Running Head: ITEM EXPOSURE STRATEGIES WITH THE GPC MODEL

Strategies for Controlling Item Exposure in Computerized Adaptive Testing with the Generalized Partial Credit Model

Laurie Laughlin Davis

Pearson Educational Measurement

Send all correspondence to:

Laurie Laughlin Davis
Pearson Educational Measurement
2201 Donley Drive, Suite 100
Austin, TX 78758

Paper presented at the annual meeting of the National Council on Measurement in Education, April 21-25, 2003 Chicago, Illinois.

Abstract

Choosing a strategy for controlling item exposure has become an integral part of test development for CAT. Item exposure can be controlled through the use of a variety of algorithms that modify the CAT item selection process. The current study investigated the performance of six procedures for controlling item exposure in a series of simulated CATs under the generalized partial credit model. In addition to a no-exposure control baseline condition, the randomesque, modified-within-.10-logits, Sympson-Hetter, conditional Sympson-Hetter, a-Stratified with multiple-stratification, and enhanced a-Stratified with multiple-stratification procedures were implemented to control exposure rates. Two variations of the randomesque and modified-within-.10 logits procedures were examined which varied the size of the item group from which the next item to be administered was randomly selected. The results indicate that the randomesque and modified-within-.10-logits procedures with the six-item group variation provide the best option for controlling exposure when impact on measurement precision and ease of implementation are considered.

Strategies for Controlling Item Exposure in Computerized Adaptive Testing with the Generalized Partial Credit Model

Computerized Adaptive Testing (CAT) is based on the premise that items that are too easy or too difficult contribute little to the information about an examinee's trait level. By eliminating the need to administer items of inappropriate difficulty, CAT can shorten testing time, increase measurement precision, and reduce measurement error due to boredom, frustration, and guessing (Wainer, 1990). However, CAT creates special circumstances that can lead to overexposure of certain items within the item pool, resulting in a threat to validity. The very adaptive algorithms that make trait-level estimation in CAT so efficient also tend to produce variability in the frequency with which items are used. Because each item has its own set of characteristics definable by the parameters of a given model, using maximum information to select items will cause items to differ in terms of the desirability of these characteristics for measuring an examinee's trait level.

When combined with the near continuous nature of CAT administration necessary for large volume testing programs, uneven item use presents a problem for test security. Those items which are more popular (those with the most desirable characteristics for measuring trait level) will be administered quite frequently, whereas some items may be administered less frequently or never administered at all. This results in a differentiation between the available item pool (all items available for administration) and the functional item pool (those items that will most often be administered). Wainer and Eignor (2000) reported on the extent of this phenomenon in the GRE CAT and an experimental version of the SAT CAT. They found that as few as 12% of the available item pool can account for as much as 50% of the functional item pool and that as few as 33% of the available item pool can account for as much as 75% of the functional item pool.

Parshall, Davey, and Nering (1998) discussed the three often conflicting goals of item selection in CAT. First, item selection must maximize measurement precision, by presenting the item that maximizes information for the examinee's current trait level. Second, item selection must seek to protect the security of the item pool by limiting the degree to which items may be exposed. Third, item selection must ensure that examinees will receive a content balanced test. Stocking and Swanson (1998) added a fourth goal to this list, stating that item selection must also maximize item usage so that all items in a pool are used, thereby ensuring good economy of item development. Stocking and Lewis (2000) likened the item selection problem to squeezing an inflated balloon—pushing in at one point will cause a bulge to appear at another.

Overview of Exposure Control Methods

Different approaches to the goals of item selection will produce different testing algorithms (Stocking & Lewis, 2000). Attempts to address the second goal are denoted exposure control methodologies (Parshall, Davey, & Nering, 1998). Many algorithms exist that seek to control item exposure through constraining the administration of more popular items. These algorithms differ in terms of their complexity and the variables taken into consideration. Way (1998) discussed two types of exposure control strategies—randomization and conditional selection. Randomization strategies operate by randomly choosing the next item for administration from a set of nearly optimal items rather by than selecting the single most informative item. Conditional selection strategies operate by controlling the probability of an item being administered conditional on a given criterion, such as the expected frequency of item use. A third approach to exposure control has been proposed by Chang and Ying (1996) in which items in the pool are stratified according to their statistical properties (item parameters) and items are constrained to be administered from certain strata.

Randomization procedures are usually considered to be easy to understand and simple to implement but provide no guarantee that item exposure will be constrained to a given level. Kingsbury and Zara (1989) proposed a method for controlling item exposure, which they termed the "randomesque" procedure that randomly selects the next item to be administered from a predetermined group of the most informative items. The size of the group from which an item is randomly selected for administration is not explicitly stipulated; however, Kingsbury and Zara (1989) initially implemented the procedure with an option for item group sizes between 2 and 10.

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

Conditional selection strategies allow a preset target exposure rate to be stipulated and provide a reasonable guarantee that exposure will be constrained to this level. However, these procedures usually involve time-consuming simulations which must be run prior to operational testing. The most well-known conditional selection procedure is undoubtedly the Sympson-

Hetter (Sympson & Hetter, 1985; Hetter & Sympson, 1997), which attempts to directly control the rate of item exposure through the use of probabilistically-determined exposure control parameters assigned to each item. The goal of the Sympson-Hetter is to constrain the maximum probability of administration for an item to a predetermined level. The target exposure rate (r) stipulates what maximum proportion of the population should see each item (Hetter & Sympson, 1997).

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

Parshall, Davey, and Nering (1998) pointed out that the Sympson-Hetter tends to produce more relaxed exposure control parameters for items whose information peaks at the tails of the distribution, because of the sparseness of examinees at those trait levels. While the item may not be seen by more than the targeted proportion of the population, almost all examinees of low or high trait level will see it. This is especially problematic in test-retest scenarios in which an examinee's true trait level does not change and he or she is likely to be exposed to the same items. Additionally, test takers are more likely to have friends of similar trait level, so shared information about test questions may be more beneficial for increasing scores than the global exposure rate would predict. These observations resulted in the development of an extension to the Sympson-Hetter procedure, which establishes exposure control parameters conditional on trait level (Parshall, Davey, & Nering, 1998; Chang, Ansley, & Lin, 2000). It is important here to distinguish between conditional selection strategies that select items based on a given condition, such as frequency of administration, and conditional exposure control that provides control of item exposure conditional on trait level. While most attempts to control exposure conditional on trait level have been implemented with conditional selection strategies, it is possible to implement conditional exposure control with a randomization or stratification procedure.

The conditional Sympson-Hetter procedure is conducted in exactly the same fashion as the Sympson-Hetter procedure with the exception that the frequency of item administration is tallied separately for each of “m” discrete trait levels. The simulations to develop the exposure control parameters are then conducted separately for each level of theta. The result is a matrix of n (items) X m (theta values), which yields the conditional exposure control parameters. In operation, the examinee’s current trait estimate determines which column of parameters will be used.

The stratification procedures, typified by the a-Stratified design (Chang & Ying, 1996), require test developers to partition an item pool into “k” strata based on the value of the item discrimination parameter. Strata are then arranged in ascending order of discrimination. The test is divided into stages to match the number of strata such that a given number of items are chosen from each stratum, with the lowest discriminating stratum used at the beginning of the test and the highest discriminating stratum used at the end of the test.

A number of variations have been built on the original framework of the a-Stratified design. Chang, Qian, and Ying (1999) developed the a-Stratified design with b-blocking which divides the item pool into blocks based on item difficulty and then stratifies items by discrimination within each block. Yi and Chang (2000) implemented a multiple-stratification variant to allow for content balancing. Leung, Chang, and Hau (1999) incorporated the Sympson-Hetter into the a-Stratified design in a procedure they termed the Enhanced a-Stratified design. In this method, Sympson-Hetter item exposure control parameters are set through pre-operational CAT simulations with the item pool partitioned into the desired strata.

Exposure Control with Polytomous Items

As computerized adaptive testing moves out of its adolescence, there is a push to incorporate more performance-oriented items and simulations into large-scale CAT programs—to utilize the full potential of the technology to deliver more authentic assessments. The majority of CAT programs that have included these sorts of innovative item types continue to provide only a dichotomous score based on the end state of the task. However, the advent of complex scoring algorithms allows for online automated scoring of partial-credit items such as essay questions, teslets, simulations, and constructed-response math items that may enable the use of such items in a CAT environment (Bejar, 1991; Burstein, Kukich, Wolff, Lu, & Chodorow, 1998; Clauser, Margolis, Clyman, & Ross, 1997).

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

Two recent studies (Pastor, Chiang, Dodd, & Yockey, 1999; Davis, Pastor, Dodd, Chiang, & Fitzpatrick, 2003) examined the Sympson-Hetter under the partial credit model. The authors concluded that, while it did not have a negative impact on measurement precision, it was difficult to implement and relatively ineffective in controlling exposure rates. They concluded that difficulties encountered in the implementation of the Sympson-Hetter mechanism, especially problems with convergence of the exposure control parameters in small item pools, were not outweighed by the observed gains in test security.

Pastor, Dodd, & Chang (2002) sought to evaluate variations of the a-Stratified design as a simpler alternative to the Sympson-Hetter with the generalized partial credit model. The study compared the a-Stratified design to both the Sympson-Hetter and the enhanced a-Stratified design. In addition, conditional exposure control was examined using both the conditional Sympson-Hetter and the newly proposed conditional enhanced a-Stratified design, which incorporated the conditional Sympson-Hetter algorithm into the a-Stratified design. While the a-Stratified design controlled exposure rates better than the no-exposure control condition, other methods were generally superior. Convergence problems were again observed when exposure control parameters for the conditions involving the Sympson-Hetter and conditional Sympson-Hetter were established.

Davis and Dodd (2001) evaluated the performance of a variant of the Lunz and Stahl (1998) within-.10-logits randomization procedure with the partial credit model. The procedure was modified to work with the polytomous case. Rather than forming an item group of all items within .10 logits of the needed item difficulty, the authors used maximum information item selection to select the two most informative items at each of three points along the theta scale: estimated theta, estimated theta minus 0.10 logits, and estimated theta plus 0.10 logits. This resulted in a group of six passages from which one was randomly selected for administration. Results indicated that, while measurement precision was somewhat reduced over the no-exposure control condition, the procedure was successful in reducing exposure and overlap rates and increasing pool utilization. Somewhat promising was the finding that maximum exposure rates were reasonably controlled even without the use of exposure control parameters.

The smaller nature of polytomous item pools appears to impact the implementation of conditional selection strategies because of convergence problems observed in establishing the exposure control parameters. It is also questionable whether procedures such as the Sympson-Hetter are even effective in controlling item exposure in polytomous pools (Davis, et al., 2003). It is, therefore, highly desirable to investigate alternative procedures to the Sympson-Hetter. The results of the Davis and Dodd (2001) study indicated that the modified-within-.10-logits procedure may, in fact, yield sufficient test security with polytomous item pools. Further research is needed, however, to evaluate other randomization procedures with polytomous pools and to examine the item group size from which the next item to be administered will be randomly drawn. The current study attempts to address these issues by evaluating the performance of a series of exposure control procedures that include representatives from all three classifications (randomization, conditional, and stratification) with the generalized partial credit model (Muraki, 1992).

The Generalized Partial Credit Model

The generalized partial credit model (Muraki, 1992) is an extension of the two-parameter logistic model to the polytomous case. The probability function of scoring in category *x* on item *i* given the examinee's trait level, θ, for the partial credit model is defined as:

$$Pix(\theta) = \frac{\exp\left[\sum_{k=0}^{x} a_i(\theta - bik)\right]}{\sum_{h=0}^{mi} \exp\left[\sum_{k=0}^{h} a_i(\theta - bik)\right]}, \qquad (1)$$

where m_i is the number of score categories minus one, b_{ik} is the difficulty parameter associated with score category *x*, and a_i is the item discrimination.

Method

Overview of Techniques

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

Item pool

Data were obtained from 22 forms of the Verbal Reasoning section of the Medical College Admissions Test (MCAT), used in six administrations from April, 1996, through April, 2001, to assemble an item pool consisting of 157 polytomously scored items. Responses from examinees to the passage-based items were clustered and scored by passage in a “testlet” fashion. Due to previously observed problems with low category frequencies in the extremes of the response distribution, which caused difficulties with item calibration (Davis & Dodd, 2001), item responses were collapsed across categories and recoded. The result was an item pool that consisted of 63% 3-category items, 18.5% 4-category items, and 18.5% 5-category items. Original classifications that indicated an item’s membership in one of three content areas (humanities, social sciences, natural sciences) were retained to preserve any naturally existing differences in item difficulty by content area. Of the 157 items, 39% were classified as humanities, 37.5% as social sciences, and 23.5% as natural sciences.

Data Generation

Response data for the 157 items was generated using conventional techniques for generating responses to polytomous items (see Dodd & Koch, 1987, for more detail). Four data sets were generated for use in the current study. The first data set (N(0,1); N=7,500) was used as the calibration sample to obtain item parameter estimates for use in the CATs. The second data set (N(0,1); N=8,000) was used to set exposure control parameters for the Sympson-Hetter and enhanced a-Stratified designs. The third data set (N=15,000) was used to set exposure control parameters for the conditional Sympson-Hetter. For the conditional Sympson-Hetter, exposure control parameters are set independent of the trait level distribution, and, therefore, to attain the same level of precision for the exposure control parameters at all levels of theta, the sample of simulees should be uniformly distributed across the theta levels. For the current study, 1,000 simulees were drawn at each of fifteen discrete theta levels in order to set the exposure control parameters for the conditional Sympson-Hetter. The fourth data set (N(0,1); N=1,000) was generated for use in each of the CAT conditions.

Parameter Estimation

Responses from the N=7,500 calibration samples to the 157 items were separately submitted to PARSCALE (Muraki & Bock, 1993) for calibration according the generalized partial credit model.

Stratification of the Item Pool

For the a-Stratified and enhanced a-Stratified conditions, the pool was divided into five strata with 32 items in each of the first two strata and 31 items in each of the remaining three strata. Yi and Chang (2000) developed a modification to the a-Stratified design in which items are stratified according to multiple factors that might include not only item discrimination, but also item difficulty and item content associations. This multiple-stratification modification was designed to control for any variation in item difficulty across content areas, which might result in a poor distribution of item contents across the strata.

In the current study, it was found that a moderately strong negative correlation (-0.54) existed between the estimated item discrimination parameter and the number of categories, which resulted in the more discriminating strata containing items with fewer categories and the less discriminating strata containing items with more categories. This caused problems due to the use of a content balancing mechanism that insured that each test administered reflected an appropriate distribution of items both in terms of content and number of categories. It was determined that the Yi and Chang (2000) method of multiple-stratification could be adapted to control for this phenomenon. Therefore, before being stratified according to discrimination, items were first stratified according to content area and number of categories. A triple stratification of the pool was implemented so that each stratum would contain a sufficient number of items from each content area and of each number of categories. Combining the three levels of content with the three levels of categories resulted in nine unique sets of item characteristics. The item pool was separated into nine subpools representing each of these item types. Each of these nine subpools was then sorted independently by discrimination. The five strata were then formed by pulling items from each of the sorted subpools into the appropriate stratum—the 3 or 4 least discriminating items from each subpool were used to make the 1st stratum; the 3 or 4 highest discriminating items of each type were used to construct the 5th stratum.

Setting the exposure control parameters

Sympson-Hetter and Enhanced a-Stratified Conditions

Responses from the N=8,000 data set were used with the estimated item parameters in setting the exposure control (K_i) parameters for the Sympson-Hetter and enhanced a-Stratified conditions through a series of iterative CAT simulations. Hetter and Sympson (1997) reported that the maximum probability of administration will approach a value slightly above the target exposure rate. While several operational CAT testing programs using dichotomous items have selected a target exposure rate of .20 (Stocking, 1992), the smaller nature of the polytomous item pool in the current study necessitates a more liberal criterion. For the current research, a target exposure rate, r, of 0.39 was, therefore, set for each of the conditions to ensure that the maximum probability of administration would converge to the 0.40 level. It should be noted that this target exposure rate was even higher than the 0.30 level used in previous research exploring the

Sympson-Hetter in polytomous CAT (Pastor, et al., 1999; Davis, et al., 2003; Pastor, et al., 2002), but was chosen after initial experimentation with more restrictive levels of exposure control failed to produce convergence of the exposure control parameters.

For the first iteration, all K_i's were set equal to 1.0 so that every item that was selected through maximum information item selection would actually be administered. This provided a baseline exposure rate for all items in the pool. After each iteration, the probability of selecting each item, P(S), was computed by dividing the frequency of selection by the number of simulees (in this study, 8,000). Based on the probability of selection, new K_i's were computed such that if an item's P(S) was less than or equal to the target exposure rate (r=.39), K_i was set equal to 1.0. Otherwise, if an item's P(S) was greater than our target exposure rate (r=.39), K_i was set equal to r/P(S).

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

Conditional Sympson-Hetter

Responses from the N=15,000 data set were used with the estimated item parameters in setting the exposure control (K_i) parameters for the conditional Sympson-Hetter. This data set was designed to have 1,000 simulees at each of 15 levels of theta—from –3.5 to 3.5 logits in 0.5 logit increments. Iterations to set the exposure control parameters for the conditional Sympson-Hetter were conducted in the same fashion as those for the Sympson-Hetter, with the exception that the probability of selection was computed separately for each trait level and a separate exposure control parameter was computed for each item at each trait level. The target exposure rate for each theta level was set to 0.39. All K_i's were set equal to 1.0 for the first iteration and new K_i's computed at each iteration by comparing the probability of selection to the target exposure rate.

CAT simulations

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

The Kingsbury and Zara (1989) constrained CAT (CCAT) content balancing method was used for all conditions. Two factors were jointly balanced with this method—content area affiliation and number of categories per item. By combining the three levels of content area affiliation (humanities, social sciences, natural sciences) with the three levels of numbers of categories (3, 4, or 5), nine unique sets of item characteristics were produced (humanities with 3 categories, social sciences with 5 categories, etc). The specified content area/number of categories for the first item administered was randomly chosen. After each item administration, the proportion of each of the nine item types given was computed and compared to the target desired proportion. The next item administered was constrained to be chosen from the area with the largest discrepancy. Target proportions for each of the nine item types were defined to match the observed percentages of each characteristic in the item pool. Item selection is described for each of the conditions below.

No-Exposure Control (Maximum Information)

In the no-exposure control condition, items were chosen to maximize the information at the current trait estimate.

Randomesque

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

Modified-Within-.10-logits

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

Sympson-Hetter

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

Conditional Sympson-Hetter

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

a-Stratified design with multiple- stratification

The a-Stratified design was implemented in this study according to an adaptation of the multiple-stratification procedure (Yi & Chang, 2000). There were five stages of testing, with four items administered from each of the five increasingly discriminating strata. Within a stratum, items were selected by maximum information item selection.

Enhanced a-Stratified design with multiple-stratification

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

While the triple stratification of the pool by content, number of categories, and discrimination parameter, described above in the Stratification of the Item Pool section, ensured that enough items were present within each stratum to meet the content balancing constraints, the inclusion of the Sympson-Hetter exposure control parameters in the enhanced a-Stratified design could result in those items being unavailable for administration. In the event that no item within the current stratum met the desired content and category constraints and whose comparison between K_i and the random number allowed it to be administered, a method of backward and forward searching through the strata was implemented (Leung, Chang, & Hau, 2000; Leung, 2001). This method allowed items from other strata to be considered for administration when no item from the current stratum could be administered. Items from previous strata were first considered in a step-down fashion from the current stratum, with items from successive strata considered only if no item from either the current or all previous strata could be administered. For example, if the current stratum was 3, the order in which items from other strata would be considered would be stratum 2, stratum 1, stratum 4, and finally stratum 5.

Data Analyses

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

$$Bias = \frac{\sum_{k=1}^{n}(\hat{\theta}_k - \theta_k)}{n}, \qquad (2)$$

$$RMSE = \left[\frac{\sum_{k=1}^{n} (\hat{\theta}_k - \theta_k)^2}{n} \right]^{1/2}, \qquad (3)$$

where $\hat{\theta}_k$ is the estimate of trait level for simulee k, θ_k is the known trait level for simulee k, and n is the total number of simulees.

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

In order to measure item overlap, the audit trail of each simulee was compared to the audit trail of every other simulee. A data file containing the number of items shared among the simulees as well as the difference between their known theta values was created to obtain an index of item overlap conditional on theta. Simulees were defined to have “similar” trait levels when their known thetas differed by two logits or fewer and “different” trait levels when their known thetas differed by more than two logits (Pastor, et al., 1999; Davis, et al., 2003).

Results

Descriptive Statistics for the Item Pool

Table 1 lists the mean, standard deviation, minimum, and maximum values of the item discrimination and step difficulties for all 157 items in the pool. Most of the items (99 out of 157) had only two step difficulties; therefore, the values for the third, and fourth step difficulties are only given for the remaining 58 and 29 items, respectively. The test information function for the item pool calibrated according to the generalized partial credit model is plotted in Figure 1. The item pool information peaked at a theta value of –0.6.

[Insert Table 1 and Figure 1 about here]

Table 2 gives the mean, standard deviation, minimum, and maximum values for the item discrimination parameters broken out by strata. Due to implementation of the modified Yi and Chang (2000) multiple-stratification to the a-Stratified design, which stratified the item pool by content and number of categories as well as by discrimination, item discrimination values may overlap across strata. Yi and Chang (2000) stated that this overlap is acceptable so long as the average item discrimination increases across strata. As can be seen from the table, while there is some degree of overlap across the strata, the average 'a'-value does increase as the strata increase.

[Insert Table 2 about here]

Exposure Control Parameters

Initial attempts were made to set the target exposure rate for the Sympson-Hetter condition to more restrictive levels (0.19 and 0.29), but after 30 iterations the observed maximum probability of administration did not approach the desired level. A target exposure rate of 0.39 was, therefore, chosen to ensure converge to the 0.40 level. The maximum probability of administration after the 30th iteration with a target exposure rate of 0.39 was 0.404. A target exposure rate of 0.39 was also set for the enhanced a-Stratified condition, which yielded a maximum observed probability of administration of 0.399 after the 30th iteration. For the conditional Sympson-Hetter condition, conditional maximum exposure rates of 0.39 were chosen and 30 iterations were conducted to ensure convergence of the parameters to this level. Stocking and Lewis (1998) reported that the conditional maximum probability of administration consistently converged to between 0.05 and 0.10 above the target value. In the current study, the maximum probability of administration at each theta level converged to values slightly greater than the target (0.420-0.498).

Descriptive Statistics for the Exposure Control Conditions

A listwise deletion of 191 nonconvergent cases was conducted after all conditions had been run. A case was defined as nonconvergent if, once the end of the test had been reached, the trait estimate was greater than or equal to 4.0 or less than or equal to –4.0, or if maximum likelihood estimation had never been reached. Table 3 lists the number of nonconvergent cases for each condition. The number of nonconvergent cases was somewhat high, ranging from 34 to 47 across conditions, with the exception of the a-Stratified and enhanced a-Stratified conditions where the number of nonconvergent cases was 2 and 10, respectively. The remainder of the results are reported on the sample (N=809) of observations that remained after the nonconvergent cases had been deleted.

Table 3 also provides the average theta estimate and standard error for each exposure control condition. The mean of the known thetas was 0.03 with a standard deviation of 1.03. For all conditions a slightly increased average theta estimate resulted, with the no-exposure control condition producing the average theta estimate closest to the known (0.05) and the randomesque-6 and modified-within-.10-logits-6 conditions producing average theta estimates furthest from the known (0.08 and 0.09 respectively). Standard deviations of the theta estimates were also slightly inflated relative to the standard deviation of the known thetas and ranged from 1.09 to 1.12. As expected, the no-exposure control condition yielded the lowest average standard error, 0.28, with the a-Stratified and enhanced a-Stratified conditions yielding the highest average standard errors (0.33 for both procedures). The randomesque-3, modified-within-.10-logits-3, and Sympson-Hetter conditions all yielded average standard errors of 0.29, while the randomesque-6, modified-within-.10-logits-6, and conditional Sympson-Hetter conditions yielded average standard errors of 0.31.

[Insert Table 3 about here]

Table 4 states the correlations between known and estimated theta for each exposure control condition as well as values for the bias and RMSE. The correlation between known and estimated theta was highest for the no-exposure control condition (0.96) and lowest for the conditional Sympson-Hetter condition (0.93). Both the modified-within-.10-logits-3 and the Sympson-Hetter conditions yielded correlation coefficients comparable to the no-exposure control condition. Bias was functionally zero across conditions. The no-exposure control condition shared the lowest RMSE value with the Sympson-Hetter condition (0.29). The modified-within-.10-logits-6 and the conditional Sympson-Hetter conditions had the highest RMSEs (0.39 and 0.40, respectively).

[Insert Table 4 about here]

Pool Utilization and Exposure Rates

Table 5 presents the frequency of observed exposure rates along with the average, maximum, and standard deviation of exposure rates, and the percent of pool not administered for each condition. Given that test length was the same for all conditions studied, the observed average exposure rates did not differ and were equal to the test length (20) divided by the item pool size (157), or 0.127 across all conditions (Chen, Ankenmann, & Spray, 1999). The no-exposure control condition demonstrated the highest standard deviation of exposure rates (0.206), signifying the most uneven item use. The lowest values for standard deviation of exposure rates were those for the randomesque-6, modified-within-.10-logits-6, and conditional Sympson-Hetter conditions (0.12 for all three conditions), signifying the most even item use.

The maximum exposure rate of any item was highest for the no-exposure control condition, with a value of 0.878, and lowest for conditional Sympson-Hetter condition, with a value of 0.407. The Sympson-Hetter, enhanced a-Stratified design, randomesque-6, and modified-within-.10-logits-6 conditions also had relatively low maximum exposure rates (0.419, 0.424, 0.477, and 0.482, respectively). The randomesque-3, within .10 logits-3, and a-Stratified design conditions, on the other hand, resulted in maximum exposure rates above acceptable levels (0.705, 0.713, and 0.735, respectively).

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

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

[Insert Table 5 about here]

Item Overlap

Pairwise comparisons of examinee audit trails were made for each simulee with every other simulee, resulting in 326,836 pairwise comparisons per condition. Table 6 contains the average item overlap for all simulees, for those of different trait levels (known thetas differed by more than two logits), and for those of similar trait levels (known thetas differed by two logits or fewer) for each condition. Information regarding item overlap information is presented both in terms of average number of items shared across a 20-item test and average percent of items shared across a 20-item test.

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

[Insert Table 6 about here]

Discussion

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

Randomization Procedures

The randomization procedures, represented in this study by the randomesque and modified-within-.10-logits methods, were easy to implement and required only a minimum of additional programming over that needed to run the CAT. One of the driving questions of this study was whether these randomization procedures could be as effective for protecting test security as the conditional selection strategies, such as the Sympson-Hetter procedure. The answer to this question seems to be overwhelmingly positive. The randomesque-6 and modified within .10 logits-6 procedures provided relatively good control over exposure and overlap rates and the highest degree of pool utilization of any procedures examined in the current study.

A second question posed by this study was to address what impact item group size would have on the effectiveness of the randomization procedures for controlling exposure. It is evident from the results that item group size does have strong implications for the ability of these procedures to control exposure, with an item group size of 6 providing superior test security to an item group size of 3 for both the randomesque and modified-within-.10-logits procedures. Maximum exposure rates for the 3-item group variations were on the order of 0.20-0.25 larger than for the 6-item group variations. Overlap rates for the 3-item group variations were 9-10%

larger than for the 6-item group variations. Finally, the percent of pool not administered was twice as large for the 3-item group variations as for the 6-item group variations.

While these procedures were quite effective for improving test security, a corresponding decrease in measurement precision was observed. The average standard error increased from 0.28 to 0.31 over the no-exposure condition with the use of either the randomesque-6 or modified-within-.10-logits-6 procedures. Further, the correlation between known and estimated theta decreased and the RMSE increased.

It appears that the randomesque procedure may have a very slight edge over the modified-within-.10-logits procedure both in terms of its effectiveness in controlling exposure and in terms of its impact on measurement precision. However, these differences are extremely small and both procedures performed very well.

Conditional Selection Procedures

The Sympson-Hetter and conditional Sympson-Hetter procedures both proved to be challenging and time-consuming to implement. The simulations had to be run multiple times when convergence of the parameters to the initial target exposure rates was not achieved. In addition, the programming requirements for these procedures were substantial.

The performance of the Sympson-Hetter procedure with regard to test security was disappointing given the amount of effort required. The maximum observed exposure rate for the Sympson-Hetter condition was slightly above, but close to, the target value. Overlap rates and percent of pool not administered, while reduced over the no-exposure control condition, were still high relative to other exposure control procedures investigated in the current study.

Measurement precision for the Sympson-Hetter condition remained relatively high. However, the rather lenient target exposure rates used in this study might provide at least some explanation for this. Had more restrictive levels of exposure control been possible, it is anticipated that measurement precision would have suffered more of an impact.

The conditional Sympson-Hetter procedure faired somewhat better than its unconditional counterpart. The global maximum exposure rate was controlled to levels below those seen with the Sympson-Hetter even though the procedure only guarantees conditional target exposure rates. The conditional Sympson-Hetter method has been criticized because its lack of global exposure control can allow global exposure rates to exceed acceptable levels; however, that did not appear to be the case in this study. Further, the conditional Sympson-Hetter procedure did a much better job of controlling overlap rates than did the Sympson-Hetter procedure, providing overlap rates equivalent to those seen with the randomesque-6 and within .10 logits-6 procedures. Finally, pool utilization is much improved with the use of the conditional Sympson-Hetter procedure, yielding rates equivalent to or just slightly above those seen with the randomesque-6 and modified-within-.10-logits-6 procedures.

The tradeoff for this level of test security came in the form of lower measurement precision. The average standard error increased from 0.28 to 0.31, the correlation between known and estimated theta dropped from 0.96 to 0.93, and the RMSE increased from 0.29 to 0.40 over the no-exposure control condition.

Stratification Procedures

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

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

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

Conclusions and Directions for Future Research

The results of this study provide strong support for the use of randomization procedures with sufficient item group sizes to control test security with polytomous item pools. The randomesque-6 and modified-within-.10-logits-6 procedures proved themselves to be simple to implement and provided substantial reductions in exposure rates and item overlap as well as substantial increases in pool utilization over the no-exposure control condition. Further, when compared head-to-head with other options for controlling exposure, they demonstrated comparable or better performance on almost all measures of test security. For those measures where performance was somewhat less satisfactory than with other procedures, this small cost in performance has to be weighed against the expense of added complexity that would be necessary to adopt the competitive option. Only the conditional Sympson-Hetter procedure was competitive with the randomization procedures. While it did produce slightly lower maximum exposure rates and some comparable pool utilization and overlap rates, the conditional Sympson-Hetter procedure is the single most complex procedure examined in the current study. Further, the negative impact on measurement precision seemed to be somewhat greater for the conditional Sympson-Hetter procedure than for the randomesque-6 or modified-within-.10-logits-6 procedures. The results of the current study, therefore, indicate that the randomesque-6 and modified-within-.10-logits-6 procedures provide the best all-around options for controlling test security when ease of implementation and impact to measurement precision are considered.

However, it is important to point out that none of the procedures examined was able to control exposure rates to levels traditionally acceptable with dichotomous item pools (0.20) or to levels previously used with polytomous models (0.30) with the test structure used in this study. A maximum observed exposure rate of approximately 0.40 was the lowest rate that could be obtained with any of the procedures. This suggests that the size and structure of the item pool with regard to number of categories and content play a large role in determining the ability of test developers to control item exposure. Also, it is important to recognize that, in order to achieve these gains in test security, some measurement precision was sacrificed. As Parshall, Davey, and Nering (1998) have suggested, there seems to be an unavoidable tradeoff between exposure control and measurement precision. None of the procedures examined demonstrated the ability to overcome this tradeoff, and, therefore, a balance between these goals is the best that could be obtained.

Future research should be conducted to determine how different pool sizes and characteristics would affect the utility of the various exposure control mechanisms. The item pool size and characteristics examined in the current study were selected because they represented a realistic test structure that might be used by a large-scale testing program. However, it is quite clear that the content and category constraints imposed by this structure severely affected the ability of certain procedures to control test security.

This research has identified two strategies that seem to be useful for controlling exposure in small polytomous item pools—the randomesque and modified-within-.10-logits procedures. However, the results have also shown that the item group size from which the next item to be administered is randomly chosen has a significant impact on the effectiveness of the strategies. The goal of future research should be to provide a more detailed study of the item group size variable, both by itself and in relation to the size of the pool, to attempt to develop a set of recommendations for implementing the procedures. Other randomization procedures, such as Revuelta and Ponsoda’s (1998) progressive and restricted maximum information procedures, should be examined with polytomous item pools to determine whether they can offer similar benefits for exposure control.

References

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

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

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

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

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

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

Chen, S. Ankenmann, R.D., & Spray, J.A. (1999, April). Exploring the relationship between item exposure rate and test overlap rate in computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

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

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

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

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

Elliot, S. (2003, February). The next steps beyond computer-scored essays. Presented at The Association of Test Publishers conference—Technology in Testing: Application and Innovation, Amelia Island, FL.

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.

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

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

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.

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.

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

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

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

Pastor, D.A., Dodd, B.G., & Chang, H.H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. Applied Psychological Measurement, 26(2), 144-163.

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

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

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

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.

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

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

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

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

Table 1

Mean, Standard Deviation, Minimum, and Maximum for Item Parameter Estimates

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

Table 2

Mean, Standard Deviation, Minimum, and Maximum of Item Discrimination Parameters Across Strata

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

Table 3

Means (and Standard Deviations) for Estimated Theta and Standard Error, with Number of Nonconvergent Cases for the Exposure Control Conditions

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

*Note: For Known Thetas (N=809) Mean=0.03; SD=1.03

Table 4

Correlation Coefficients between Known and Estimated Theta, with Bias and RMSE

Exposure Control Condition	***Correlation***	***Bias***	***RMSE***
No-Exposure Control	0.96	-0.03	0.29
Randomesque-3	0.95	-0.05	0.36
Randomesque-6	0.95	-0.06	0.35
Within .10 Logits-3	0.96	-0.04	0.30
Within .10 Logits-6	0.94	-0.06	0.39
Sympson-Hetter	0.96	-0.03	0.29
Conditional Sympson-Hetter	0.93	-0.05	0.40
A-Stratified	0.95	-0.03	0.34
Enhanced A-Stratified	0.95	-0.04	0.34

Table 5

Pool Utilization and Exposure Rates for the Exposure Control Conditions

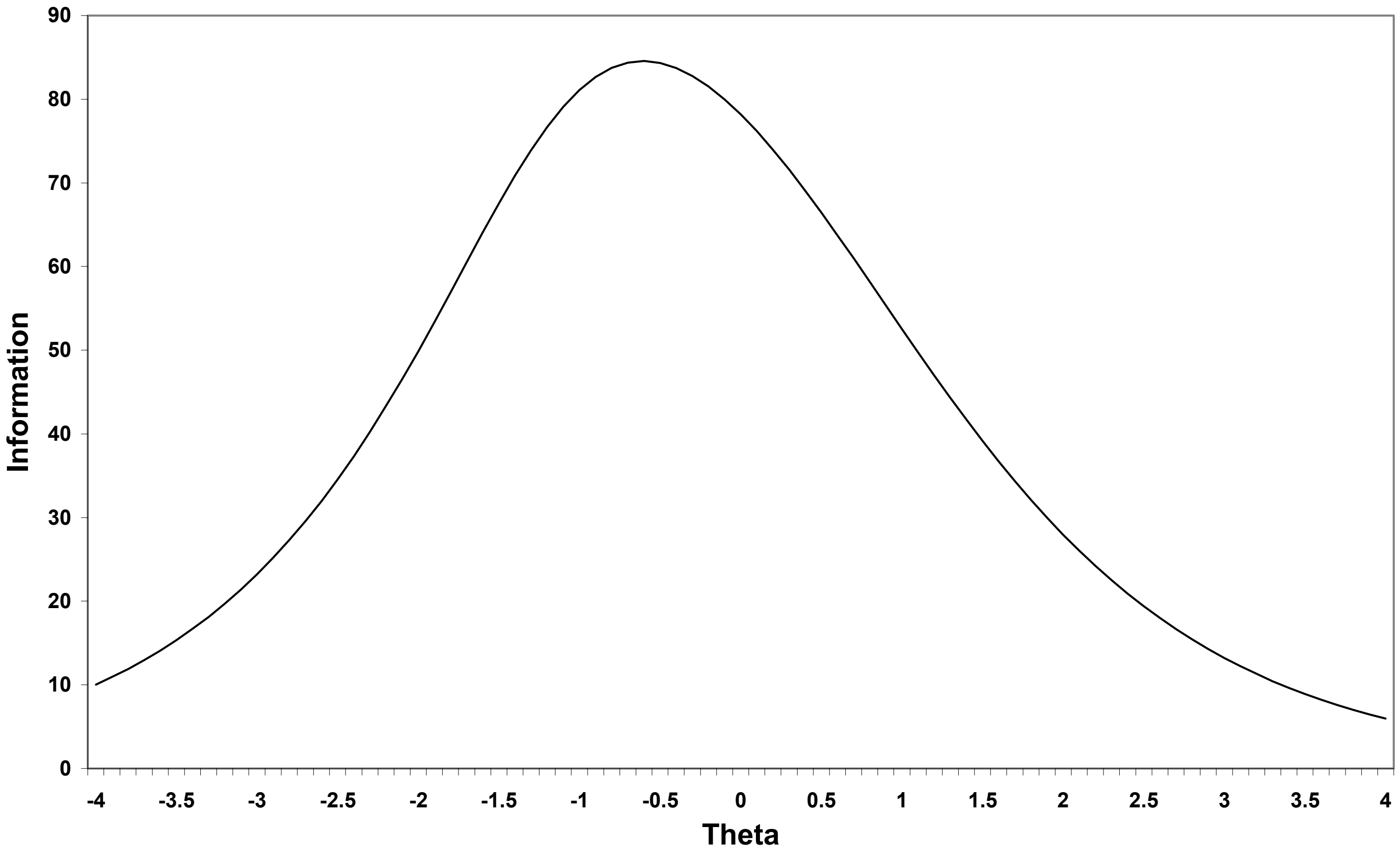
Exposure Control Conditions	***No-Exposure Control***	***Randomesque 3***	***Randomesque 6***	***Within .10 Logits-3***	***Within.10 Logits-6***	***Sympson Hetter***	***Conditional Sympson-Hetter***	***A-Stratified***	***Enhanced A-Stratified***
Exposure Rate									
1	0	0	0	0	0	0	0	0	0
.91-.99	0	0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

Table 6

Item Overlap Rates for the Exposure Control Conditions

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

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


Information
Theta