns reflect a pre-specified number of discrete ability points along the ability distribution. The simulation procedure described in the previous section is then employed to obtain the *n x m* matrix of exposure control parameters conditioned on ability. Therefore, each item has *m* exposure control parameters, one at each of the discrete ability points.

Parshall, Davey, and Nering (1998) compared the Sympson-Hetter and the conditional Sympson-Hetter. The conditional Sympson-Hetter used more of the item

pool than the Sympson-Hetter, although some of the items were overexposed more for the conditional Sympson-Hetter than the Sympson-Hetter. Descriptive statistics on test length indicated the Sympson-Hetter yielded average test lengths of 30 items across the ability distribution, while the conditional Sympson-Hetter yielded longer tests at the tails of the ability distribution and shorter tests in the middle of the distribution.

Davey-Parshall Procedure. Davey and Parshall (1995) proposed the Davey-Parshall procedure as an alternative conditional procedure that would condition on items rather than on ability. This procedure focuses on minimizing test overlap across examinees. Similarly to the Sympson-Hetter procedure, the Davey-Parshall procedure requires setting exposure control parameters through simulations prior to live CAT administrations. The exposure control parameters are created in an *n x n* matrix, where *n* represents the number of items in the item pool. Conditional parameters in the off diagonal matrix control the frequency with which items appear together based on pairwise comparisons during simulations thereby controlling test overlap. The diagonal of the matrix has exposure control parameters that perform similarly to the Sympson-Hetter parameters.

Parshall, Davey, and Nering (1998) also compared the Davey-Parshall procedure to the Sympson-Hetter and conditional Sympson-Hetter procedures. Although, the Davey-Parshall and Sympson-Hetter procedures performed similarly in terms of pool usage, test length, and item overlap, the Davey-Parshall procedure consistently performed better. The Davey-Parshall procedure used more items in the

item pool than the Sympson-Hetter procedure, but fewer items than the conditional Sympson-Hetter procedure. The conditional Sympson-Hetter procedure repeatedly performed better than the other two procedures.

Parshall, Hogarty, and Kromrey (1999) examined a combination of all three procedures, the Sympson-Hetter, conditional Sympson-Hetter, and the Davey-Parshall, into a single procedure called the Tri-Conditional procedure. While examining the Tri-Conditional and Sympson-Hetter procedures, the Tri-Conditional showed better measurement precision and pool usage. The authors noted that the Tri-Conditional model is computationally more complex than the Sympson-Hetter, which may limit its use in practical applications (Parshall, Hogarty, & Kromrey, 1999).

Stocking and Lewis Multinomial Procedure. Stocking (1992) noted that the exposure control parameters had difficulty converging for the Sympson-Hetter procedure due to artificially setting some of the K_i values equal to one in order to have enough items to administer the test. Due to convergence problems when setting the exposure control parameters for the Sympson-Hetter procedure, Stocking and Lewis developed a multinomial procedure (Stocking & Lewis, 1995). Initially the exposure control parameters are set using the Sympson-Hetter procedure. Secondly, rather than selecting items based on optimal item selection, this procedure employs a multinomial model for item selection. Multinomial probabilities are calculated to determine the probability of selection based on all previous items not being selected (Stocking & Lewis, 1995).

Restricted Maximum Information Procedure. Revuelta and Ponsoda (1998) proposed the restricted maximum information procedure as a way of specifying a maximum exposure rate, without requiring the prior simulations needed to set the K_i values for the Sympson-Hetter procedure. For the restricted maximum information procedure, a pre-specified maximum exposure rate, *k,* limits the number of times an item, *i,* can be administered across all tests. For all tests already administered, *t,* the exposure rate of item, *i,* is equal to the number of times the item has been administered, a_i, divided by the current number of administered tests (a_i/t). All items with a current exposure rate below the maximum exposure rate, *k,* are included in the item pool from which an item is selected based on maximum information. If an item's current exposure rate (a_i/t) is larger than the maximum exposure rate, *k,* then the item will be removed from the item pool and not considered for selection. Once more tests have been administered the item's exposure rate (a_i/t) will reduce thereby allowing it to be reinstated in the item pool once it falls below the maximum exposure rate, *k*.

In the same study discussed previously, Revuelta and Ponsoda (1998) examined the restricted maximum information procedure with a maximum exposure rate restricted to 0.40 along with the progressive and randomesque procedures. In terms of precision, the restricted maximum information procedure required administration of three more items compared to the progressive and randomesque procedures in order to obtain the same level of precision. The restricted maximum information procedure successfully maintained a maximum item exposure rate of 0.40, but 15% of the item pool was never administered. Due to the progressive

procedure's administration of all of the items and the restricted maximum information procedure's ability to control the maximum item exposure rate, Revuelta and Ponsoda (1998) decided to combine these procedures.

Progressive Restricted Procedure. Revuelta and Ponsoda (1998) combined the restricted maximum information procedure and the progressive procedure to create the progressive restricted procedure. Before administration of a CAT, the available items are determined by the restricted procedure, such that no item will exceed the maximum exposure rate, *k*. Once the item pool is determined for a CAT, the progressive procedure is used to select an item for administration.

The progressive restricted procedure was examined with two maximum exposure rates $k = .40$ and $k = .15$. Overall the combination of the restricted maximum information procedure and the progressive procedure resulted in most if not all items in the item pool being administered and maintained a maximum exposure rate equal to *k* (Revuelta & Ponsoda, 1998).

Stratification Procedures

The discrimination parameter impacts the calculation of information, such that higher discrimination parameters yield higher information functions. This relationship results in items with high discrimination values being exposed more than items with low discrimination values (Chang & Ying 1999). The stratification procedures take advantage of the relationship between information and discrimination. In the following sections, two stratification procedures are described, along with a discussion on current research for each procedure.

a-Stratified Procedure. Chang and Ying (1999) developed the a-Stratified procedure to include the item discrimination parameters more directly in the exposure control procedure. Rather than focusing on maximum information for item selection, Chang and Ying (1999) propose that at the beginning of the CAT when little is known about the examinee's ability, lower discriminating items should be administered. The a-Stratified procedure stratifies the item pool based on the discrimination parameter, *a*. The item bank is partitioned into *K* levels based on the item discrimination, *a,* values. The test is also partitioned into *K* stages with the items denoted $n_1, n_2,\ldots, n_K$. Within the *k*th stage, the n_k item is selected whose difficulty level, *b*, is most similar to the expected ability value, $\hat{\theta}$. This procedure is repeated for $k = 1, 2,\ldots, K$. As the CAT is administered, the examinee's ability estimate will come closer to approximating the examinee's known ability. At this point items with higher discriminating values will be administered (Chang & Ying, 1999).

The *a*-Stratified procedure has four possible advantages. First, by restricting the use of highly discriminating items until the examinee's ability is well estimated, a more efficient method for estimation may result. Secondly, the stratification of the item bank may lead to more evenly distributed exposure rates. Thirdly, items with low discriminating values will be administered more often. Finally, the simpler method does not require extensive computational simulations prior to live testing. This allows for items to be added and removed from the item pool more easily (Hau & Chang, 1998).

Chang and Ying (1999) observed that the *a*-Stratified procedure had lower test overlap and used more of the item pool than the Sympson-Hetter with Fisher information or the Sympson-Hetter with Bayesian item selection. In addition, the *a*-Stratified procedure administered more items with low discrimination than either of the Sympson-Hetter with Fisher information or the Sympson-Hetter with Bayesian item selection procedures. Tang, Jiang, and Chang (1998) observed better item pool usage and more distributed exposure rates for the *a*-Stratified procedure than the Sympson-Hetter procedure for real item data when examinee ability matches the item pool difficulty. When examinee ability does not match item pool difficulty, the *a*-Stratified procedure does not select items as well. The authors suggested increasing the number of strata when examinee ability and item pool difficulty do not match. In addition, Tang, Jiang, and Chang (1998) discussed incorporating maximum information selection in the last strata as future research to overcome this disadvantage.

Stocking (1998) observed that stratification based only on the discrimination parameter may lead to lower strata with wider ranges of difficulty and higher strata with narrower ranges of difficulty due to the correlation between difficulty and discrimination parameters. In response, Chang, Qian, and Ying (2001) modified the *a*-Stratified procedure to have *b*-blocking. By incorporating *b*-blocking, the examinee ability level will be able to match a difficulty level for each stratum. This procedure consists of arranging the items by difficulty level and then dividing the item pool into *M* blocks. Then within each of the *M* blocks, partition it into *K* strata based on the

discrimination values. The items are then recombined such that there are $j = 1, 2,\ldots,$ K strata with each strata containing items with a range of difficulty values. Chang, Qian, and Ying (2001) reported that the *a*-Stratified procedure with *b*-blocking improved the control of item exposure rates and yielded lower mean squared errors in comparison to the *a*-Stratified procedure.

Enhanced Stratified Procedure. Leung, Chang, and Hau (1999) combined the a-Stratified procedure and the Sympson-Hetter procedure to create the enhanced stratified exposure control procedure. This procedure involves partitioning the item pool into K strata and identifying a maximum exposure rate, r. Simulated CATs are run using the a-Stratified method during which each item starts with an exposure control parameter equal to one. As the CATs are administered, the exposure control parameters are adjusted. Once exposure control parameters have been identified for each item, the live CATs consist of selecting the item with the difficulty level closest to the examinee's estimated ability within a stage. The item is administered if its exposure control parameter is greater than a random number from a uniform distribution; otherwise another item is selected for administration. This procedure continues throughout the stage and then moves to the next stage until the test is complete.

In comparison to the a-Stratified and the Sympson-Hetter procedures, the enhanced stratified procedure performed well at controlling the exposure of items. All the items were used during the CAT administrations, but they were not overly exposed. The Sympson-Hetter procedure reported as many as 166 of 400 items that

were not administered. Also, the test overlap rates were lowest for the enhanced stratified procedure (Leung et al., 1999).

Exposure Control Procedures for Polytomous Models

Research evaluating exposure control procedures for polytomous models is not as extensive as that with the dichotomous models. The following sections discuss the current research in determining the efficacy of exposure control procedures with polytomous models in the context of CATs. The polytomous models include the graded response model, the partial credit model, and the generalized partial credit model. Although some of the CATs vary in terms of content balancing and stopping rules, only maximum likelihood estimation has been investigated with exposure control procedures in polytomous CATs. The exposure control procedures discussed in this section are the randomesque procedure, the modified within .10 logits procedure, the Sympson-Hetter procedure, the conditional Sympson-Hetter procedure, the a-Stratified procedure, and the enhanced stratified procedure.

Randomesque Procedure

Davis (2002) examined the randomesque procedure with three polytomous models: graded response model (GR), partial credit model (PC) and the generalized partial credit model (GPC). Each model was examined for two item group levels, randomesque-3 and randomesque-6, such that an item was randomly selected from the three or six most informative items. The randomesque-3 procedure did not maintain item security, reporting maximum exposure rates of 0.73 for GR, 0.50 for PC, and 0.71 for GPC. In addition, the randomesque-3 did not administer 23% of the

items for GR, 20% for PC and 26% for GPC. In terms of item overlap, the randomesque-3 reported overall average overlap of 36% for GR, 24% for PC, and 33% for GPC. For each model, examinees with similar abilities had the highest percent of items overlapping.

The randomesque-6 procedure performed better than the randomesque-3 procedure across the three models. The maximum exposure rate for the three models was 0.49 for GR, 0.40 for PC, and 0.48 for GPC. The randomesque-6 also used more of the item pool resulting in the percentage of items not administered equaling 11% for GR, 8% for PC, and 13% for GPC. Item overlap rates for the randomesque-6 were lower than the randomesque-3 procedure, but on average 20-25% of the items were similar across examinees.

Modified Within .10 Logits Procedure

Since polytomous items do not have a single difficulty level, the selection procedure for the within .10 logits procedure (Lunz & Stahl, 1998) was modified to select a number of informative items based on ability level. Davis and Dodd (2001) developed the modified within .10 logits procedure to select a range of informative items around the examinee's current ability level. Items with the most information are selected at the ability level minus 0.10 logits, at the ability level and at the ability level plus 0.10 logits, and then an item is randomly selected from those items for administration. Davis (2002) looked specifically at selecting a total of three and six items for the graded response model, partial credit model, and generalized partial credit model. For the three-item group, the modified within .10 logits-3 selected the

most informative item at each of the ability levels. For the six-item group, the modified within .10 logits-6 selected the two most informative items at each of the ability levels. A single item was then randomly chosen from the three/six selected items for administration.

The modified within .10 logits-3 procedure did not maintain item security, reporting maximum exposure rates of 0.74 for GR, 0.54 for PC, and 0.71 for GPC (Davis, 2002). In addition, the modified within .10 logits-3 did not administer 22% of the items for GR, 20% for PC and 26% for GPC. In terms of item overlap, the modified within .10 logits-3 reported overall average overlap of 36% for GR, 25% for PC, and 33% for GPC. For each model, examinees with similar abilities had the highest percent of items overlapping. The modified within .10 logits-3 performed similarly to the randomesque-3 procedure.

The modified within .10 logits-6 performed better than the modified within .10 logits-3 procedure (Davis, 2002). The maximum exposure rate for the three models was 0.50 for GR, 0.40 for PC, and 0.48 for GPC. The modified within .10 logits-6 also used more of the item pool resulting in the percentage of items not administered equaling 11% for GR, 8% for PC, and 14% for GPC. Item overlap rates for the modified within .10 logits-6 were lower than the modified within .10 logits-3, but on average 20-26% of the items were similar across examinees. The modified within .10 logits-6 performed similarly to the randomesque-6 procedure.

Sympson-Hetter Procedure

Pastor, Chiang, Dodd, and Yockey (1999) investigated the extent to which the Sympson-Hetter procedure, with maximum exposure rate equal to 0.30, controlled item exposure for CAT systems based on the partial credit model that varied in terms of item pool size (60, 120) and stopping rule (variable, fixed). The Sympson-Hetter procedure showed negligible differences in measurement precision compared to a non-exposure control condition. The Sympson-Hetter also administered more items and reduced the amount of item overlap across examinees (Pastor et al., 1999).

Davis, Pastor, Dodd, Chiang, and Fitzpatrick (2000) enhanced the complexity of the partial credit model CAT system in the Pastor et al. (1999) study to include rotated content balancing and examined the performance of the Sympson-Hetter procedure. The Sympson-Hetter procedure administered more of the item pool for both content and non-content CATs compared to a no exposure control CAT. It also maintained the item exposure rates around or below the maximum allowed exposure rate.

In the Davis (2002) study mentioned previously, the author also investigated the Sympson-Hetter procedure with GR, PC, and GPC models. The target maximum exposure rate was equal to 0.39 and the maximum exposure rates for the models equaled 0.42 for GR, 0.43 for PC, and 0.42 for GPC. The Sympson-Hetter was able to control the maximum exposure rates, unlike the randomesque and modified within .10 logits procedures. Yet, the Sympson-Hetter procedure had a much higher percentage of items never administered compared to the randomization models. For

item overlap, almost a third of the items were similar across examinees. If the examinees had similar abilities, then even more items were alike.

Pastor, Dodd, and Chang (2002) compared several exposure control procedures for the generalized partial credit model. For the Sympson-Hetter procedure with a target exposure rate set equal to 0.30, the maximum exposure rate was 0.34 and 28% of the item pool was never administered. Compared to the other procedures, which will be discussed in the following sections, the Sympson-Hetter procedure did not perform well in controlling item exposure (Pastor et al., 2002).

Conditional Sympson-Hetter Procedure

Davis (2002) also investigated the conditional Sympson-Hetter procedure. The conditional Sympson-Hetter procedure showed marked improvement compared to the Sympson-Hetter procedure. The maximum exposure rate was 0.40 for GR and PC, and 0.41 for GPC. The percent of pool not administered for the three models was 15% for GR and PC, and 14% for GPC. The percent of item overlap was 26% for GR, 22% for PC, and 24% for GPC. In the Pastor et al. (2002) study, the conditional Sympson-Hetter procedure constrained the maximum exposure rate to 0.32 and administered most of the item pool. The percent of pool not administered equaled 3%.

a-Stratified Procedure

Davis (2002) investigated the utility of the a-Stratified procedure for controlling item exposure in CATs for the graded response model and the generalized partial credit model. The partial credit model was not included because it assumes the discrimination parameters are equal and thus are not modeled in the probability

function. The item pool was partitioned into five strata with 32 items in the first two strata and 31 items in the remaining three strata. The author noted a negative relationship between discrimination parameter and the number of categories for the polytomous items. Therefore the pool was stratified according to content area and number of categories, and then the pool was stratified by discrimination.

The Davis (2002) study results indicated poor exposure control for the a-Stratified procedure. The maximum exposure rate for GR was 0.90 and for GPC was 0.74. Over half of the items were never administered for both models. In addition, 40-46% of the items overlapped across examinees. Pastor et al. (2002) reported better pool utilization for the a-Stratified procedure with GPC ranging from 13-28% of the items never administered, but the maximum exposure rate was 1.00. Davis (2002) noted these results might be due to the prior stratification by content area.

Enhanced Stratified Procedure

Davis (2002) also investigated the utility of the enhanced stratified procedure for controlling item exposure in CATs for the graded response model and the generalized partial credit model. The enhanced stratified procedure used more of the item pool compared to the a-Stratified procedure. The enhanced stratified procedure reported much lower maximum exposure rates of 0.42 for both measurement models. This reflects the integration of the Sympson-Hetter procedure in the model. The percent of item overlap across examinees mirrored that of the Sympson-Hetter procedure. Pastor et al. (2002) also observed an increase in exposure control for the enhanced stratified procedure compared to the a-Stratified procedure. The maximum

exposure rate was 0.34 - 0.35 and the percent of pool not administered reduced to 2% and 13% depending on pool size.

Statement of Problem

Exposure control procedures protect CAT item pools from being compromised by restricting the exposure of some items and administering a larger percentage of the items in the item pools. Yet, implementation of an exposure control procedure in a CAT system reduces the precision of measurement of the test since the most optimal items are not necessarily administered. Test administrators need to weigh the benefits and costs of including exposure control procedures in CAT systems.

Previous research conducted by Revuelta and Ponsoda (1998) for dichotomously scored CATs indicates that the precision of measurement is not significantly impacted when using the randomesque procedure, progressive procedure, and the Sympson-Hetter procedure with a maximum exposure rate equal to 0.40. The restricted maximum information procedure did appear to decrease the precision of measurement when included in a CAT system. In terms of exposure rate, the Sympson-Hetter procedure and the restricted maximum information procedure provided the lowest maximum exposure rate. Yet, the progressive procedure used more of the item pool compared to the other procedures. In combination, the progressive restricted procedure obtained high levels of measurement precision and controlled the exposure of items (Revuelta & Ponsoda, 1998).

Chang (1998) also compared several exposure control procedures in a dichotomously scored CAT system, but in addition included the Kingsbury and Zara procedure (1989) for content balancing to create a more realistic CAT system. The 5-4-3-2-1 randomization procedure poorly controlled the exposure of items, but produced the least bias results. Although the Sympson-Hetter conditions controlled the item exposure rate to the designate level, a Sympson-Hetter procedure with a maximum exposure rate of 0.20 did not significantly impact the precision of measurement, while a Sympson-Hetter procedure with a maximum exposure rate of 0.10 did significantly impact the precision of measurement. In comparison to the Sympson-Hetter conditions, Chang (1998) noted that the Stocking and Lewis unconditional multinomial procedure yielded similar results, but was less efficient in determining the exposure control parameters than the Sympson-Hetter procedure. Chang (1998) reported the best item exposure control for the Davey-Parshall procedure and Stocking and Lewis conditional multinomial procedure, yet each of these led to poor precision of measurement.

Davis (2002) examined exposure control procedures for polytomously scored CAT systems with the Kingsbury and Zara content balancing procedure. In comparison to the randomesque procedure, modified within .10 logits procedure, conditional and unconditional Sympson-Hetter procedures, a-Stratified procedure and enhanced stratified procedure, the modified within .10 logits procedure and randomesque procedure with group size of six items provided low exposure rates and

low item overlap rates. In addition, these procedures administered a large portion of the item pool compared to a non-exposure control condition (Davis, 2002).

Exposure control procedures appear to perform differently for dichotomously scored CAT systems versus polytomously scored CAT systems with content balancing. In the dichotomous case, conditional selection procedures appear to be the optimal choice (Chang, 1998), while randomization procedures perform best for polytomous CATs (Davis, 2002). Exposure control procedures have yet to be examined with testlet-based CAT systems modeled using testlet response theory.

This dissertation examined various exposure control procedures in CAT systems based on the three-parameter logistic testlet response theory model and the partial credit model. Recommended exposure control procedures for dichotomously scored CAT systems and polytomously scored CAT systems were combined with the Kingsbury and Zara procedure (1989) for content balancing and expected a posteriori estimation of ability. The exposure control procedures are the randomesque procedure, the modified within .10 logits procedure, two levels of the progressive restricted procedure, and two levels of the Sympson-Hetter procedure. Each of these are compared to a baseline no exposure control procedure, maximum information. Specifically, this dissertation sought to answer the following questions:

1. To what extent do the exposure control procedures impact the precision of measurement for the CAT systems based on either the three-parameter logistic testlet response theory model or the partial credit item response theory model?

2. To what extent do the exposure control procedures control testlet exposure and testlet pool utilization for a CAT system based on either the three-parameter logistic testlet response theory model or the partial credit item response theory model?
3. Which is the optimal exposure control procedure for a CAT system based on the three-parameter logistic testlet response theory model?
4. Does the optimal exposure control procedure for a CAT system based on the three-parameter logistic testlet response theory model differ from the optimal exposure control procedure for a CAT system based on the partial credit item response theory model?

CHAPTER THREE: METHODOLOGY

Two measurement models appropriate for testlets were used to evaluate the relative merits of seven exposure control conditions in the context of computerized adaptive testing systems. The measurement models were the three-parameter logistic testlet response theory (TRT) model and the partial credit (PC) model. The exposure control procedures investigated were the randomesque procedure (Kingsbury & Zara, 1989), two levels of the progressive restricted procedure (Revuelta & Ponsoda, 1998), two levels of the Sympson-Hetter procedure (Sympson & Hetter, 1985), the modified within .10 logits procedure (Davis & Dodd, 2001), and a maximum information procedure. The maximum information procedure was used for a no exposure control condition in order to provide a baseline from which to compare the six other exposure control conditions for each measurement model.

Each CAT system consisted of maximum information testlet selection contingent on an exposure control procedure and content balancing for passage type and the number of items per passage; expected a posteriori (EAP) ability estimation for the interim and final ability estimations; and a fixed length stopping rule of seven passages totaling fifty multiple-choice items. Measurement precision and exposure rates were examined to evaluate the effectiveness of the exposure control procedures for each measurement model.

Item Pool

The data used to obtain item parameters for the item pool consisted of examinee responses from 22 forms of the Verbal Reasoning section of the Medical

College Admissions Test administered from April 1996 to April 2001. The average number of examinees per form was 7,234 examinees with a minimum of 2,510 and a maximum of 14,439 examinees. Each form contained eight reading passages and 55 multiple-choice items. The reading passages differed by content (humanities, social science, or natural science) and the number of multiple-choice items associated with the reading passages (6, 7, 8, or 10 items).

Previous research conducted through the Medical College Admissions Test (MCAT) Graduate Student Research Program on the reading passages of the MCAT indicated the presence of local item dependence on the Verbal Reasoning section and to a lesser extent on the Biological Sciences and Physical Sciences sections (Zenisky, Hambleton, & Sireci, 2002). Local item dependency was measured using the Q_3 statistic (Yen, 1984). The Verbal Reasoning section yielded positive Q_3 statistics for reading passage testlets ranging from 0.009 to 0.058 across two forms.

For the partial credit model, the testlet pool contained 149 reading passages (testlets) scored as polytomous items. When the PC model was applied to the testlet data, each item within a given testlet was scored correct or incorrect and summed to create the polytomous score for the testlet. The PC model testlet pool consisted of 40% humanities, 36% social science, and 24% natural science reading passages. In terms of the number of items per testlet, the PC model testlet pool consisted of 68% six-item, 20% seven-item, 7% eight-item, and 5% ten-item reading passages.

For the testlet response theory model, the testlet pool contained 176 reading passages (testlets) with a total of 1,210 dichotomous multiple-choice items. Each item

under the TRT model was scored dichotomously: correct (1) or incorrect (0). The TRT model testlet pool consisted of 37.5% humanities, 37.5% social science, and 25% natural science reading passages. In terms of the number of items per passage, the TRT model testlet pool consisted of 60% six-item, 18% seven-item, 10% eight-item, and 12% ten-item reading passages.

The discrepancy in the number of testlets for the PC and TRT models was due to the low category frequencies and convergence problems when estimating the testlet parameters for some of the passages under the PC model (Davis & Dodd, 2001). The TRT model used all the available testlets rather than mirror those testlets used by the PC model, thereby taking full advantage of the properties of the TRT model.

Parameter Estimation

The testlet parameters were estimated separately for the PC model and the TRT model. Each form was calibrated independently under each measurement model due to non-overlapping testlets across forms. The resulting testlet parameter estimates were combined to create the testlet pool. This process mirrored the randomly equivalent groups design used in the real test administrations.

The estimated testlet parameters for the PC model were obtained from the Davis and Dodd (2001) study. In that research, the same data for the MCAT forms described above were calibrated using the PARSCALE software program (Muraki & Bock, 1993). PARSCALE applies a two-step marginal maximum likelihood EM algorithm to estimate parameters. The two-step process is iterative until the testlet parameter estimates stabilize. The first step involves calculating the provisional

expected frequency and sample size. The second step estimates the marginal maximum likelihood. For each testlet, the number of step difficulty parameters, b_{ik}, is equal to the number of multiple-choice items associated with the reading passages. The research by Davis and Dodd (2001) found that the passages were similar in terms of discrimination power and therefore the more general GPC model was not needed. They used the most parsimonious model, the PC model, for testlet calibrations.

For the TRT model, the testlet parameters were estimated with the SCORIGHT software program (Wang, Bradlow, & Wainer, 2001). The three-parameter logistic TRT model with the testlet effect, $\gamma_{jd(i)}$, was used, resulting in four parameter estimates: difficulty (*b*), discrimination (*a*), guessing (*c*), and the testlet effect parameter ($\gamma_{jd(i)}$). The testlet effect was allowed to vary across testlets for all examinees. SCORIGHT employs a Markov Chain Monte Carlo technique with Gibbs sampling to draw inferences from the posterior distribution of the parameters to estimate the parameters of the model. The MCAT data was calibrated using 8000 iterations of which the first 7000 iterations were dropped. Every fifth-iteration of the remaining 1000 iterations was selected to create the posterior distribution of the parameters from which the inferences were drawn.

Data Generation

The PC model testlet response data were generated using the IRTGEN SAS macro (Whittaker, Fitzpatrick, Williams, & Dodd, in press). Response data were generated for ten samples, each with 1,000 simulees. Each simulee was assigned a

known theta value by randomly selecting theta from a normal distribution with mean zero and standard deviation equal to one. Based on the parameter estimates obtained from the calibration of the MCAT data and the simulee's known theta value, the probability of responding in each category for a given testlet was calculated. The category probabilities for a given testlet were then summed to create cumulative subtotal probabilities for each response category. A random number was selected from a uniform distribution that ranges from 0 to 1 and compared to the cumulative subtotal probabilities. If the random number was less than the subtotal probability for a given category, the simulee's response was that category score. This process was repeated for every testlet and every simulee in each of the ten samples. The resulting ten generated response data sets were used for each PC model CAT condition.

The TRT item response data were generated for the same ten samples as the PC model, each with 1,000 simulees. Each simulee was assigned a known theta value by randomly selecting theta from a normal distribution with mean zero and standard deviation equal to one. The probability of responding to an item was based on the simulee's theta value, the item parameter estimates obtained from SCORIGHT, and a generated person-specific testlet effect. The testlet effect parameter was determined by selecting a random variable from a normal distribution with mean zero and standard deviation equal to the square root of the variance of the testlet effect for a given testlet that was obtained from SCORIGHT. The selected random number was used as the testlet effect parameter in the probability model for all items in a testlet for that simulee. In order to introduce random error, the simulee's response was

compared to a randomly selected number from a uniform distribution that ranges from 0 to 1. The simulee received a correct response (1) if the random number was less than the simulee's response and an incorrect response (0) otherwise. This process was repeated for every item and every person in each of the ten samples. The same ten generated response data sets were used for each CAT condition based on the TRT model.

CAT Simulations

The CAT simulations were based on modifications made to a SAS program created by Chen, Hou, and Dodd (1998) and modified by Davis and Dodd (2001). Each CAT consisted of testlet selection based on maximum information contingent on content balancing and exposure control procedures. The ability and the person-specific testlet effects were estimated using expected a posteriori (EAP) estimation after each testlet was administered. The stopping rule for test administration was seven reading passages resulting in the administration of 50 multiple-choice items. Each CAT condition was repeated for the ten data sets. The following outlines the steps in the CAT algorithm, after which a more detailed explanation of each step is provided.

1. Content balancing for passage type and number of items per passage. (The initial content selection was randomly selected.)

2. Information calculated for applicable testlets. (The initial ability estimate, $\hat{\theta}$, was set to zero representing the mean of the population. Subsequent ability estimates were estimated through EAP estimation.)
3. Exposure control procedure leading to testlet selection and administration.
    a. Maximum information (MI)
    b. Randomesque procedure (RA)
    c. Modified within .10 logits (MW)
    d. Progressive restricted with .20 maximum exposure rate (PR20)
    e. Progressive restricted with .30 maximum exposure rate (PR30)
    f. Sympson-Hetter with .20 maximum exposure rate (SH20)
    g. Sympson-Hetter with .30 maximum exposure rate (SH30)
4. EAP estimation of interim ability and the testlet effect parameter.
5. Calculate item information, testlet information, test information, and standard error.
6. Repeat steps 1 through 5, unless number of testlets administered equals the stopping rule of seven testlets.
7. Estimate the final ability and testlet effect parameters with EAP estimation.

For administration of the first reading passage, the type of content and the number of items per passage were randomly selected for each examinee. The remaining reading passages were selected using the Kingsbury and Zara (1989)

procedure, which compares the target proportions for content balancing to the actual proportions during test administrations and selects the next testlet from the content with the largest discrepancy between the target and actual proportions. Therefore, each simulated test consisted of 40% humanities, 36% social science, and 23% natural science reading passages. Concurrently, the Kingsbury and Zara (1989) procedure controlled the number of items per passage such that each simulated test consisted of 42% six-item, 28% seven-item, 14% eight-item and 14% ten-item reading passages. For the remaining testlets that met the requirements for the Kingsbury and Zara (1989) procedures, the testlet information was calculated using the examinee's current ability estimate.

For the TRT model, the testlet information, which is the sum of the item informations, was used along with the exposure control procedure to determine which testlet was selected for administration to the examinee. Since, the testlet effect was unknown at this point in the CAT, the testlet effect was set to zero for testlet selection. This was equivalent to using the three-parameter logistic IRT model for item selection. The initial ability was equal zero (the mean of the population) and the interim ability estimates were employed for subsequent calculations of information.

Expected a posteriori (EAP) estimation was based on a normal prior distribution ranging from -4 to +4. Thirty evenly spaced quadrature points were used to calculate each of the weights to determine the posterior distribution. EAP was used to estimate the examinee's abilities and the testlet effect parameter. EAP was used for both the interim and final estimates.

Exposure Control Procedures

Seven exposure control conditions, based on five exposure control procedures, were examined for the partial credit model and the three-parameter logistic testlet response theory model. The resulting fourteen conditions were repeated ten times, once for each of the ten simulated data sets. The following discusses the seven exposure control conditions.

Maximum Information

Maximum information (MI) with no exposure control served as the baseline condition. The testlet with the most information was selected for administration throughout the CAT.

Randomesque Procedure

The randomesque procedure (RA) chose six of the most informative testlets from which one was randomly selected for administration. This procedure continued throughout the CAT.

Progressive Restricted Procedure

The progressive restricted procedure combined the restricted and progressive procedures. First, the restricted procedure defined a new testlet pool for the next examinee, and then the progressive procedure selected a testlet from this newly defined testlet pool. This procedure was evaluated at two levels: progressive restricted maximum information with a maximum exposure rate restricted to .20 (PR20) and progressive restricted maximum information with a maximum exposure rate restricted to .30 (PR30).

Modified Within .10 Logits Procedure

The modified within .10 logits (MW) procedure selected six testlets from which one testlet was randomly selected for administration. The testlets were selected based on testlet information for the examinee's estimated ability. The two most informative testlets were selected at the estimated ability. The two most informative testlets were selected at the estimated ability plus .10 logits. And the two most informative testlets were selected at the estimated ability minus .10 logits.

Sympson-Hetter Procedure

The exposure control parameters were set separately for the two Sympson-Hetter conditions. For each, a random sample of 10,000 simulees from a normal distribution with mean zero and standard deviation equal to one was generated. The resulting exposure control parameters were used to examine the Sympson-Hetter procedure at two levels: a maximum exposure rate equal to .20 (SH20) and a maximum exposure rate equal to .30 (SH30).

Each CAT test stopped after seven reading passages (testlets) were administered. A fixed length stopping rule, rather than a variable length stopping rule, was used to make comparisons across the exposure control procedures easier. With fixed length tests, exposure control procedures can be compared on measurement precision; and testlet overlap is easier to calculate.

Data Analyses

Assessment of the CAT systems was based on retrieval of simulees' known theta values and the effectiveness of the exposure control procedures. The degree to

which the CAT systems recovered the known theta values was evaluated through descriptive statistics, the Pearson product-moment correlation, bias, standardized difference between means (SDM), root mean squared error (RMSE), standardized root mean squared difference (SRMSD), and average absolute difference (AAD). The following equations illustrate the computation of bias, RMSE, SDM, SRMSD, and AAD:

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

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

$$SDM = \frac{\bar{\hat{\theta}} - \bar{\theta}}{\sqrt{\frac{s^2\hat{\theta} + s^2\theta}{2}}} \quad , \tag{3}$$

$$SRMSD = \sqrt{\frac{\frac{1}{n}\sum_{k=1}^{N}(\hat{\theta}_k - \theta_k)^2}{\frac{s^2\hat{\theta} + s^2\theta}{2}}} \text{ , and} \tag{4}$$

$$AAD = \frac{\sum_{i=1}^{N}\left|\hat{\theta}_k - \theta_k\right|}{n} \quad , \tag{5}$$

where $\hat{\theta}_k$ is the estimated ability obtained from the CAT and θ_k is the known ability used to generate the response data for examinee *k*. Each of the descriptive statistics was averaged across the results for the ten data sets.

Evaluation of the exposure control procedures was based on descriptive statistics of the testlet exposure rates including frequency, mean, standard deviation, and maximum exposure rate. Examinees' audit trails were examined to determine the frequency with which a testlet was administered in each CAT condition. The testlet exposure rate represented the number of times a testlet was administered to examinees divided by the total number of examinees. The percentage of testlets not administered during any of the CAT administrations represented pool utilization. In addition, each testlet was evaluated for test overlap across all examinees, examinees with similar abilities and examinees with different abilities. Examinees' audit trails were compared to determine the test overlap. Examinees' with similar abilities were defined in two ways: examinees having theta values within two logits and examinees having theta values within one logit. Examinees with different abilities were defined as examinees with discrepancy in theta values larger than two logits or examinees with discrepancy in theta values larger than one logit. Each of the measures used to evaluate the exposure control procedures was averaged across the results for the ten data sets.

CHAPTER FOUR: RESULTS

The results for the seven exposure control conditions are discussed separately for the two measurement models: the partial credit model and the three-parameter logistic testlet response theory model. For each model, the exposure control conditions are evaluated based on descriptive statistics, exposure rates, and test overlap. The results presented in this chapter are averages of the results for the ten data samples within each exposure control condition.

Partial Credit Model

Six of the seven CAT conditions were successfully completed for all ten data samples for the partial credit model. The CAT condition for the progressive restricted procedure with the maximum exposure rate restricted to 0.20 (PR20) was not successful for eight of the ten PC model data samples. For PR20, the process of setting exposure control parameters after each CAT administration, thereby creating a unique testlet pool for each examinee, resulted in two of the reading passage content areas with ten multiple-choice items having no available testlets to administer. The original PC testlet pool contained only one ten-item social science reading passage and one ten-item natural science reading passage. When the progressive restricted maximum exposure rate was 0.20, these testlets, after being administered, would not be available again until after four examinees completed CATs. For example, if the first examinee receives item 1, then item 1 has an exposure rate of 0.50 for the second examinee, 0.33 for the third examinee, 0.25 for the fourth examinee, and finally 0.20 for the fifth examinee. If the ten-item social science content area or ten-item natural

science content area is randomly selected as the content for the second, third, or fourth examinee, than the CAT program will automatically end due to a failure to meet the content balancing requirements for the CAT.

Since content balancing randomly selects the reading passage and number of items associated with the reading passages, this was a problem for some of the data samples, but not all. When the progressive restricted maximum exposure rate was restricted to 0.30, all the CAT conditions for the ten data samples were successfully administered. The following sections discuss the descriptive statistics, exposure rates, and test overlap for the exposure control conditions using the partial credit model.

Descriptive Statistics

The PC model testlet pool consisted of 149 testlets with 6, 7, 8, or 10 multiple-choice items. The mean, standard deviation, minimum and maximum for the step values are listed in Table 1. Due to the varying number of multiple-choice items associated with a testlet, the number of step values also varied. The PC model testlet pool contained 101 six-item passages, 30 seven-item passages, 10 eight-item passages, and 8 ten-item passages.

The grand mean, standard error of the mean, minimum mean, and maximum mean for the estimated thetas across ten replications for each of the exposure control conditions are listed in Table 2 for the partial credit model. The grand mean across ten replications for the known theta was -0.002 with a minimum mean of -0.053 and a maximum mean of 0.023.

TABLE 1: Descriptive Statistics of the Parameter Estimates Calibrated Using the Partial Credit Model

Partial Credit Model					
Testlet Parameter	N	Mean	Std Dev	Minimum	Maximum
Step Value 1	149	-2.217	0.574	-3.334	-0.863
Step Value 2	149	-1.440	0.473	-2.737	-0.354
Step Value 3	149	-0.938	0.499	-2.155	0.288
Step Value 4	149	-0.502	0.570	-1.809	1.165
Step Value 5	149	0.036	0.685	-1.604	1.908
Step Value 6	149	0.895	0.861	-1.129	2.739
Step Value 7	48	0.881	0.955	-0.888	3.085
Step Value 8	18	0.798	0.977	-0.603	2.760
Step Value 9	8	0.617	0.474	-0.009	1.103
Step Value 10	8	1.611	0.473	0.770	2.172

TABLE 2: Descriptive Statistics of the Estimated Thetas Yielded by the Partial Credit Model Across Ten Replications

Partial Credit Model				
	Estimated Thetas[a]			
Exposure Control Condition	Grand Mean	Standard Error of the Mean	Minimum Mean	Maximum Mean
Maximum Information	-0.011	0.021	-0.039	0.026
Randomesque	-0.010	0.026	-0.048	0.037
Modified Within .10 Logits	-0.007	0.025	-0.043	0.035
Progressive Restricted (.20)[b]	-0.017	0.017	-0.028	-0.005
Progressive Restricted (.30)	-0.010	0.018	-0.049	0.017
Sympson-Hetter (.20)	-0.014	0.021	-0.058	0.019
Sympson-Hetter (.30)	-0.015	0.018	-0.046	0.014

Note: Each replication contained 1,000 observations.
[a]Known Thetas: grand mean = -0.012, standard error of the mean = 0.023, minimum mean = -0.053, and maximum mean = 0.027
[b]Statistics are based on the two of ten replications that were successfully completed.

For the standard deviation of the estimated thetas, Table 3 lists the mean, minimum, and maximum across ten replications. The means of the standard deviation of the estimated thetas were slightly less than the means of the standard deviation for the known thetas. This reflects the tendency for EAP estimation to regress toward the mean of the prior distribution (Kim & Nicewander, 1993; Weiss, 1982). The estimated thetas and standard deviations for each condition under the partial credit model approximated a normal distribution with mean zero and standard deviation approximately one.

The standard error of the ability estimate reflects the precision of measurement of the exposure control conditions. The mean, minimum, and maximum for the standard errors across ten replications for each of the exposure control conditions are listed in Table 4 for the partial credit model. Maximum information (MI), the no exposure control condition, yielded the lowest mean of the standard errors (0.280). The Sympson-Hetter conditions yielded slightly higher means of the standard errors than the MI, with the Sympson-Hetter restricted to maximum exposure rate of 0.30 (SH30) reporting better measurement precision then the Sympson-Hetter restricted to maximum exposure rate of 0.20 (SH20). The progressive restricted procedure with maximum exposure rate of 0.30 (PR30) yielded a mean of the standard errors equal to 0.300 and the progressive restricted procedure with maximum exposure rate of 0.20 (PR20) yielded a mean of the standard errors equal to 0.302. The randomesque (RA) and modified within 0.10 logits (MW)

TABLE 3: Descriptive Statistics of the Standard Deviation of the Estimated Thetas Yielded by the Partial Credit Model Across Ten Replications

Partial Credit Model			
	Standard Deviation of Estimated Thetas[a]		
Exposure Control Condition	Mean	Minimum	Maximum
Maximum Information	0.956	0.926	0.989
Randomesque	0.946	0.911	0.972
Modified Within .10 Logits	0.947	0.906	0.970
Progressive Restricted (.20)[b]	0.939	0.936	0.943
Progressive Restricted (.30)	0.947	0.918	0.982
Sympson-Hetter (.20)	0.945	0.906	0.980
Sympson-Hetter (.30)	0.949	0.919	0.975

Note: Each replication contained 1,000 observations.
[a]Standard Deviation of Known Thetas: mean = 0.992, minimum = 0.954, and maximum = 1.022
[b]Statistics are based on the two of ten replications that were successfully completed.

TABLE 4: Descriptive Statistics of the Standard Errors Yielded by the Partial Credit Model Across Ten Replications

Partial Credit Model			
	Standard Errors		
Exposure Control Condition	Mean	Minimum	Maximum
Maximum Information	0.280	0.278	0.282
Randomesque	0.309	0.307	0.311
Modified Within .10 Logits	0.309	0.307	0.311
Progressive Restricted (.20)[a]	0.302	0.302	0.303
Progressive Restricted (.30)	0.300	0.298	0.302
Sympson-Hetter (.20)	0.290	0.288	0.292
Sympson-Hetter (.30)	0.284	0.282	0.286

Note: Each replication contained 1,000 observations.
[a]Statistics are based on the two of ten replications that were successfully completed.

procedures yielded the poorest measurement precision with means of the standard errors equal to 0.309. In practical terms, the differences between the exposure control conditions' means of the standard errors are negligible.

The correlation coefficients between the known and estimated thetas were calculated for the ten data samples within each exposure control condition. The mean, minimum, and maximum of the ten correlation coefficients for each exposure control condition are listed in Table 5. For the seven exposure control conditions, the correlations yielded similar results, ranging from 0.95 to 0.96. The exposure control conditions did not negatively impact the estimation of examinees' ability relative to the no exposure control condition.

The measurement statistics, bias, standardized difference between means (SDM), average absolute difference (AAD), root mean squared error (RMSE), and standardized root mean squared difference (SRMSD), are reported in Table 6 and Table 7 for the PC model. For each statistic, the mean, minimum, and maximum across the ten data samples are listed. The bias and SDM statistics are functionally zero when rounded to the second decimal place for each condition. The means for the AAD statistic range from 0.217 to 0.233. The means for the RMSE statistic range from 0.276 to 0.298. The means for the SRMSD statistic range from 0.539 to 0.565. The small differences in the measurement statistics across the exposure control conditions are not practically significant. In comparison to the no exposure control condition, MI, incorporation of the exposure control conditions in the PC CAT systems did not significantly decrease measurement precision.

TABLE 5: Descriptive Statistics of the Correlation Coefficients Between Known and Estimated Thetas for the Partial Credit Model Across Ten Replications

Partial Credit Model			
	Correlation Coefficient		
Exposure Control Condition	Mean	Minimum	Maximum
Maximum Information	0.961	0.956	0.964
Randomesque	0.954	0.950	0.958
Modified Within .10 Logits	0.954	0.950	0.958
Progressive Restricted (.20)[a]	0.954	0.952	0.955
Progressive Restricted (.30)	0.955	0.952	0.958
Sympson-Hetter (.20)	0.958	0.956	0.961
Sympson-Hetter (.30)	0.960	0.955	0.965

Note: Each replication contained 1,000 observations.
[a]Statistics are based on the two of ten replications that were successfully completed.

TABLE 6: Descriptive Statistics of the Bias, Standardized Difference Between Means (SDM) and Average Absolute Difference (AAD) for the Partial Credit Model Across Ten Replications

Partial Credit Model			
Exposure Control Condition	Bias Mean (Min, Max)	SDM Mean (Min, Max)	AAD Mean (Min, Max)
Maximum Information	0.000 (-0.014, 0.024)	0.001 (-0.024, 0.015)	0.217 (0.208, 0.224)
Randomesque	-0.002 (-0.029, 0.021)	0.002 (-0.022, 0.030)	0.233 (0.227, 0.239)
Modified Within .10 Logits	-0.005 (-0.032, 0.020)	0.005 (-0.020, 0.033)	0.234 (0.227, 0.239)
Progressive Restricted (.20)[a]	0.007 (0.003, 0.011)	-0.007 (-0.011, -0.003)	0.233 (0.231, 0.235)
Progressive Restricted (.30)	-0.002 (-0.011, 0.019)	0.002 (-0.019, 0.012)	0.231 (0.225, 0.235)
Sympson-Hetter (.20)	0.003 (-0.006, 0.013)	-0.003 (-0.013, 0.006)	0.223 (0.216, 0.231)
Sympson-Hetter (.30)	0.004 (-0.007, 0.015)	-0.004 (-0.015, 0.007)	0.220 (0.209, 0.228)

Note: Each replication contained 1,000 observations.
[a]Statistics are based on the two of ten replications that were successfully completed.

TABLE 7: Descriptive Statistics of the Root Mean Squared Error (RMSE) and Standardized Root Mean Squared Difference (SRMSD) for the Partial Credit Model Across Ten Replications

Partial Credit Model		
Exposure Control Condition	RMSE Mean (Min, Max)	SRMSD Mean (Min, Max)
Maximum Information	0.276 (0.263, 0.283)	0.539 (0.527, 0.561)
Randomesque	0.297 (0.287, 0.304)	0.562 (0.544, 0.518)
Modified Within .10 Logits	0.298 (0.284, 0.306)	0.563 (0.547, 0.586)
Progressive Restricted (.20)[a]	0.295 (0.293, 0.298)	0.565 (0.559, 0.571)
Progressive Restricted (.30)	0.294 (0.286, 0.299)	0.559 (0.545, 0.574)
Sympson-Hetter (.20)	0.284 (0.277, 0.296)	0.550 (0.532, 0.566)
Sympson-Hetter (.30)	0.279 (0.264, 0.286)	0.544 (0.520, 0.567)

Note: Each replication contained 1,000 observations.
[a]Statistics are based on the two of ten replications that were successfully completed.

Exposure Rates

The exposure rate is the proportion of the number of times an item is administered to the total number of CATs administered. When examining exposure rates, a key indicator for the effectiveness of the exposure control condition is the maximum exposure rate. Table 8 lists the grand mean, minimum mean and maximum mean of the exposure rates for each of the exposure control conditions. As expected the maximum information (MI) condition yielded the highest maximum exposure rate, .617, indicating that some testlets were exposed to as many as 62% of the examinees. The maximum testlet exposure rate for the randomesque (RA) condition was .195 and for the modified within .10 logits (MW) condition the maximum exposure rate was .198. The maximum testlet exposure rates for the RA and MW conditions reflect a high level of exposure control. The progressive restricted conditions and Sympson-Hetter conditions yielded maximum testlet exposure rates equivalent to their restricted exposure rates assigned to the conditions, 0.20 or 0.30.

The exposure control procedures have the same mean exposure rate because it is the ratio of test length to testlet pool size. However, the exposure control procedures do not provide the same level of exposure control. This is evident in the standard deviation of the exposure rates. Table 9 lists the mean, minimum, and maximum for the standard deviation of the exposure rates for each of the exposure control conditions. The MI condition yielded the highest average standard deviation (0.105) of the exposure rate, indicating the most variability in exposure rates across the testlets. The PR20 condition reported the lowest average standard deviation

TABLE 8: Descriptive Statistics of Testlet Exposure Rates for the Partial Credit Model Across Ten Replications

Partial Credit Model			
	Testlet Exposure Rates		
Exposure Control Condition	Grand Mean	Minimum Testlet Exposure Rate Mean	Maximum Testlet Exposure Rate Mean
Maximum Information	0.047	0.000	0.617
Randomesque	0.047	0.000	0.195
Modified Within .10 Logits	0.047	0.000	0.198
Progressive Restricted (.20)[a]	0.047	0.002	0.201
Progressive Restricted (.30)	0.047	0.001	0.300
Sympson-Hetter (.20)	0.047	0.000	0.214
Sympson-Hetter (.30)	0.047	0.000	0.315

Note: Each replication contained 1,000 observations.
[a]Statistics are based on the two of ten replications that were successfully completed.

TABLE 9: Descriptive Statistics of Standard Deviation of the Exposure Rates for the Partial Credit Model Across Ten Replications

Partial Credit Model			
	Standard Deviation of the Exposure Rates		
Exposure Control Condition	Mean	Minimum	Maximum
Maximum Information	0.105	0.104	0.107
Randomesque	0.050	0.050	0.051
Modified Within .10 Logits	0.051	0.051	0.052
Progressive Restricted (.20)[a]	0.049	0.049	0.049
Progressive Restricted (.30)	0.055	0.054	0.056
Sympson-Hetter (.20)	0.072	0.071	0.073
Sympson-Hetter (.30)	0.084	0.083	0.085

Note: Each replication contained 1,000 observations.
[a]Statistics are based on the two of ten replications that were successfully completed.

(0.049) of the exposure rate, indicating the least variability in exposure rates across the testlets. Ideally, the standard deviation of the exposure rates will be low indicating an even usage of testlets throughout the testlet pool.

Table 10 lists the frequency of testlet exposure rates averaged across ten data samples for each of the exposure control conditions. For all the exposure control conditions except MI, the majority of the testlets yielded exposure rates less than 0.20. Pool utilization represents the percentage of testlets not administered during any of the CAT administrations. As anticipated, the MI condition yielded the highest percentage of testlets not administered (62%). Both the SH20 and SH30 procedures yielded high percentage of testlets never administered, 52% and 57% respectively. This indicates that examinees' performance is estimated using less than half of the testlet pool. The RA and MW procedures performed much better with percentage of testlets never administered being 28% for both procedures. Surprisingly, the PR20 and PR30 procedures used the entire testlet pool, yielding 0% of the pool never being administered.

Test Overlap

Test overlap, or testlet overlap, refers to the number of testlets that two examinees have in common. Each testlet was evaluated for test overlap across all examinees, examinees with similar abilities, and examinees with different abilities. Examinees' audit trails were compared to determine the test overlap. Examinees' with similar abilities were defined in two ways: examinees having theta values within two logits and examinees having theta values within one logit. Examinees with different

TABLE 10: Frequency of Testlet Exposure Rates for the Partial Credit Model Averaged Across Ten Replications

Partial Credit Model							
	Exposure Control Condition						
Exposure Rate	MI	RA	MW	PR20[a]	PR30	SH20	SH30
.71-1.0	0	0	0	0	0	0	0
.61-.70	1	0	0	0	0	0	0
.51-.60	0	0	0	0	0	0	0
.41-.50	3	0	0	0	0	0	0
.36-.40	3	0	0	0	0	0	0
.31-.35	0	0	0	0	0	0	2
.26-.30	0	0	0	0	3	0	6
.21-.25	5	0	0	1	1	3	6
.16-.20	4	6	7	8	4	24	7
.11-.15	6	19	19	16	9	4	7
.06-.10	13	33	31	23	29	14	13
.01-.05	23	49	50	103	102	28	25
0.0 (Not Admin.)	92 (62%)	41 (28%)	42 (28%)	0 (0%)	0 (0%)	77 (52%)	84 (57%)

Note: Each replication contained 1,000 observations.
[a]Statistics are based on the two of ten replications that were successfully completed.

abilities were defined as examinees with discrepancy in theta values larger than two logits or examinees with discrepancy in theta values larger than one logit.

Table 11 shows the mean, minimum, and maximum for test overlap for all examinees, examinees with similar abilities (within two logits), and examinees with different abilities (greater than two logits). The largest overlap values are reported by the MI condition with examinees overall sharing as many as two testlets on average. Yet, when ability is taken into consideration for the MI condition, the examinees with similar abilities are sharing two testlets on average while examinees with different abilities have less than a testlet in common. Both the SH20 and SH30 procedures yield average overall overlap values above one indicating that on average examinees have a testlet in common. On closer examination for the SH20 and SH30 procedures, examinees with similar abilities (within two logits) are receiving one to two testlets in common, while different ability (greater than two logits) examinees are not likely to have testlets in common. The other four conditions, RA, MW, PR20, and PR30, yield average overall overlap rates below one indicating that on average examinees have less than a testlet in common. For examinees with similar abilities, these four conditions still yield overlap rates below one.

TABLE 11: Descriptive Statistics of Test Overlap for the Partial Credit Model Across Ten Replications Using Two Logits to Define Ability Groups

Partial Credit Model			
	Test Overlap		
Exposure Control Condition	Overall Overlap Mean (Min, Max)	Similar[b] Abilities Mean (Min, Max)	Different[c] Abilities Mean (Min, Max)
Maximum Information	1.962 (1.923, 2.001)	2.234 (2.194, 2.270)	0.473 (0.422, 0.518)
Randomesque	0.700 (0.691, 0.707)	0.755 (0.749, 0.762)	0.397 (0.377, 0.410)
Modified Within .10 Logits	0.714 (0.705, 0.724)	0.770 (0.763, 0.781)	0.407 (0.392, 0.421)
Progressive Restricted (.20)[a]	0.681 (0.6809, 0.681)	0.739 (0.734, 0.744)	0.354 (0.337, 0.372)
Progressive Restricted (.30)	0.764 (0.750, 0.780)	0.838 (0.825, 0.854)	0.354 (0.336, 0.378)
Sympson-Hetter (.20)	1.082 (1.062, 1.104)	1.224 (1.203, 1.239)	0.306 (0.285, 0.342)
Sympson-Hetter (.30)	1.364 (1.331, 1.390)	1.545 (1.527, 1.567)	0.368 (0.344, 0.399)

Note: Each replication contained 1,000 observations.
[a]Statistics are based on the two of ten replications that were successfully completed.
[b]Similar = abilities within two logits.
[c]Different = abilities greater than two logits.

In order to examine overlap rates more closely across examinees with similar abilities, the definition for similar abilities was restricted to a smaller range. Table 12 lists the mean, minimum, and maximum overlap rates for examinees with similar abilities (within one logit) and examinees with different abilities (greater than one logit). The largest overlap values are reported by the MI condition with similar examinees sharing as many as three testlets on average. The overlap rates for the SH20 and SH30 increased with the new definition of similarity. Examinees within one logit of each on the ability scale shared two testlets on average. The other four conditions, RA, MW, PR20, and PR30, maintained overlap rates below one indicating that on average examinees have less than a testlet in common.

TABLE 12: Descriptive Statistics of Test Overlap for the Partial Credit Model Across Ten Replications Using One Logit to Define Ability Groups

Partial Credit Model		
	Test Overlap	
Exposure Control Condition	Similar[b] Abilities Mean (Min, Max)	Different[c] Abilities Mean (Min, Max)
Maximum Information	2.818 (2.773, 2.865)	1.017 (0.978, 1.050)
Randomesque	0.860 (0.855, 0.867)	0.523 (0.509, 0.533)
Modified Within .10 Logits	0.874 (0.869, 0.880)	0.538 (0.525, 0.552)
Progressive Restricted (.20)[a]	0.868 (0.858, 0.878)	0.470 (0.463, 0.477)
Progressive Restricted (.30)	1.006 (0.994, 1.040)	0.496 (0.477, 0.518)
Sympson-Hetter (.20)	1.540 (1.523, 1.578)	0.577 (0.549, 0.621)
Sympson-Hetter (.30)	1.949 (1.926, 1.988)	0.716 (0.685, 0.743)

Note: Each replication contained 1,000 observations.
[a]Statistics are based on the two of ten replications that were successfully completed.
[b]Similar = abilities within one logit.
[c]Different = abilities greater than one logit.

Testlet Response Theory

The seven exposure control conditions were successfully completed for the testlet response theory model across all ten replications. Unlike the PC model, the TRT model contained more than one item for the ten-item social science content area and ten-item natural science content area. The following sections discuss the descriptive statistics, exposure rates, and test overlap for the exposure control conditions using the testlet response theory model.

Descriptive Statistics

The TRT model testlet pool consisted of 176 testlets with a total of 1,210 dichotomous multiple-choice items: 105 six-item passages, 33 seven-item passages, 17 eight-item passages, and 22 ten-item passages. The mean, standard deviation, minimum and maximum for the item parameters are listed in Table 13. The difficulty item parameter estimate yielded a mean of –0.760 and minimum and maximum estimates in the normal range for this parameter, -4.759 and 3.469 respectively. Although, the mean item parameter estimate of the discrimination, 1.018, is within an acceptable range, the minimum value of 0.178 is very low indicating that this item performed poorly in distinguishing between examinees with high levels of ability and examinees with low levels of ability. Ideally, the psuedo-guessing item parameter should be less than the probability of selecting the correct response by chance. The multiple-choice items in this study had four response options; therefore the probability of selecting the correct response by chance is 0.25. The mean for the

TABLE 13: Descriptive Statistics of the Item Parameter Estimates Calibrated Using the Testlet Response Theory Model

Testlet Response Theory Model					
Item Parameter	N	Mean	Std Dev	Minimum	Maximum
Difficulty	1210	-0.760	1.390	-4.759	3.469
Discrimination	1210	1.018	0.441	0.178	5.112
Psuedo-Guessing	1210	0.097	0.112	0.0	0.806

psuedo-guessing item parameter estimate was 0.097 indicating that on average examinees were unlikely to obtain a correct response by guessing The psuedo-guessing item parameter estimate had a maximum value of 0.806, which is a concern because examinees were very likely to obtain a correct response based on guessing for this particular item.

Table 14 shows the mean, standard deviation, minimum and maximum for the variance of the testlet effect by reading passage content area. The degree of dependency present in the testlets used in the current research was examined for the TRT model. The mean of the variance of the testlet effect was 0.49 with a minimum of 0.01 and a maximum of 1.67. Since the testlet effect was allowed to vary across testlets, testlets were examined for differences in the testlet effects. Specifically, the variances of the testlet effect parameters were compared across content of the reading passages and number of items per reading passage. A significant difference ($F(2,173) = 6.25$, $p = 0.0024$) in the estimates of the testlet effect parameter variance estimates was found between the reading passages (humanities, social science, and natural science). A post-hoc Tukey's test indicated a significant difference between the mean testlet effect variances between humanities (mean = 0.588) and natural science reading passages (mean = 0.358). The mean for social science was 0.489. Analysis of variance yielded no significant differences in the means for the number of items per reading passage (6, 7, 8, and 10); F = .22, df = 3, 172, and p-value = 0.8806.

TABLE 14: Descriptive Statistics of the Testlet Parameter Estimates Calibrated Using the Testlet Response Theory Model

Testlet Response Theory Model				
		Testlet Effect Variance		
Reading Passage	N	Mean	Minimum	Maximum
Humanities	66	0.588	0.022	1.582
Social Sciences	66	0.489	0.010	1.504
Natural Sciences	44	0.358	0.033	1.673

The grand mean, standard error of the mean, minimum mean, and maximum mean for the estimated thetas across ten replications for each of the exposure control conditions are listed in Table 15 for the TRT model. The grand mean across ten replications for the known theta was -0.002 with a minimum mean of -0.053 and a maximum mean of 0.023. For the standard deviation of the estimated thetas, Table 16 lists the mean, minimum, and maximum across ten replications. The means of the standard deviation of the estimated thetas for the exposure control conditions were lower than the means of the standard deviation for the known thetas. Since the exposure control conditions yielded similar means of the standard deviations as the maximum information (MI) condition, the exposure controls do not appear to be responsible for the lower means of the standard deviation of the estimated thetas. The TRT model may be more susceptible to the tendency for EAP estimation to regress toward the mean of the prior distribution (Kim & Nicewander, 1993; Weiss, 1982).

The standard error of the ability estimate reflects the precision of measurement of the exposure control conditions. The mean, minimum, and maximum of the standard errors across ten replications for each of the exposure control conditions are listed in Table 17 for the testlet response theory model. The MI, the no exposure control condition, yielded the lowest mean standard error of 0.311. The Sympson-Hetter conditions yielded slightly higher means of the standard errors than the MI, with the Sympson-Hetter condition restricted to a maximum exposure rate of .30 (SH30) reporting better measurement precision then the Sympson-Hetter

TABLE 15: Descriptive Statistics of the Estimated Thetas Yielded by the Testlet Response Theory Model Across Ten Replications

Testlet Response Theory Model				
	Estimated Thetas[a]			
Exposure Control Condition	Grand Mean	Standard Error of the Mean	Minimum Mean	Maximum Mean
Maximum Information	0.007	0.022	-0.030	0.039
Randomesque	-0.018	0.021	-0.059	0.017
Modified Within .10 Logits	-0.009	0.021	-0.049	0.021
Progressive Restricted (.20)	-0.007	0.022	-0.048	0.023
Progressive Restricted (.30)	-0.011	0.022	-0.047	0.025
Sympson-Hetter (.20)	-0.011	0.019	-0.041	0.012
Sympson-Hetter (.30)	-0.001	0.017	-0.035	0.023

Note: Each replication contained 1,000 observations.
[a]Known Thetas: grand mean = -0.012, standard error of the mean = 0.023, minimum mean = -0.053, and maximum mean = 0.027

TABLE 16: Descriptive Statistics of the Standard Deviation of the Estimated Thetas Yielded by the Testlet Response Theory Model Across Ten Replications

Testlet Response Theory Model			
	Standard Deviation of Estimated Thetas[a]		
Exposure Control Condition	Mean	Minimum	Maximum
Maximum Information	0.884	0.836	0.904
Randomesque	0.875	0.855	0.889
Modified Within .10 Logits	0.877	0.858	0.911
Progressive Restricted (.20)	0.883	0.852	0.911
Progressive Restricted (.30)	0.878	0.857	0.895
Sympson-Hetter (.20)	0.891	0.846	0.911
Sympson-Hetter (.30)	0.890	0.849	0.905

Note: Each replication contained 1,000 observations.
[a]Standard Deviation of Known Thetas: mean = 0.992, minimum = 0.954, and maximum = 1.022

TABLE 17: Descriptive Statistics of the Standard Errors Yielded by the Testlet Response Theory Model Across Ten Replications

Testlet Response Theory Model			
	Standard Errors		
Exposure Control Condition	Mean	Minimum	Maximum
Maximum Information	0.311	0.310	0.312
Randomesque	0.355	0.353	0.357
Modified Within .10 Logits	0.355	0.353	0.356
Progressive Restricted (.20)	0.350	0.349	0.351
Progressive Restricted (.30)	0.346	0.346	0.348
Sympson-Hetter (.20)	0.326	0.326	0.328
Sympson-Hetter (.30)	0.317	0.316	0.318

Note: Each replication contained 1,000 observations.

condition restricted to a maximum exposure rate of .20 (SH20). The progressive restricted condition restricted to a maximum exposure rate of .30 (PR30) yielded a mean of the standard errors equal to 0.346 and the progressive restricted condition restricted to a maximum exposure rate of .20 (PR20) yielded a mean of the standard errors equal to 0.350. The randomesque (RA) and modified within .10 logits (MW) procedures yielded the poorest measurement precision with a mean of the standard errors equal to 0.355 for both. In practical terms, the differences between the exposure control conditions' mean standard errors are negligible.

The correlation coefficients between the known and estimated thetas were calculated for the ten data samples within each exposure control condition. The mean, minimum, and maximum of the ten correlation coefficients for each exposure control condition are listed in Table 18. For the seven exposure control conditions, the mean of the correlations yielded similar results ranging from 0.91 to 0.92.

The measurement statistics, bias, standardized difference between means (SDM), average absolute difference (AAD), root mean squared error (RMSE), and standardized root mean squared difference (SRMSD), are reported in Table 19 and Table 20 for the PC model. For each statistic, the mean, minimum, and maximum across the ten data samples are listed. The bias and SDM statistics are functionally zero when rounded to the second decimal place for each condition except MI, which is 0.02, and SH30, which is 0.01. The means for the AAD statistic range from 0.311 to 0.334. The means for the RMSE statistic range from 0.392 to 0.420. The grand

TABLE 18: Descriptive Statistics of Correlation Coefficients Between Known and Estimated Thetas for the Testlet Response Theory Model Across Ten Replications

Testlet Response Theory Model			
	Correlation Coefficient		
Exposure Control Condition	Mean	Minimum	Maximum
Maximum Information	0.919	0.915	0.924
Randomesque	0.907	0.897	0.916
Modified Within .10 Logits	0.907	0.900	0.915
Progressive Restricted (.20)	0.908	0.900	0.917
Progressive Restricted (.30)	0.909	0.900	0.918
Sympson-Hetter (.20)	0.916	0.904	0.923
Sympson-Hetter (.30)	0.921	0.910	0.935

Note: Each replication contained 1,000 observations.

TABLE 19: Descriptive Statistics for the Bias, Standardized Difference Between Means (SDM) and Average Absolute Difference (AAD) for the Testlet Response Theory Model Across Ten Replications

Testlet Response Theory Model			
Exposure Control Condition	Bias Mean (Min, Max)	SDM Mean (Min, Max)	AAD Mean (Min, Max)
Maximum Information	-0.019 (-0.045, 0.000)	0.020 (0.000, 0.048)	0.311 (0.297, 0.322)
Randomesque	0.006 (-0.018, 0.029)	-0.007 (-0.030, 0.019)	0.334 (0.321, 0.349)
Modified Within .10 Logits	-0.002 (-0.026, 0.012)	0.003 (-0.013, 0.028)	0.332 (0.321, 0.342)
Progressive Restricted (.20)	-0.005 (-0.017, 0.011)	0.005 (-0.012, 0.019)	0.330 (0.319, 0.340)
Progressive Restricted (.30)	0.001 (-0.012, 0.014)	0.001 (-0.015, 0.013)	0.328 (0.318, 0.348)
Sympson-Hetter (.20)	-0.001 (-0.027, 0.030)	0.001 (-0.032, 0.030)	0.318 (.307, 0.334)
Sympson-Hetter (.30)	-0.011 (-0.029, 0.005)	0.012 (-0.005, 0.031)	0.307 (0.291, 0.316)

Note: Each replication contained 1,000 observations.

TABLE 20: Descriptive Statistics of the Root Mean Squared Error (RMSE) and Standardized Root Mean Squared Difference (SRMSD) for the Testlet Response Theory Model Across Ten Replications

Testlet Response Theory Model		
Exposure Control Condition	RMSE Mean (Min, Max)	SRMSD Mean (Min, Max)
Maximum Information	0.392 (0.378, 0.401)	0.666 (0.652, 0.695)
Randomesque	0.420 (0.402, 0.435)	0.692 (0.673, 0.712)
Modified Within .10 Logits	0.418 (0.406, 0.427)	0.690 (0.663, 0.707)
Progressive Restricted (.20)	0.415 (0.403, 0.426)	0.686 (0.669, 0.713)
Progressive Restricted (.30)	0.413 (0.399, 0.431)	0.686 (0.667, 0.702)
Sympson-Hetter (.20)	0.400 (0.389, 0.417)	0.670 (0.651, 0.709)
Sympson-Hetter (.30)	0.387 (0.367, 0.400)	0.660 (0.628, 0.698)

Note: Each replication contained 1,000 observations.

means for the SRMSD statistic range from 0.666 to 0.692. The small differences in the measurement statistics across the exposure control conditions are not practically significant. Overall, the exposure control conditions did not significantly impact the precision of measurement.

Exposure Rates

The exposure rate is the proportion of the number of times an item is administered to the total number of CATs administered. When examining exposure rates, a key indicator for the effectiveness of the exposure control condition is the maximum exposure rate. Table 21 lists the grand mean, minimum and maximum of the testlet exposure rates for each of the exposure control conditions. As expected the maximum information (MI) condition yielded the highest maximum testlet exposure rate, .710, indicating that some testlets were exposed to as many as 71% of the examinees. The maximum testlet exposure rate for the randomesque (RA) condition was .234 and for the modified within .10 logits (MW) condition the maximum testlet exposure rate was .233. The maximum testlet exposure rates for the RA and MW conditions reflect a high level of exposure control. The progressive restricted conditions and Sympson-Hetter conditions yielded maximum exposure rates equivalent to their restricted exposure rates assigned to the conditions, 0.20 or 0.30.

The exposure control procedures have the same grand mean because it is the ratio of test length to testlet pool size. However, the exposure control procedures do not provide the same level of exposure control. This is evident in the standard deviation of the exposure rates. Table 22 lists the mean, minimum, and maximum for

TABLE 21: Descriptive Statistics of Exposure Rates for the Testlet Response Theory Model Across Ten Replications

Testlet Response Theory Model			
	Testlet Exposure Rates		
Exposure Control Condition	Grand Mean	Minimum Testlet Exposure Rate Mean	Maximum Testlet Exposure Rate Mean
Maximum Information	0.040	0.000	0.710
Randomesque	0.040	0.000	0.234
Modified Within .10 Logits	0.040	0.000	0.233
Progressive Restricted (.20)	0.040	0.001	0.201
Progressive Restricted (.30)	0.040	0.001	0.300
Sympson-Hetter (.20)	0.040	0.000	0.216
Sympson-Hetter (.30)	0.040	0.000	0.316

Note: Each replication contained 1,000 observations.

TABLE 22: Descriptive Statistics of the Standard Deviation of the Exposure Rates for the Testlet Response Theory Model Across Ten Replications

Testlet Response Theory Model			
	Standard Deviation of the Exposure Rates		
Exposure Control Condition	Mean	Minimum	Maximum
Maximum Information	0.113	0.112	0.115
Randomesque	0.054	0.054	0.055
Modified Within .10 Logits	0.054	0.053	0.055
Progressive Restricted (.20)	0.050	0.050	0.051
Progressive Restricted (.30)	0.057	0.057	0.058
Sympson-Hetter (.20)	0.069	0.068	0.069
Sympson-Hetter (.30)	0.085	0.084	0.086

Note: each replication contained 1,000 observations.

the standard deviation of the exposure rates for each of the exposure control conditions. The MI condition yielded the highest average standard deviation (0.113) of the exposure rate, indicating the most variability in exposure rates across the testlets. The PR20 condition reported the lowest average standard deviation (0.050) of the exposure rate, indicating the least variability in exposure rates across the testlets. Ideally, the standard deviation of the exposure rates will be low indicating an even usage of testlets throughout the testlet pool.

The pool utilization represents the percentage of items not administered during any of the CAT administrations. Table 23 lists the average pool utilization and the average frequency of exposure rates for the ten data samples for each of the exposure control conditions. As anticipated, the MI condition yielded the highest percentage of testlets not administered (71%). Both the SH20 and SH30 procedures yielded high percentage of testlets never administered, 59% and 64% respectively. This indicates that examinees' performance is estimated using less than half of the testlet pool. The RA and MW procedures performed much better with percentage of testlets never administered being 32% for both procedures. Surprisingly, the PR20 and PR30 procedures used the entire testlet pool, yielding 0% of the pool never being administered. Table 23 also reports the frequency of the exposure rates. For all the exposure control conditions except MI, the majority of the testlets had exposure rates less than 0.20.

TABLE 23: Frequency of Exposure Rates for the Testlet Response Theory Model Averaged Across Ten Replications

Testlet Response Theory Model							
	Exposure Control Condition						
Exposure Rate	MI	RA	MW	PR20	PR30	SH20	SH30
.71-1.0	0	0	0	0	0	0	0
.61-.70	2	0	0	0	0	0	0
.51-.60	1	0	0	0	0	0	0
.41-.50	0	0	0	0	0	0	0
.36-.40	1	0	0	0	0	0	0
.31-.35	2	0	0	0	0	0	3
.26-.30	2	0	0	0	4	0	10
.21-.25	4	5	4	1	4	5	4
.16-.20	3	5	6	10	1	22	3
.11-.15	5	16	13	9	9	6	2
.06-.10	11	24	28	19	20	12	14
.01-.05	21	71	69	137	138	28	27
.00 (Not Admin.)	124 (71%)	56 (32%)	56 (32%)	0 (0%)	0 (0%)	104 (59%)	112 (64%)

Note: Each replication contained 1,000 observations.

Test Overlap

Test overlap, or testlet overlap, refers to the number of testlets that two examinees have in common. Each testlet was evaluated for test overlap across all examinees, examinees with similar abilities, and examinees with different abilities. Examinees' audit trails were compared to determine the test overlap. Examinees' with similar abilities were defined in two ways: examinees having theta values within two logits and examinees having theta values within one logit. Examinees with different abilities were defined as examinees with discrepancy in theta values larger than two logits or examinees with discrepancy in theta values larger than one logit.

Table 24 lists the mean, minimum, and maximum test overlap rates for all examinees, examinees with similar abilities (within two logits), and examinees with different abilities (greater than two logits). The largest overlap rates are reported by the MI condition with examinees sharing as many as three testlets on average. Yet, when ability is taken into consideration for the MI condition, the examinees with similar abilities are sharing as many as three testlets on average while examinees with different abilities have one testlet in common on average. Both the SH20 and SH30 procedures yield average overall overlap rates above one indicating that on average examinees have a similar testlet in common. When considering ability, the number of testlets shared by examinees with different abilities decreases to less than one testlet in common for the SH20 and SH30 procedures The other four conditions, RA, MW, PR20, and PR30, yielded average overall overlap rates below one indicating that on

TABLE 24: Descriptive Statistics of Test Overlap for the Testlet Response Theory Model Across Ten Replications Using Two Logits to Define Ability Groups

Testlet Response Theory Model			
	Test Overlap		
Exposure Control Condition	Overall Overlap Mean (Min, Max)	Similar[a] Abilities Mean (Min, Max)	Different[b] Abilities Mean (Min, Max)
Maximum Information	2.503 (2.453, 2.571)	2.760 (2.717, 2.809)	1.095 (1.024, 1.161)
Randomesque	0.789 (0.777, 0.802)	0.830 (0.820, 0.840)	0.565 (0.549, 0.586)
Modified Within .10 Logits	0.790 (0.772, 0.805)	0.831 (0.814, 0.844)	0.566 (0.547, 0.597)
Progressive Restricted (.20)	0.714 (0.709, 0.724)	0.771 (0.765, 0.80)	0.407 (0.388, 0.428)
Progressive Restricted (.30)	0.844 (0.837, 0.853)	0.918 (0.908, 0.933)	0.442 (0.430, 0.454)
Sympson-Hetter (.20)	1.098 (1.089, 1.109)	1.207 (1.194, 1.233)	0.504 (0.493, 0.529)
Sympson-Hetter (.30)	1.529 (1.513, 1.560)	1.699 (1.670, 1.717)	0.603 (0.561, 0.656)

Note: Each replication contained 1,000 observations.
[a]Similar = abilities within one logit.
[b]Different = abilities greater than one logit.

average examinees have less than one testlet in common. For examinees with similar abilities, these four conditions still yield overlap rates below one testlet.

In order to examine overlap rates more closely across examinees with similar abilities, the definition for similar abilities was restricted to a smaller range. Table 25 lists the mean, minimum, and maximum for examinees with similar abilities (within one logit) and examinees with different abilities (greater than one logit). The largest overlap rates are reported by the MI condition with similar examinees sharing as many as three testlets on average. The overlap rate for the SH20 and PR30 increased with the new definition of similarity. Examinees within one logit of each other on the ability scale shared two testlets on average for the SH20 condition and one testlet for the PR30 condition. The other conditions, RA, MW, and PR20, maintained overlap rates below one indicating that on average examinees have less than one testlet in common.

TABLE 25: Descriptive Statistics of Test Overlap for the Testlet Response Theory Model Across Ten Replications Using One Logit to Define Ability Groups

Testlet Response Theory Model		
	Test Overlap	
Exposure Control Condition	Similar[a] Abilities Mean (Min, Max)	Different[b] Abilities Mean (Min, Max)
Maximum Information	3.152 (3.090, 3.219)	1.774 (1.729, 1.834)
Randomesque	0.881 (0.872, 0.890)	0.688 (0.673, 0.704)
Modified Within .10 Logits	0.884 (0.866, 0.897)	0.685 (0.674, 0.711)
Progressive Restricted (.20)[b]	0.860 (0.850, 0.871)	0.553 (0.542, 0.564)
Progressive Restricted (.30)	1.041 (1.024, 1.067)	0.628 (0.617, 0.642)
Sympson-Hetter (.20)	1.379 (1.364, 1.419)	0.788 (0.777, 0.799)
Sympson-Hetter (.30)	1.968 (1.913, 2.008)	1.046 (1.010, 1.080)

Note: Each replication contained 1,000 observations.
[a]Similar = abilities within one logit.
[b]Different = abilities greater than one logit.

CHAPTER FIVE: DISCUSSION

The discussion is divided into three sections. First, the four research questions listed in the problem statement are addressed based on the results of this study. Secondly, applications for this research to practical issues are described. Finally, directions for future research are presented, limitations of this study are noted, and conclusions are drawn.

Research Questions

To what extent do the exposure control procedures impact the precision of measurement for the CAT systems based on either the three-parameter logistic testlet response theory model or the partial credit item response theory model?

Test administrators face two competing goals when implementing a CAT system. They need to ensure that examinees' performance is estimated accurately and they need to protect the item pool. Protecting the item pool involves controlling the frequency of item administrations, usually through an exposure control procedure. This often leads to administering items that are not optimal for the examinees' current ability level. The trade off for item security is less precise measurements of ability. This research investigated the impact of various exposure control procedures on the precision of measurement when using the partial credit model and the three-parameter logistic testlet response theory model.

The partial credit model yielded high levels of measurement precision across all seven exposure control conditions. The estimated thetas, standard deviations of the estimated thetas and the standard errors of ability estimates reported negligible

differences across the exposure control conditions. The bias, standardized difference between means, average absolute difference, root mean squared error, and standardized root mean squared error statistics yielded similar results across the exposure control conditions when compared to the maximum information, no exposure control condition. These results indicate that the addition of these exposure control conditions to CATs using the partial credit model to score testlets does not significantly impact precision of measurement. The results reflect that of previous research with the partial credit model (Chen, Hou & Dodd, 1998; Davis & Dodd, 2001; Davis, 2002; & Pastor, Dodd, & Chang, 2002).

The testlet response theory model yielded similar results across the exposure control conditions for precision of measurement. The exposure control conditions yielded accurate estimates of the thetas and the descriptive statistics: bias, standardized difference between means, average absolute difference, root mean squared error, and standardized root mean squared error. The results of the average absolute difference are similar to the average absolute difference results found by Wainer, Bradlow, and Du (2000) for a 3PL-TRT CAT with unequal testlet effect variances. The estimates of the standard deviations of the estimated thetas were lower than the standard deviations of the known thetas. This may be due to the use of EAP estimation in the CAT systems. The correlations between the known thetas and the estimated thetas ranged from .90 to .94. These results mirror the correlations reported by Wang, Bradlow, and Wainer (2002) when the variance of the testlet effect was 0.5

or 1.0. Overall, the inclusion of exposure controls in three-parameter logistic testlet response theory CAT systems does not significantly reduce measurement precision.

To what extent do the exposure control procedures control testlet exposure and testlet pool utilization for a CAT system based on either the three-parameter logistic testlet response theory model or the partial credit item response theory model?

The purpose of exposure control is to limit the frequency with which testlets are administered to examinees thereby maintaining the integrity of the testlet pool. If testlets are seen too often, the testlets may be compromised due to examinees sharing information about the test or by an examinee seeing the testlet again when retaking the test. The effectiveness of the exposure control procedures is measured by the exposure rates of the testlets and the overlap of testlets across examinees. In this dissertation, seven exposure control conditions, including a no exposure control condition, were investigated for use with two testlet-based measurement models.

The grand mean exposure rate was 0.047 for all exposure control conditions for the partial credit model and 0.040 for all exposure control conditions for the three-parameter logistic testlet response theory model. The grand mean exposure rate is a constant for all the exposure control conditions due to the CAT systems having fixed test lengths and fixed testlet pool sizes for the PC model (Chen, Ankenmann, & Spray, 1999).

The partial credit model yielded anticipated results for maximum exposure rates. The highest level of maximum exposure, .62, resulted from the maximum

information condition (MI), which by definition always selects the most informative items for administration. In comparison to the MI condition, the other six exposure conditions controlled the rate of testlet exposure very well. On average, the randomesque (RA) and modified within .10 logits (MW) conditions reduced the maximum exposure rate of a testlet to .20. Both of these procedures are randomization procedures and therefore easy to implement with simple calculations. The reduction in the maximum exposure rate from .62 to .20 indicated that, at most, similar testlets were being seen by 20% of examinees when implementing RA and MW. The progressive restricted and Sympson-Hetter conditions are conditional procedures, therefore by definition; they will restrict the maximum exposure rate to a pre-specified value. For the partial credit model, both of these procedures were successful at restricting the maximum exposure rates.

The differentiation in the performance of the exposure control conditions in the partial credit CAT systems is revealed through pool utilization and test overlap rates. Ideally, all the testlets in the testlet pool will be administered. This is not the case for five of the seven exposure control conditions. As anticipated, the MI condition did not use all of the testlet pool. An unexpected result was the high percentage of testlets not administered for the Sympson-Hetter conditions. For the SH20 condition, 52% of the testlet pool was never administered and for the SH30 procedure, 57% of the testlet pool was never administered. Therefore, only 48% and 43%, respectively, of the testlets were ever seen by examinees. Considering the immense cost involved in developing testlets, this does not appear to be a favorable

outcome. The RA and MW conditions performed better, yielding pool utilization rates of 28% for both. This is an improvement over the Sympson-Hetter conditions, but not ideal. The progressive restricted conditions, PR20 and PR30, met the goal of using all of the available testlets in the pool, on average.

Test overlap is another indication of the effectiveness of exposure control procedures, by determining the frequency with which pairs of examinees see the same testlet. Over all possible examinee pairs, the RA, MW, PR20, and PR30 conditions reported the least number of testlet overlap, indicating that on average examinees have less than one testlet in common. The Sympson-Hetter conditions yielded on average at least one testlet in common across pairs of examinees of the seven testlets administered. When examined more closely the discrepancy between the Sympson-Hetter procedures and the other exposure control conditions widens. Testlet overlap for pairs of examinees that have ability estimates within two logits and those with ability estimates within one logit yielded an overlap of one and half to two testlets, respectively, for the Sympson-Hetter condition, while the RA, MW, PR20, and PR30 conditions maintained a testlet overlap rate of less than one testlet on average.

The results for exposure control conditions within three-parameter logistic testlet response theory CAT systems mirror those of the partial credit model. The testlet response theory model yielded anticipated results for maximum exposure rates. The highest level of maximum exposure, .71, resulted from the maximum information condition (MI). In comparison to the MI condition, the other six exposure conditions controlled the rate of testlet exposure very well. On average, the randomesque (RA)

and modified within .10 logits (MW) conditions reduced the maximum exposure rate of a testlet to .23. Both of these procedures are randomization procedures and therefore easy to implement with simple calculations. The reduction in the maximum exposure rate from .71 to .23 indicated that, at most, similar testlets are being seen by 23% of examinees when implementing either the RA or the MW conditions. The progressive restricted and Sympson-Hetter conditions are conditional procedures; therefore they will restrict the maximum exposure rate to a pre-specified value. For the TRT model, both of these procedures were successful at restricting the maximum exposure rates.

The differentiation in the performance of the exposure control conditions in the testlet response theory CAT systems is revealed through pool utilization and test overlap rates. Ideally, all the testlets in the testlet pool will be administered. This is not the case for five of the seven exposure control conditions. As anticipated, the MI condition does not use all of the testlet pool. An unexpected result was the high percentage of testlets not administered for the Sympson-Hetter conditions. For the SH20 condition, 59% of the testlet pool was never administered and for the SH30 procedure, 64% of the testlet pool was never administered. The RA and MW conditions performed better, yielding pool utilization rates of 32% for both. This is an improvement over the Sympson-Hetter conditions, but not ideal. The progressive restricted conditions, PR20 and PR30, met the goal of using all of the available testlets in the pool, on average.

Test overlap is another indication of the effectiveness of exposure control procedures, by determining the frequency with which pairs of examinees see the same testlet. Over all possible examinee pairs, the RA, MW, PR20, and PR30 conditions reported the least number of testlet overlap, indicating that on average examinees have less than one testlet in common. The Sympson-Hetter conditions reported at least one testlet in common across pairs of examinees of the seven testlets administered. When examined more closely the discrepancy between the Sympson-Hetter procedures and the other exposure control conditions widens. Test overlap for pairs of examinees that have ability estimates within two logits and those with ability estimates within one logit yielded an overlap of one to two testlets, respectively, for the Sympson-Hetter condition, while the RA, MW, PR20, and PR30 conditions maintained a testlet overlap rate of less than one testlet on average.

Which is the optimal exposure control procedure for a CAT system based on the three-parameter logistic testlet response theory model?

The exposure control conditions incorporated in the three-parameter logistic testlet response theory CAT systems yielded similar levels of measurement precision and did not significantly decrease in the precision of measurement when compared to the no exposure control condition, MI. The use of several descriptive statistics across ten replications for each condition provides confidence in these results. The differentiation between the performances of the exposure control conditions is based on the exposure rates and their impact on pool utilization. The progressive restricted procedures restricted to a maximum exposure rate of .20 or .30 yielded the best

results. In addition, the progressive restricted exposure control procedure is simple to implement compared to other conditional selection procedures, such as the Sympson-Hetter procedure. The utilization of all the testlets and the restriction of the maximum exposure rates distinguish the PR20 and PR30 as the optimal exposure control procedures for CAT systems modeled with three-parameter logistic testlet response theory.

The choice between implementing the PR20 or PR30 should be based on the magnitude of the test and the size of the testlet pool compared to content balancing. High stakes tests that are continuously administered to many examinees often require a more restrictive exposure rate. Although the testlet response theory CAT systems did not have problems with the PR20 condition, the partial credit CAT systems, revealed a limitation to the progressive restricted procedure. Three solutions would keep the CAT systems from ending prematurely. One solution is to change the CAT algorithm to select a new content area if there are no items available for the targeted content area. Yet, this would alter the test specifications and may not be an acceptable solution for the test administrators. A second solution is to create more items for these content areas. A third solution would be to use a fixed order for selecting the content area. When selecting the maximum exposure rate for the PR procedure, it is necessary to consider the testlet pool and the frequency of available testlets within the content areas.

Does the optimal exposure control procedure for a CAT system based on the three-parameter logistic testlet response theory model differ from the optimal exposure control procedure for a CAT system based on the partial credit item response theory model?

The optimal exposure control procedure for the CAT system modeled with three-parameter logistic testlet response theory is the progressive restricted procedure. This procedure is also the optimal exposure control procedure for the partial credit model. Although, the partial credit CAT system with the PR20 procedure did not work for the majority of the replications, this is an issue dealing with the number of testlets in the pool. Recall that twenty-seven testlets were dropped due to convergence problems when calibrating the data using the partial credit model. And evidently the problem with PR20 is resolved by increasing the maximum exposure rate to .30.

Practical Applications

Recent advancements in technologies have permitted new visions in the use of computers as assessment tools. Computers provide a medium for innovative item formats including the use of interactive multimedia for graphics and sound. Test developers are no longer restricted to multiple-choice item formats, such as those frequently used in paper-and-pencil tests. Computers provide a repository for collecting data such as examinees' responses and duration of time spent on an item. Schnipke and Scrams (2002) investigated response-time analyses within CATs and recommended directions for future research that utilize such information for scoring. In CATS, items are scored immediately or in "real time." Computers allow for the

administration of performance-based items in which examinees are required to demonstrate their knowledge through practical applications. Usually, performance-based items consist of several steps or tasks to be completed successfully; therefore the steps or tasks often demonstrate local dependency. These item formats are appropriate for polytomous or testlet scoring which account for local dependencies among items. The current research lays the foundation for developing appropriate measurement models for performance-based items with local dependency and incorporating them into a CAT system based on the assessment of a unidimensional trait. Often performance-based items assess multiple abilities; therefore future research in the realm of exposure control needs to be extended to multidimensional measurement models.

The availability for examinees to experience CATs in "real time" and in the safe environment of a computer allows test administrators and test developers to pursue additional uses for CATs. For example, a CAT can be used for diagnostic purposes, such that as examinees complete items they can receive instant feedback on their performance. The feedback can provide examinees with future directions for study or validation of their abilities.

For test developers, this research provides valuable information about the use of exposure control procedures with the partial credit model and the three-parameter logistic testlet response theory model. Test developers may apply these procedures to their tests to determine the optimal methods for their item pools. The computer programming that takes place "behind-the-scenes" not only provides accurate ability

estimates for admissions to medical schools, universities, colleges, etc. It also protects test developers and test administrators from law suits from disgruntled examinees.

Conclusions, Limitations, and Directions for Future Research

This dissertation compares seven exposure control conditions in CAT systems based on the three-parameter logistic testlet response theory (TRT) model and the partial credit (PC) model. The exposure control procedures are the randomesque procedure (Kingsbury & Zara, 1989), two levels of the progressive restricted procedure (Revuelta & Ponsoda, 1998), two levels of the Sympson-Hetter procedure (Sympson & Hetter, 1985), the modified within .10 logits procedure (Davis & Dodd, 2001), and a maximum information procedure. Through realistic CAT simulations that include content balancing and expected a posteriori estimation, these exposure control conditions are evaluated based on precision of measurement and testlet exposure control. Precision of measurement is found to be similar across the exposure control conditions within the model-based CAT systems.

As anticipated, maximum information yields the best measure of precision, the highest exposure rates, and the highest percentage of items not administered. The Sympson-Hetter conditions, which are conditional procedures, maintain the pre-specified maximum exposure rate, but perform very poorly in terms of pool utilization. The randomization procedures, randomesque and modified within .10 logits, yield low maximum exposure rates, but use only about 70% of the testlet pool. Surprisingly, the progressive restricted procedure, which is a combination of both a conditional and randomization procedure, yields the best results in its ability to

maintain and control the maximum exposure rate and use the entire testlet pool. The progressive restricted procedures are the optimal procedures for both the partial credit CAT systems and the three-parameter logistic testlet response theory CAT systems. A future direction for research in the area of exposure control procedures would be to provide summary statistics for comparing exposure control procedures.

Revuelta and Ponsoda (1998) developed the progressive restricted exposure control procedure with CAT systems modeled with three-parameter logistic item response theory and ability estimation based on maximum likelihood estimation. The progressive restricted exposure control procedure needs to be examined for other CAT systems. This research expanded the application of the progressive restricted procedure to CAT systems using expected a posteriori (EAP) estimation with the three-parameter logistic testlet response theory model and the partial credit model. Further research is needed to explore the effectiveness of the progressive restricted procedure with additional polytomous models such as the graded response model and the generalized partial credit model. Each of these should be examined with both expected a posteriori estimation and maximum likelihood estimation.

This research is based on the Verbal Reasoning section of the MCAT; therefore the results of this study are limited in the generalizations that can be made. The Verbal Reasoning section consists entirely of testlets in the form of reading passages with multiple-choice items. The Biological Science and Physical Science sections of the MCAT consist partly of testlets and partly of independent multiple-choice items. The inclusion of the independent items in CAT systems with content

balancing and exposure controls similar to this study may find different optimal exposure control procedures. Continued research with simulated CATs based on real test administrations is needed not only for the other MCAT sections, but also for tests developed by other test developers. These results provide a blueprint from which other research studies can expand upon.

The three-parameter logistic testlet response theory CATs require further investigation with estimation procedures due to the reduction in the standard deviations of the estimated thetas and the correlations between known thetas and estimated thetas. Further research with maximum likelihood estimation rather then expected a posteriori estimation would inform whether the standard deviations of the estimated thetas is lower due to the TRT model or the EAP estimation.

In comparing the three-parameter logistic testlet response theory model and the partial credit model, only the correlations between the estimated thetas and known thetas can be directly examined since the models are on different scales. The partial credit model yielded higher levels of measurement precision across all of the exposure control conditions with correlations ranging from 0.95 to 0.96. The TRT model reported lower correlations ranging from 0.91 to 0.92. On a more general note, the standard errors are larger for the TRT model given a smaller range of theta then the PC model, indicating that the PC model is more accurate.

Although, the PC model provided better measurement precision, the decision to use one model over the other depends on the testing company's willingness to sacrifice measurement precision for testlet pool size. The TRT model's available

testlet pool size was 176 testlets and the PC model's available testlet pool size was 149 testlets. A larger testlet pool may be more important for test security issues than the difference in correlations between estimated thetas and known thetas for the two models. In this dissertation the impact of a larger testlet pool for the TRT model versus the PC model is not as clear when examining the test overlap results. The TRT model yielded an average of two and half testlets overlapping across examinees for the maximum information condition compared to an average of two testlets overlapping for the PC model. When taking into account the exposure control conditions, the test overlap is similar for the TRT and PC models, although slightly higher for the TRT model.

The three-parameter logistic testlet response theory model offers an advantage over the partial credit model by keeping the item as the unit of measurement, rather than the testlet being the unit of measurement. The CAT system based on the TRT model adapts the test at the testlet level rather than at the item within the testlet level. CATs based on one of the TRT models that allow selecting items adaptively *within* a testlet might further expand the functional item pool size. In addition, adapting items within a testlet may increase the precision of measurement of the three-parameter logistic testlet response theory model. Further research is warranted to fully understand the potential of the testlet response theory model.

REFERENCES

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

Chang, H. H., Qian, J., & Ying, Z. (2001). A-stratified multistage CAT with b blocking. *Applied Psychological Measurement, 25,* 333-341.

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, H. H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67 (3),* 387-398.

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, PQ, 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.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich College.

Davey, T., & Nering, M. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, & J. J. Fremer (Eds.), *Computer-Based*

Testing: Building the Foundation for Future Assessments (pp. 165-191). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Davey T., & Parshall, C. G. (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.

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

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

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.

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

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

Fennessy, L. M. (1995). *The impact of local dependencies on various IRT outcomes.* Unpublished doctoral dissertation, University of Massachusetts at Amherst. [Dissertation Abstracts International, 56-03A, p.899.]

Flaugher, R. (2000). Item pools. In Wainer, H. (Ed), *Computerized adaptive testing: A primer* (2nd ed.) (pp. 37-59). Mahwah, NH: Lawrence Erlbaum Associates.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston, MA: Allyn and Bacon.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: 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.

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.

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

Kim, J. K. & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika, 58*, 587-599.

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

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1,* 95-100.

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

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.

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

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

McBride, J. R. & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), New horizons in testing (pp.223-226). New York, Academic Press.

Mislevy, R. J., & Bock, R. D. (1983). BILOG: Item and test scoring with binary logistic models [computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J. & Stocking, (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57-75

Morrison, C. A., Subhiyah, R. G., & Nungester, R. J. (1995). *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. (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.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Bulletin 69-92). Princeton NJ: educational Testing Service.

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

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

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21,* 25-36.

Schnipke, D. L. & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer Based Testing: Building the Foundation for Future Assessments* (pp. 237-266). Mahwah, NH: Lawrence Erlbaum Associates.

Segall, D. O. & Moreno, K. E. (1999). Development of the computerized adaptive testing version of the armed services vocational aptitude battery. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in Computerized Assessment* (pp. 35-65). Mahwah, NH: Lawrence Erlbaum Associates.

Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14,* 299-311.

Sireci, S. G., Wainer, H., & Thissen, D. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247.

Smith, R. W., Plake, B. S., & De Ayala, R. J. (2001). *Item and passage selection algorithm simulations for a computerized adaptive version of the verbal section of the Medical College Admission Test (MCAT)*. MCAT Monograph Series.

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. Glass (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 (3)*, 277-292.

Stocking, M. L., & Swanson, L. (1998). Optimal design of item banks for computerized adaptive tests. *Applied Psychological Measurement, 22 (3),* 271-279.

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 to the annual meeting of the American Educational Research Association, San Diego.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association,* 82, 528-540.

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

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

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NH: Lawrence Erlbaum Associates.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Psychological Measurement*, 8 (2), 157-187.

Wainer H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glass (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 245-269). Netherlands: Kluwer Academic Publishers.

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

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

Wainer, H. & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NH: Lawrence Erlbaum Associates.

Wang, X., Bradlow, E. T., & Wainer, H. (2001). The SCORIGHT computer program [Computer program]. Princeton, NJ: Educational Testing Service.

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

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473-492.

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

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213.

Yen, W. (1994). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8 (2),* 125-145.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement, 39(4),* 291-309.

VITA

Aimee Michelle Boyd was born in St. Louis, Missouri on September 21, 1974, the daughter of Loretta Margaret Summers and Donald Lee Summers. After completing her work at Carroll High School, Southlake, Texas, in 1993, she entered Saint Mary's College, Notre Dame, Indiana. During the 1994-1995 school year, she attended Saint Patrick's College, Maynooth, Ireland. She received the degree of Bachelor of Science from Texas A&M University in May 1998. During the following year she was employed as a research assistant at PRO-ED, Inc. In January 1999, she entered the Graduate School at The University of Texas. She worked as a graduate research assistant at the IC^2 Institute at The University of Texas from September 1999 until May 2003. She received the degree of Master of Arts from The University of Texas at Austin in August 2001.

Permanent Address: 7920 San Felipe Blvd.
Apt. 1806
Austin, Texas 78729

This report was typed by the author.