# *An Examination of Testlet Scoring and Item Exposure Constraints in the Verbal Reasoning Section of the MCAT*

Laurie Laughlin Davis
Barbara G. Dodd
University of Texas at Austin

This research was funded by the Graduate Student Research Program sponsored by the Association of American Medical Colleges—Medical College Admissions Test

## Abstract

The current study examined item exposure control procedures for testlet scoring of reading passages in the Verbal Reasoning section of the Medical College Admissions Test (MCAT) with four computerized adaptive testing (CAT) systems using the partial credit model. The first system was a traditional CAT using maximum information item selection. The second employed random item selection to provide a baseline for optimal exposure rates. The third implemented a variation of Lunz and Stahl's (1998) randomization procedure. The fourth examined Luecht & Nungester's (1998) computerized adaptive sequential testing (CAST) system. Content and item type balancing were controlled with the Kingsbury and Zara (1989) procedure in the first three conditions and were internally controlled through panel construction for the CAST condition. A series of simulated fixed length CATs were run to determine the optimal item selection procedure. Results indicated that both the randomization procedure and CAST performed well in terms of exposure control and measurement precision, with the CAST system providing the best overall solution when all variables were taken into consideration.

# An Examination of Testlet Scoring and Item Exposure Constraints in the Verbal Reasoning Section of the MCAT

Reading comprehension passages have long been a familiar item type on the landscape of verbal ability tests. These items usually consist of a reading passage followed by a series of items about the passage. However, these items have provided a special challenge for testing programs wishing to convert to item response theory (IRT) based computerized adaptive testing (CAT). This stems from the fact that groups of items which refer back to the same stimuli (in this case the reading passage) often violate the IRT assumption of local independence which states that an examinee's responses to a series of items should be statistically independent from one another. When this assumption is violated, ability estimates may be inaccurate due to overestimation of item information.

Wainer and Lewis (1990) discuss three ways of dealing with conditional dependencies among reading passage based items. First, the format may be altered so that only one item is asked per passage. While this resolves the issue, it is inefficient in terms of the small amount of information acquired for the large commitment of an examinee's time. The second method suggests that the dependencies simply be ignored and items within each passage scored individually in a dichotomous fashion. With such a solution, the amount of information obtained from each item may be overestimated, causing inaccuracy in ability estimation. Finally, the approach preferred by the authors was to score all items attached to a given passage as a single item or "testlet" using a polytomous IRT model. Thus, if four items were associated with a passage, the possible testlet score would range from 0 to 4, indicating the number of items answered correctly.

Another issue which plagues CAT administration is the threat to validity posed by over exposure of the item pool's most informative items. In an unconstrained CAT, maximum information item selection is used to optimize measurement precision by administering the most informative item for an examinee. However, this method of item selection often conflicts with the goal of maintaining test security through reducing overexposure of the most informative items. Alternatively, item usage is most uniform in the case that items are selected for administration randomly, however, measurement precision is sacrificed. Many different algorithms have been proposed to strike a balance between these two extremes.

***Controlling for item exposure***

Parshall, Davey, and Nering (1998) discuss the three often conflicting goals of item selection in CAT. First, item selection must maximize measurement precision, by selecting the item which maximizes information or posterior precision for the examinee's current ability 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. Different approaches to the goals of item selection will produce different testing algorithms (Stocking & Lewis, 2000). Attempts to address Parshall, Davey, and Nering's (1998) third goal are denoted exposure control methodologies.

Way (1998) discusses two types of exposure control strategies—randomization and conditional selection. Randomization strategies randomly choose the next item for administration from a set of nearly optimal items rather than selecting the single most informative item. Conditional selection strategies are those in which the probability of an item being administered is controlled conditional on a given criterion. Randomization procedures are usually considered to be simple to implement, but provide no guarantee that item exposure will

be constrained to a given level. Conditional strategies provide a guaranteed maximum exposure rate, but usually require complex and time-consuming simulations prior to operational use to derive exposure control parameters. In addition, these simulations may have to be rerun as test conditions change. A third approach to exposure control has recently been proposed by Chang and Ying (1996) in which items in the pool are stratified according to their statistical properties (item parameter estimates) and items are constrained to be administered from certain strata.

Randomization procedures were among the first proposed mechanisms for controlling item exposure. The 5-4-3-2-1 technique (McBride & Martin, 1983; Hetter & Sympson, 1997) was first proposed to alter the initial ordering of items in the CAT-ASVAB. In this procedure, the first item for administration is randomly chosen from the five most informative items, the second is chosen from the four most informative items, and so on until the fifth item when maximum information item selection resumes. This procedure focuses on controlling item exposure at the beginning of the test when examinees are most likely to see common items due to the use of a common initial ability estimate. Kingsbury and Zara (1989) proposed the randomesque method which randomly selects the next item to be administered from a group of the most informative items. Thus, the next item to be administered may be randomly chosen from the 5 or 10 most informative items for a given theta estimate. Thomasson (1998) examined the performance of a "choose one of three" randomization procedure whereby the next item for administration was chosen from the three most informative items and the two items not chosen were blocked from further administration.

Rather than choose from a fixed number of items, another possibility would be to choose from all items within a certain distance of the target difficulty value (Lunz & Stahl, 1998). Lunz and Stahl (1998) created a randomization procedure whereby all items within 0.10 logits of the

needed item difficulty are available for selection and the item to be administered is randomly chosen from among them. This procedure, which was designed to work with the Rasch model, does not utilize information in item selection, but rather matches the current ability estimate directly to the difficulty values of the items. The number of items which will appear in any selection grouping will depend on the distribution of item difficulties in the bank and within a given content area. If there are no items within 0.10 logits of the required item difficulty, the algorithm will randomly select the item having the closest difficulty to the target.

The best known conditional selection procedure is the Sympson-Hetter technique (Sympson & Hetter, 1985) which uses the frequency with which an item is administered in a simulated sample of examinees to determine the rate of exposure for each item. This iterative process results in each item being assigned an exposure control parameter ($K_i$) with a value between zero and one. These parameters are then used in live testing to constrain the probability of administering an item. The advantage of the Sympson-Hetter is that it allows a preset target exposure rate (r) to be reasonably ensured as long as the characteristics of the sample to which the test is given are distributed the same as those from which the parameters were derived. Many new conditional selection procedures have been built on the foundation of the Sympson-Hetter—extending and modifying it to fit different testing scenarios and to address limitations in the original procedure (Stocking & Lewis, 1995; Stocking & Lewis, 1998; Davey & Parshall, 1995).

Chang & Ying (1999) challenged the idea that maximum information item selection provides optimal measurement precision with their a-Stratified design which requires the item pool to be partitioned into different strata based on the value of the item discrimination parameter. Strata are then arranged in ascending order of discrimination. The test is divided into stages to match the number of strata such that a given number items are chosen from each

stratum, with the lowest discriminating strata used at the beginning of the test and the highest discriminating strata used at the end of the test. The a-Stratified design evens out exposure rates, because items with low and high discrimination values have an equal likelihood of being selected.

***Computerized Adaptive Sequential Testing***

An alternative to implementing an exposure control procedure to modify optimal item selection, is to control exposure apriori by preconstructing adaptive test forms. Computerized Adaptive Sequential Testing, or CAST, (Luecht & Nungester, 1998) involves the preconstruction of modules which contain groups of items and the arrangement of these modules into multistaged panels. Modules within each stage of a panel are segregated by item difficulty such that one module may contain easy items, one average items, and one hard items. As an examinee moves through the stages of testing, he or she is routed to a certain difficulty of module depending upon the current ability estimate. Modules and panels are constructed to meet certain statistical specifications such as desired test information, content coverage, and item exposure level. CAST incorporates the adaptive nature of CAT by allowing ability routing decisions between stages, while providing for quality control over a fixed number of adaptive test forms or "paths" through the panels. The framework for CAST harkens back to early adaptive testing procedures such as two stage testing (Lord, 1971b), pyramidal testing (Lord, 1974), and stradaptive testing (Weiss, 1974a) which used fixed branching through a structured item pool. CAST updates these methodologies by incorporating the IRT concept of information as well as the item selection and ability estimation procedures commonly used in CAT.

CAST allows great flexibility in assembling adaptive test forms in terms of the number of panels, number of stages, number of modules per stage, and number of items per module that can

be used. As such, CAST can be customized to meet the needs of each particular testing program. CAST panel assembly is usually undertaken using automated test assembly (ATA) software which allows the phrasing of statistical and nonstatistical goals in terms of mathematical functions to be minimized or maximized. Figure 1 presents several example panels which might be constructed with the CAST system. Typically, once a particular panel structure is decided upon, items will be assembled into modules to create multiple panels with the same structure. Luecht and Nungester (1998) discuss two strategies for panel assembly—bottom up and top down. With the bottom up strategy, items are assembled into modules such that each module as a self contained unit meets the requisite information, content, and item feature targets selected for the test. With this method, modules are interchangeable and can be mixed and matched to create multiple overlapping panels. The top down strategy requires only test level specifications of the statistical and nonstatistical targets. Modules are assembled in such a fashion that any path through the panel will result in a test of appropriate precision, content, and item type, however, modules are not exchangeable either within or across panels.

In operational testing, examinees are randomly assigned a panel to take. Testing begins with the module in the first stage. After all items in the first stage module have been completed, an examinee's ability is estimated using one of the typical CAT estimation procedures (MLE; Bayes Modal; EAP) and one of the modules in the second stage is selected for administration according to one of the typical CAT item selection procedures (maximum information; minimum posterior variance). Note that the module is the unit of test administration, so rather than item selection, CAST requires module selection and the module whose items as a grouped entity will provide the most information or the least posterior variance will be selected. The process of

ability estimation and module selection repeats for each stage of testing until the last stage of testing is completed.

***Exposure control research with polytomous IRT models***

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

While the research investigating the extent that measurement precision is affected when using such constraints in dichotomous item pools is extensive, only recently have researchers begun to address the effects exposure control when using polytomous items. The results of research on dichotomous items is not necessarily generalizable to the polytomous case because polytomously scored items yield a higher modal level of information across a larger span of the theta scale than dichotomously scored items (Koch & Dodd, 1989). Therefore, it is uncertain whether the negative impact on measurement precision observed in the dichotomous case of administering suboptimal items will be observed. In addition, polytomous item pools tend to be smaller than dichotomous item pools. The 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. Although, Koch and Dodd (1989) concluded that it was possible for a polytomous CAT to

perform well with item pools as small as 30 items, this finding did not take into consideration the threat posed to the test's validity if the item pool were to be compromised.

Pastor, Chiang, Dodd, and Yockey (1999), examined the performance of the Sympson-Hetter exposure control algorithm in fairly small (60,120) item pools using the partial credit model and concluded that it provided some protection against item exposure with minimal reduction in measurement precision. Davis, Pastor, Dodd, Chiang, and Fitzpatrick (2000) replicated these results with regard to measurement precision, but concluded that difficulties incurred in the implementation of the Sympson-Hetter mechanism, especially problems with convergence of the exposure control parameters in small item pools, were not outweighed by the observed gains in test security.

Pastor, Dodd, and Chang (in press) examined a broader range of exposure control mechanisms using the generalized partial credit model. The authors sought to evaluate the a-Stratified design (Chang & Ying, 1996) as a more simplistic alternative to the Sympson-Hetter. The study compared the a-Stratified design to both the Sympson-Hetter and the enhanced a-Stratified design (Leung, Chang, & Hau, 1999). In addition, exposure control conditional on ability was examined using both the conditional Sympson-Hetter (Stocking & Lewis, 1998; Parshall, Davey, & Nering, 1998) and the newly proposed conditional enhanced a-Stratified design, which incorporated the conditional Sympson-Hetter algorithm into the a-Stratified design. In contrast to the previous two studies, the results demonstrated a noticeable decrease in measurement precision as exposure control became more restrictive. As in previous research, convergence problems were observed when establishing exposure control parameters for the conditions incorporating the Sympson-Hetter or conditional Sympson-Hetter. The authors concluded that a more simplistic approach to exposure control such as the a-Stratified design

would be most appropriate with low or medium stakes tests or when the item pool to test length ratio was small. However, when high stakes testing necessitates tighter control of item exposure, the more restrictive conditional selection procedures should be considered.

The current research evaluates the utility of a modified Lunz and Stahl (1998) within .10 logits randomization procedure and the CAST framework for controlling item exposure in the context of testlet scoring with the partial credit model, using items and passages from the Verbal Reasoning section of the Medical College Admissions Test (MCAT). Zenisky, Hambleton, & Sirici (2000) explored the extent to which item dependencies occurred in the three passage based sections (Verbal Reasoning, Physical Sciences, and Biological Sciences) of the MCAT. Their findings indicated that item dependencies were most severe in the Verbal Reasoning section and that observed differences in reliability and ability estimates between conditions in which items were scored dichotomously (ignoring item dependencies) and polytomously (accounting for item dependencies) could produce meaningful differences in examinee scores.

***The partial credit model***

The partial credit model (Masters, 1982) is an extension of the one-parameter logistic (Rasch) model to the case where items may be scored polytomously as would be appropriate when partial credit is awarded for responses. The probability function of scoring in category $x$ on item $i$ given the examinee's ability, $\theta$, for the partial credit model is defined as;

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{k=0}^{x}(\theta - b_{ik})\right]}{\sum_{h=0}^{m_i}\exp\left[\sum_{k=0}^{h}(\theta - b_{ik})\right]}, \qquad (1)$$

where $m_i$ is the number of score categories minus one and $b_{ik}$ is the difficulty parameter associated with score category $x$.

## Method

### *Overview of techniques*

Four computerized adaptive testing systems were examined in this study. The first system was a traditional CAT using maximum information item selection. The second employed random item selection to provide a baseline for optimal exposure rates. The third implemented a variation of Lunz and Stahl's (1998) randomization procedure. The fourth examined Luecht & Nungester's (1998) computerized adaptive sequential testing (CAST) system. Content and item type balancing were controlled with the Kingsbury and Zara (1989) procedure in the first three conditions and were internally controlled through panel construction for the CAST condition. A series of simulated fixed length CATs were run to determine the optimal item selection procedure. Conditions were compared in terms of measurement precision, exposure rates, item overlap, and pool utilization.

### *Item pool*

Data were obtained from 22 forms of the MCAT collected during six separate administrations occurring from April 1996 through April 2001. Each form of the Verbal Reasoning section contained 55 multiple choice items grouped into eight passages for a total possible 176 passages. Passages contained either six, seven, eight, or ten multiple choice items each. Items were grouped together according to their passage identification number and the 0/1 scores for each item were summed to create passage testlet scores ranging from zero to the number of items per passage.

Inspection of the frequency distribution of the resulting passage testlet scores, indicated a problem with low category frequencies for certain passages. In other words, the number of individuals obtaining a given passage score was exceedingly small (in some cases zero). In order

to provide a reliable parameter estimate for each step value, the IRT calibration program requires a minimum number of observations for each score category. Previous experience has demonstrated that having ten observations per score category is sufficient to estimate the parameter. Twenty-five passages had to be dropped from the item pool due to low category frequencies. Another two passages were dropped due to convergence problems during calibration. The remaining 149 passages comprised the item pool which was used in all CAT conditions. A plot of test information for the item pool calibrated according to the partial credit model is presented in Figure 2. The item pool information exhibits a slight positive skew with information peaking at a theta value of –0.8.

Passages within the item pool were classified according to both content area and passage type as defined by the number of multiple choice items per passage. Of the 149 passages, 44% represented the content area of Humanities, 31% represented the content area of Social Science, and 25% represented the content area of Natural Science. Passage type was distributed as follows: 68% were six item passages, 20% were seven item passages, 7% were eight item passages, and 5% were ten item passages. While, MCAT also provides target percentages for item type by cognitive category (i.e. comprehension, evaluation, application, and incorporation) in it's test specifications for the Verbal Reasoning section, this value varied at the item level within a passage, making it impossible to include it in this study which treats the passage as the functional unit of measurement. A good distribution of cognitive categories within the items associated with a passage, however, may provide sufficient balance with respect to this variable.

***Parameter Estimation***

Passage scores were submitted to PARSCALE (Muraki & Bock, 1993) for calibration according to the partial credit model (Masters, 1982). Due to limitations of the MCAT equating

design, no attempt to equate the passages was made. Each form was calibrated separately and the resulting item parameters grouped together to create the CAT item pool, without adjustment to put them on the same scale. Since this was a simulation study and the goal of item calibration was simply to obtain realistic item parameters for the Verbal Reasoning section of the MCAT, the effects of this decision were judged to be minimal. However, this could not be advised for operational implementation. The number of observations per test form ranged from 2,510 to 14,949 depending upon the availability of data from alternate form orderings and multiple administrations within the obtained testing window.

PARSCALE employs a marginal maximum likelihood EM algorithm for parameter estimation that consists of two steps: first, the provisional expected frequency and sample size are calculated, and second, the marginal maximum likelihood is estimated. These steps continue through a series of iterations until item parameter estimates stabilize. For the current study, a convergence criterion of largest change of 0.001 or a maximum of 50 iterations was used (parameter estimates for all forms converged with fewer than 25 iterations). As noted above, two passages were discarded from the item pool due to convergence problems during calibration.

***Data generation***

The calibrated item parameters from PARSCALE were used as input to the data generation program. Item responses to the 149 passages were generated for 1000 simulees using conventional techniques. A random number was drawn from a normal distribution (0,1) to represent the known ability for each simulee. The probability of responding in each category given a simulee's ability was then computed for each item according to the partial credit model (Masters, 1982). These probabilities were then summed to create a cumulative probability of response ranging from 0 to 1. A random number was drawn from a uniform distribution and

compared to the cumulative response probability. The simulee was assigned the score which corresponded to the location in the cumulative response distribution that the random number fell at or below. This procedure was repeated for all simulees and all passages.

***CAT simulations***

A program originally developed by Chen, Hou, and Dodd (1998) was modified to meet the specifications of each CAT condition. Experimentation with the maximum information condition determined that a seven passage fixed test length would provide the best option for meeting measurement precision and non-statistical goals. 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 maximum likelihood estimation thereafter. Content and passage type were balanced for the three CAT conditions using the Kingsbury and Zara (1989) constrained CAT (CCAT) method. According to this method, after each item administration, the proportion of items given in each area is computed and compared to the target desired proportion. The next item administered is constrained to be chosen from the area with the largest discrepancy. In the current study, target proportions for both content and passage type were defined to match the observed percentages of each characteristic in the item pool. In the maximum information item selection condition, items were chosen to maximize the information at the current ability estimate, as constrained by content and passage type restrictions. In the random item selection condition, items were selected at random from the pool subject to content and passage type constraints.

As originally proposed, the Lunz and Stahl (1998) within .10 logits procedure randomly selects the item to be administered from among all items falling within 0.10 logits of the needed

item difficulty. Since, in the dichotomous case of the Rasch model, information peaks at the point on the ability scale where theta equals the item difficulty, selecting the next item according to item difficulty generally provides the same result as maximum information item selection and is computationally easier to implement. Polytomous items, however, do not have a single item difficulty value, but rather multiple step values needed to describe the probability of obtaining a score in a particular category. Therefore, modifications were needed to allow the use of the within .10 logits procedure in the polytomous case.

In the current study, the within .10 logits procedure was implemented by using maximum information item selection to select the two most informative passages at each of three points along the ability metric: estimated theta, estimated theta minus 0.10, and estimated theta plus 0.10. This resulted in a group of six passages from which one was randomly selected to be administered. The within .10 logits item selection procedure was conducted in combination with content and passage type balancing. For some combinations of content and passage type, there were fewer than six passages available in the pool. Specifically, this occurred for the eight and ten item passages. In these cases, the item to be administered was randomly selected from among all unadministered passages which met the content and passage type requirements. Implications of this decision for exposure control are presented in the results and discussion sections.

***CAST panel construction***

Drawing from the 149 calibrated passages in the CAT item pool, eight CAST panels were assembled using the top down method. Each panel contained three stages, with one module for the first stage and three modules each for the second and third stages. Passages were classified according to difficulty as either easy, average, or hard. Classifications were made based on an

examination of the characteristics of the step value parameters for each passage. Passages were assigned to modules and panels by hand, without the use of automated test assembly (ATA) software. The panel structure used as well as the distribution of passages within panels by difficulty, content, and passage type was chosen to work within the constraints of the available item pool. Had the characteristics of the pool been different or ATA software been available, other, more optimal, passage arrangements may have been determined. The first stage module contained three passages, one at each level of difficulty. The second and third stage modules were segregated by passage difficulty, with one module at each stage of easy, average, or hard passages. The second and third stage modules each contained two passages, yielding a total test length of seven passages. The structure of the panels used in this study is similar to that of the first diagram in Figure 1.

Panels were also constructed to meet content and passage type specifications. The first stage module contained one passage from each of the three content areas. The second stage modules each contained one passage each from Humanities and Natural Science. The third stage modules each contained one passage each from Humanities and Social Science. This provided each examinee with a test meeting the desired proportion of content coverage from each area. While the number of passages administered through the CAST design was consistent for all simulees, it was necessary to arrange passages within panels such that the total number of multiple choice items administered to each simulee would also be consistent. Passage type was, therefore, also considered in assigning passages to panels. The first stage module contained one passage each with eight and ten multiple choice items. This allowed the pools most informative passages to be administered in the first stage, thereby providing for more precision in the second stage routing decision. In addition, this also made the most efficient use of the small number of

eight and ten item passages in the pool. The third passage administered in the first stage contained either six or seven multiple choice items. This flexibility was necessary to meet the content and difficulty constraints with the available pool. The second stage modules contained either all six item passages or one six and one seven item passage depending upon the passage type of the third passage in the first stage. For example, for some panels, the first stage administers, a ten item, an eight item, and a seven item passage, with six item passages only in the second stage. However, other panels, give a ten item, an eight item, and a six item passage in the first stage with a six item and a seven item passage given in the second stage. All passages in the third stage contained six items. In spite of the variability in first and second stage passage types across panels, the resulting test length in terms of multiple choice items was the same for all panels.

The CAST structure chosen made use of 120 of the available 149 passages, leaving 29 passages unused by the procedure. Despite this excess of passages, meeting the target panel structure in terms of difficulty, content, and passage type was difficult with this pool. Unused passages, for the most part, represented those characteristics which the pool provided in abundance, leaving holes in the panel structure in certain areas. Information plots for each panel (shown in Figure 3) were judged to be similar enough to provide approximately equal measurement precision for all simulees regardless of which panel they were administered.

***CAST administration***

Simulations for the CAST condition, were also conducted by making modifications to the Chen, Hou, and Dodd (1998) program. Simulees were randomly assigned to take one of the eight panels. Once a panel had been assigned, examinees were administered the three passages in the first stage. Only after all three passages had been administered was ability estimated using

MLE, with a provision for using variable stepsize should category scores on the first three passages be identical. Simulees were then routed to one of three modules in the second stage containing either easy, average, or hard passages. This routing decision was made by adding together the information of passages within each module and selecting the module which provided the most information at the current ability estimate. After simulees completed both passages in the second stage, ability was again estimated, and simulees were routed to a module in the third stage. The third stage routing decision was made in the same way as the second stage routing decision with the exception that simulees could only be routed to a module of the same or adjacent difficulty. For example, simulees administered the easy module in the second stage, could be administered either the easy or the average module in the third stage, but not the hard module. This concurs with the examples given by Luecht and Nungester (1998) and prevents the possibility of any negative psychological impact which might be expected to occur from jumping from easy to hard items or vice versa.

***Data analyses***

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

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

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

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

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

$$AAD = \frac{\sum_{i=1}^{N} \left| \hat{\theta}k - \theta k \right|}{n}, \quad (5)$$

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

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

In order to measure test overlap, the audit trails of each simulee were compared to the audit trails of every other simulee. A data file containing the number of items shared among the simulees as well as the difference between their known theta values was created to obtain an

index of item overlap conditional on theta. Simulees were defined to have "similar" ability when their known thetas differed by two logits or fewer and "different" ability when their known thetas differed by more than two logits (Pastor, Chiang, Dodd, & Yockey, 1999; Davis, Pastor, Dodd, Chiang, & Fitzpatrick, 2000; Pastor, Dodd, & Chang, 2001).

## Results

### *Descriptive Statistics*

After all conditions had been run, a listwise deletion of 166 nonconvergent cases was performed. A case was defined as nonconvergent if, once the end of the test had been reached, the ability 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. The number of nonconvergent cases for each condition is listed in Table 1. As expected, the random item selection condition produced the most nonconvergent cases with 115. The CAST condition had no nonconvergent cases. The within .10 logits procedure and maximum information item selection procedure were in between these extremes, with within .10 logits producing slightly more nonconvergent cases (44) than the maximum information item selection procedure (27). The remaining results are reported on the sample (N=834) of observations which remained after the nonconvergent cases had been deleted.

Table 1 also contains the average theta estimate and standard error for each condition. The mean of the known thetas was 0.09 with a standard deviation of 1.03. The maximum information condition produced results which were reasonably close to these values. The random item selection condition yielded an average theta estimate lower than expected with a higher standard deviation than for the known values. The within .10 logits and CAST conditions both resulted in average theta estimates and standard deviations higher than for the known

values. As expected, maximum information item selection yielded the lowest standard error (0.29) with random item selection yielding the highest standard error (0.36). The within .10 logits and CAST conditions both yielded the same standard error (0.33) which fell between these two extremes.

Table 2 presents the correlations between known and estimated theta for each condition as well as statistics for bias, SDM, RMSE, SRMSD, and AAD. Random item selection produced the lowest correlation between known and estimated theta (0.93). The remaining three conditions all yielded higher and similar values (0.95 to 0.96). Bias was zero for maximum information item selection, and very slightly negative for the other three conditions (-0.02 to –0.04). Results for SDM mirrored those for bias. The RMSE, SRMSD, and AAD statistics all revealed the same pattern of results across conditions. Values were lowest for maximum information item selection and highest for random item selection. CAST resulted in a slightly higher values than did within .10 logits.

***Pool utilization and exposure rates***

Table 3 contains the frequency of observed exposure rates along with the average, maximum, and standard deviation of exposure rates, and the percent of pool not administered for each condition. Table 3 is partitioned into two sections. The top portion represents the results when the exposure rates for the entire item pool are considered. The bottom portion represents the results when the eight and ten item passages have been removed from the item pool. This dual presentation is necessary to illustrate the impact of content and passage type constraints on exposure rates. So few of the eight and ten item passages were available in the item pool that in order to fulfill the nonstatistical constraints of the test specifications, these passages were forced to be overexposed.

Chen, Ankenmann, & Spray (1999) state that the average exposure rate for any fixed length test will always be constant and mathematically equal to the ratio of test length to pool size. Since test length was the same for all conditions studied, the observed average exposure rates, therefore, did not differ across conditions with the same size item pools. The only differences in average exposure rate occurred for the CAST condition, when fewer passages were in the item pool, forcing average exposure rates to increase slightly. The standard deviation of exposure rates was highest for maximum information item selection and second highest for the within .10 logits procedure in both the top and bottom portions of the table. The relative position of random item selection and the CAST procedure flipped from the top portion to the bottom portion of the table. When all passages were included, the CAST procedure yielded a slightly lower standard deviation of exposure rates than did random item selection, however, when eight and ten item passages were removed from the analyses random item selection produced the lowest standard deviation of exposure rates. This same pattern occurred for maximum exposure rates with maximum information yielding the highest maximums (.513 entire pool; .474 reduced pool), within .10 logits the second highest maximums (.444 entire pool; .191 reduced pool), and the relative position of CAST (.165 entire pool; .165 reduced pool) and random item selection (.428 entire pool; .104 reduced pool) reversing when eight and ten item passages were removed.

The percent of pool not administered was highest for maximum information item selection (62% entire pool; 66% reduced pool) and second highest, though greatly reduced, for the within .10 logits procedure (18% entire pool; 21% reduced pool). Random item selection administered all items in the pool and the CAST procedure administered all items from the reduced (N=120) pool.

### *Item Overlap*

Audit trails for each simulee were compared to the audit trails of every other simulee resulting in 347,361 pairwise comparisons per conditions. Table 4 contains the average item overlap for all simulees, those of different abilities (known thetas differed by more than two logits), and those of similar abilities (known thetas differed by two logits or fewer) for each condition. Table 4 is also partitioned into two sections, with the top portion representing the results for the entire item pool and the bottom portion representing the results when the eight and ten item passages have been removed from the item pool. Maximum information item selection produces the highest overall overlap rates with an average of 26% overlap in the entire pool and 17% overlap in the reduced pool. While, the relative position of random item selection and the CAST procedure alternate in the top and bottom portions of the table, CAST consistently results in lower overlap rates than the within .10 logits procedure (9% and 5% for CAST vs. 13% and 6% for within .10 logits). Results for examinees of similar ability demonstrate the same pattern with highest overlap rates occurring with maximum information item selection (30% overlap entire pool; 20% overlap reduced pool) and overlap rates for CAST (10% overlap entire pool; 6% overlap reduced pool) lower than overlap rates for within .10 logits (14% overlap entire pool; 7% overlap reduced pool). Results for examinees of different abilities produce a different pattern, but the overlap rates are uniformly so small (6-10% for entire pool; 2-4% for reduced pool) that comparison of conditions is not warranted.

### *CAST implementation*

Table 5 contains the frequency with which different paths through the panels were taken. As can be seen by the row totals, distribution of simulees across each of the eight panels was relatively even, with any variability attributable to the removal of nonconvergent cases from the

sample. However, distribution of simulees across paths within the panels was skewed toward the extremes with more than 50% of the simulees being routed to the hard modules and almost 20% of the simulees being routed to the easy modules in both the second and third stages. Figures 4 and 5 present plots of the module level information at the second and third stages for panels 5 and 8. As can be seen in the figures, the average difficulty modules were only the most informative over a very limited range of the theta scale, and, in certain cases (such as stage 2 of panels 5 and 8), the average difficulty modules never provided the most information. The information plots presented here are typical of the plots observed for the remainder of the panels. Therefore, the bias seen in Table 5 toward the extremes can be explained by the relatively small amounts of information provided by the average difficulty modules compared to the easy and hard difficulty modules.

## Discussion

### *Item Pool*

While having an item pool of sufficient size to estimate ability and minimize item exposure is important, Stocking and Lewis (2000) emphasize the need for the available item pool to adequately reflect test specifications for content, item type, and other nonstatistical properties. Their research has demonstrated that item pools which do not have a sufficient number of items to match these specifications make the use of conditional exposure control strategies such as the Sympson-Hetter difficult, if not impossible, to implement because of problems in obtaining convergence of the exposure control parameters. Since practical and economic issues often make expansion of an item pool infeasible, alternatives to the conditional exposure control strategies must be sought. The procedures examined in this study provide two reasonable alternatives for controlling exposure with a less than optimal item pool. The result of doing so,

however, is evidenced by the higher exposure rates of the passage types in short supply, as seen in the top portions of Tables 3 and 4.

In the current study, test specifications were set equal to the proportion of item type occurring in the calibrated pool in an attempt to avoid the mismatched situation described above. However, so few passages were given in each test that the rounding of the target proportions to whole passage units, forced a discrepancy between target proportions and observed proportions, and consequently a mismatch with the available pool. For example, the target number of ten item passages to be administered in an examinee's test was 5%. Since the test length was 7 passages, this translated to roughly one-third of a passage. Operationally, this was rounded to one whole passage, or 14% of the test. Because the item pool contained only 5% ten item passages, those 5% were forced to be overexposed. It is, therefore, recommended for future research and operational use that test length be considered when determining test specifications and item pool makeup, such that the target proportion of item type to be administered and the available percentage of that item type in the pool result in whole number units. Given the current test specifications, it is recommended that the item pool be supplemented with additional eight and ten item passages.

Calibration problems with low category frequencies, forced the removal of many passages from the item pool. However, these problems occurred disproportionately with the eight and ten item passages. Of the 27 passages discarded from the item pool, 22 were eight or ten item passages. The presence of additional categories clearly requires an increase in the sample size required for calibration, however, this alone cannot account for the problem, as the sample sizes for most forms were judged to be sufficient. An alternative explanation may stem from the testlet scoring procedures themselves. Unlike true polytomously scored items in which

a single item is assigned a score along a continuum indicating quality or number of steps successfully completed, testlet scores are derived by summing the correct or incorrect response strings to a series of related items. During test administration, each item is completed independently. The probability of answering all items incorrectly, thus receiving a score of zero, is quite small due to the accumulated effects of guessing across the set of items. The more items that are included in a set, the more likely that an examinee will answer at least one item correctly. Since most of the observed low category frequencies occurred for the 0 and 1 score categories, this explanation seems likely. One possible solution to this dilemma may be found in the work of Wilson and Masters (1993) in which a method for calibrating polytomous items with low category frequencies is discussed.

***Random item selection***

Random selection of items for administration was presented in this study only as a means of providing a baseline comparison to gauge optimal exposure rates. In other words, this method of item selection would not be recommended for operational implementation. However, the results from this condition do highlight an interesting phenomenon. Even with items being chosen for administration completely at random, with no reference whatsoever to ability, the correlation between known and estimated theta only dropped to 0.93 and the standard error was only 0.36. The condition provided the worst results in terms of measurement precision among the four studied conditions, but the results were not that far below those of the optimal condition. Two possible explanations are posited. One possibility might be that these results stem from the compensatory nature of polytomously scored items. These items yield a higher modal level of information across a larger span of the theta scale than dichotomously scored items, making negligible the impact on measurement precision of substituting of one item for another. Another

possibility is that results may differ for Rasch and non-Rasch models due to the influence, or lack thereof, of item discrimination on information, and thus, maximum information item selection. Way (1998) discusses the differential impact of the underlying measurement model on CAT performance, citing research that demonstrates that Rasch based CATs for dichotomously scored items tend to be robust to modifications in the item pool and item selection algorithms which may cause the administration of sub-optimal items (Haynie & Way, 1994; Way, Zara, & Leahy, 1996). In a Rasch model, when all items are assumed to be of equal discrimination, the substitution of one item for another may make less of an impact on ability estimation than in the case where items vary in terms of discrimination because they are all equally informative, only differing in their location along the difficulty scale. Further research is necessary to delineate between these possible explanations.

***Within .10 logits and CAST***

Bergstrom and Lunz (1999) demonstrated the utility of the within .10 logits procedure with dichotomously scored items from the ASCP CAT with a 900 item bank, concluding that the maximum exposure rate was less than 30% for most items, with only a few items near the pass point and in short supply content areas being administered with higher frequency. Results of the current study reflect the same pattern, with exposure rates being controlled to below 20% except for a couple of items where overexposure was necessary in order to meet nonstatistical constraints. These findings suggest that while maximum exposure rates still cannot be guaranteed as with a conditional selection strategy, with careful development and supplementation of the item pool exposure rates can be held to acceptable levels. The procedure is simple to implement, controls exposure relatively well, and can be used with less than optimal item pools.

The performance of the CAST procedure was better than anticipated, given that panels were constructed without the aid of automated test assembly software. While overall information was comparable across panels, module level information functions point to the nonuniform distribution of information across difficulty levels within stages which caused a bias towards the extremes in module selection. The fact that ability was recovered as accurately as it was, with no nonconvergent cases, and that item exposure and overlap rates remained low even in the face of this bias, is testimony to the robust nature of the CAST system. It is anticipated that results for the CAST method would even improve with a larger, more targeted item pool and the use of ATA software.

In comparing the two viable solutions for controlling exposure presented in the current study, it should first be stated that both methods performed well and provided a good measure of control over exposure and overlap rates with an acceptably low decrease in measurement precision. Either procedure can be recommended for use with a similar item pool size and test structure to the one used here. However, there were differences in how the procedures performed, and while these differences were not completely consistent in pointing to a superior procedure across dependent measures, the weight of evidence suggests that the CAST system may provide the best overall solution. CAST outperformed the within .10 logits procedure in terms of exposure and overlap rates and pool utilization. Differences in the maximum exposure rates between the two procedures were most exaggerated when the entire pool was considered for analysis. While this difference was largely mitigated by the exclusion of the eight and ten item passages, CAST remained superior. The same pattern emerged when the two procedures were compared on overlap rates, with CAST outperforming within .10 logits by 4% when the entire pool was considered and by a smaller 1% when the eight and ten item passages were

removed. The most stunning result, however, in terms of test security was in regard to pool utilization. While, the within .10 logits procedure did reduce the percent of pool not administered by 44-45% over maximum information item selection, the CAST procedure consistently administered all available passages.

In terms of the descriptive statistics, results were mixed with the CAST system providing superior values for bias and SDM and the within .10 logits procedure yielding better performance for RMSE, SRMSD, and AAD statistics. The correlation between known and estimated theta was marginally higher for within .10 logits than for CAST, but this difference is too small to be of practical significance. Estimated theta values for CAST were slightly closer to those of the known thetas, but again the difference was small. Results for the two procedures in terms of standard error were identical. The most notable result, however, was that the CAST system produced no nonconvergent cases, whereas the within .10 logits procedure produced 44.

CAST yielded overall superior performance for test security, especially in terms of pool utilization. While the within .10 logits procedure did have identical or superior values for some descriptive measures, differences in favor of the procedure were small relative to the CAST statistics. Finally, CAST demonstrated it's superiority in it's ability to estimate theta values for all 1000 simulees in the original sample.

***Conclusions***

Testlet scoring of passage based items provides a clear advantage over dichotomous scoring when conditional item dependencies are found. However, this advantage does not come without the cost of the added complexity of using polytomous IRT models and calibration difficulties due to low category frequencies when large numbers of items are associated with a passage. It is, therefore, recommended that testlet scoring only be considered when strong

conditional dependencies have been found among the passage based items, as was the case with the Verbal Reasoning section of the MCAT.

Two methods for controlling item exposure when testlet scoring is used were proposed in the current research. Both the within .10 logits and CAST procedures performed well in terms of test security and measurement precision and are both certainly preferable to the no exposure control alternative. When all variables are considered, however, the CAST system appears to be the more flexible and robust option. In addition to providing, superior results in terms of test security, CAST has the advantage of apriori construction of test forms, enabling test developers to execute a higher level of quality control of the measurement and content related properties of each test form. While one particular CAST structure was examined in the current study, many other structures are possible. Further research should examine the impact of CAST structure on measurement and test security variables.

## References

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

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., 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.

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.

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

Davis, L.L., 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.

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

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

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

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

Lord, F.M. (1971b). A theoretical study of two-stage testing. Psychometrika, 36, 227-241.

Lord, F.M. (1974). Individualized testing and item characteristic curve theory. In Krantz, Atkinson, Luce, & Suppes, Contemporary developments in mathematical psychology, 2 (pp. 106-126). San Francisco, CA: W.H. Freeman.

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.

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.

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.

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.

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

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

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

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

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

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.

Thomasson, G.L. (1998). CAT Item exposure control: New evaluation tools, alternate methods and integration into a total CAT program. Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.

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

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

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

Weiss, D.J. (1974a). Strategies of adaptive ability measurement. (RR 74-5Z). Minneapolis, MN: Psychometric Methods Program, Department of Psychology, University of Minnesota. (NTIS No. AD-A004 270).

Wilson, M., & Masters, G. (1993). The partial credit model and null categories. Psychometrika, 58, 87-99.

Zenisky, A.L., Hambelton, R.K., & Sireci, S.G. (2000, April). Effects of local item dependence on the validity of IRT item, test, and ability statistics. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

TABLE 1

*Means (and Standard Deviations) for Estimated Theta and Standard Error (N=834)*

| Passage Selection Condition | Theta* Estimate | Standard Error |
|---|---|---|
| Maximum Information | 0.09 | 0.29 |
| | (1.09) | (0.06) |
| Random | 0.01 | 0.36 |
| | (1.14) | (0.12) |
| Within .10 Logits | 0.13 | 0.33 |
| | (1.11) | (0.08) |
| CAST | 0.11 | 0.33 |
| | (1.11) | (0.08) |

*Note: Mean and SD for Known Thetas were Mean=0.09; SD=1.03

TABLE 2

*Correlation Coefficients Between Known and Estimated Theta, Bias, Standardized Difference Between Means (SDM), Root Mean Squared Error (RMSE), Standardized Root Mean Squared Difference (SRMSD), and Average Absolute Difference (AAD) Statistics for Each of the Four Passage Selection Conditions (N=834)*

| Passage Selection Condition | Correlation | Bias | SDM | RMSE | SRMSD | AAD |
|---|---|---|---|---|---|---|
| Maximum Information | 0.96 | 0.00 | 0.00 | 0.31 | 0.52 | 0.24 |
| Random | 0.93 | -0.04 | 0.04 | 0.41 | 0.59 | 0.30 |
| Within .10 Logits | 0.96 | -0.04 | 0.04 | 0.33 | 0.54 | 0.25 |
| CAST | 0.95 | -0.02 | 0.02 | 0.35 | 0.56 | 0.27 |

TABLE 3

*Pool Utilization and Exposure Rates for the Four Passage Selection Conditions*

| **All Passages Included** | | | | |
|---|---|---|---|---|
| | Maximum Information | Random | Within .10 Logits | CAST* |
| Number of Passages | 149 | 149 | 149 | 120 |
| Exposure Rate | | | | |
| 1 | 0 | 0 | 0 | 0 |
| .91-.99 | 0 | 0 | 0 | 0 |
| .81-.90 | 0 | 0 | 0 | 0 |
| .71-.80 | 0 | 0 | 0 | 0 |
| .61-.70 | 0 | 0 | 0 | 0 |
| .51-.60 | 1 | 0 | 0 | 0 |
| .41-.50 | 3 | 2 | 1 | 0 |
| .36-.40 | 2 | 0 | 1 | 0 |
| .31-.35 | 0 | 0 | 0 | 0 |
| .26-.30 | 3 | 0 | 0 | 0 |
| .21-.25 | 1 | 0 | 0 | 0 |
| .16-.20 | 5 | 1 | 7 | 3 |
| .11-.15 | 6 | 6 | 11 | 15 |
| .06-.10 | 14 | 21 | 25 | 34 |
| .01-.05 | 21 | 119 | 77 | 68 |
| Not Administered | 93 | 0 | 27 | 0 |
| Exposure Rate AVG | 0.047 | 0.047 | 0.047 | 0.058 |
| Exposure Rate SD | 0.101 | 0.052 | 0.063 | 0.044 |
| Exposure Rate MAX | 0.513 | 0.428 | 0.444 | 0.165 |
| % of Pool Not Administered | 62% | 0% | 18% | 0% |
| **8 and 10 Item Passages Removed From Analyses** | | | | |
| | Maximum Information | Random | Within .10 Logits | CAST* |
| Number of Passages | 131 | 131 | 131 | 104 |
| Exposure Rate | | | | |
| 1 | 0 | 0 | 0 | 0 |
| .91-.99 | 0 | 0 | 0 | 0 |
| .81-.90 | 0 | 0 | 0 | 0 |
| .71-.80 | 0 | 0 | 0 | 0 |
| .61-.70 | 0 | 0 | 0 | 0 |
| .51-.60 | 0 | 0 | 0 | 0 |
| .41-.50 | 2 | 0 | 0 | 0 |
| .36-.40 | 2 | 0 | 0 | 0 |
| .31-.35 | 0 | 0 | 0 | 0 |
| .26-.30 | 2 | 0 | 0 | 0 |
| .21-.25 | 1 | 0 | 0 | 0 |
| .16-.20 | 5 | 0 | 5 | 1 |
| .11-.15 | 3 | 1 | 7 | 5 |
| .06-.10 | 10 | 19 | 23 | 30 |
| .01-.05 | 20 | 111 | 69 | 69 |
| Not Administered | 86 | 0 | 27 | 0 |
| Exposure Rate AVG | 0.038 | 0.038 | 0.038 | 0.048 |
| Exposure Rate SD | 0.088 | 0.019 | 0.042 | 0.036 |
| Exposure Rate MAX | 0.474 | 0.104 | 0.191 | 0.165 |
| % of Pool Not Administered | 66% | 0% | 21% | 0% |

*Note: CAST panel construction did not use the entire available item bank

TABLE 4

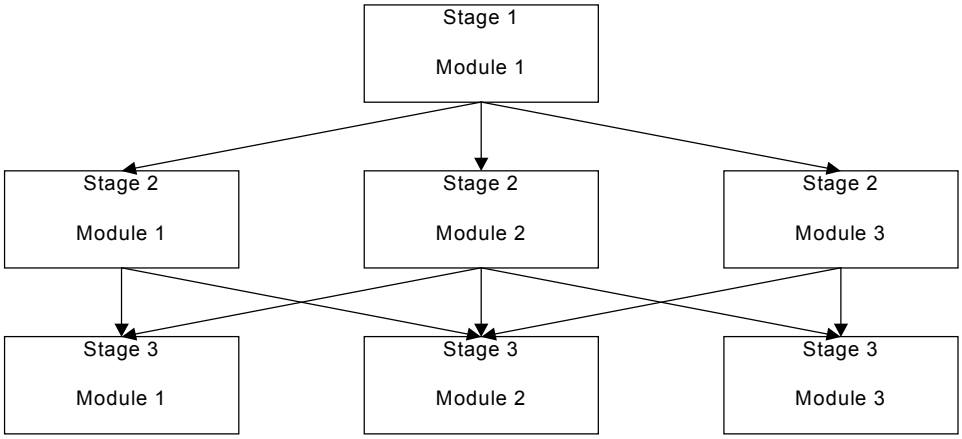
*Item Overlap for the Four Passage Selection Conditions*

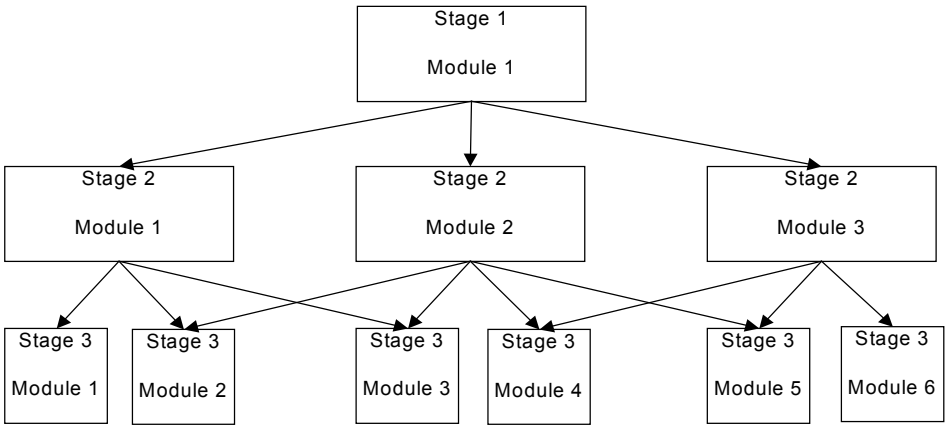
| **All Passages Included** | | | |
|---|---|---|---|
| | Overall Average Overlap (N=347,361) | Different Abilities Average Overlap (N=59,471) | Similar Abilities Average Overlap (N=287,890) |
| ***Maximum Information*** | 1.84 | 0.49 | 2.12 |
| (149 Paragraphs) | 26% | 7% | 30% |
| ***Random*** | 0.72 | 0.71 | 0.72 |
| (149 Paragraphs) | 10% | 10% | 10% |
| ***Within .10 Logits*** | 0.90 | 0.63 | 0.96 |
| (149 Paragraphs) | 13% | 9% | 14% |
| ***CAST*** | 0.63 | 0.45 | 0.67 |
| (120 Paragraphs) | 9% | 6% | 10% |
| **8 and 10 Passages Items Removed from Analyses** | | | |
| | Overall Average Overlap (N=347,361) | Different Abilities Average Overlap (N=59,471) | Similar Abilities Average Overlap (N=287,890) |
| ***Maximum Information*** | 1.20 | 0.26 | 1.40 |
| (131 Paragraphs) | 17% | 4% | 20% |
| ***Random*** | 0.23 | 0.23 | 0.23 |
| (131 Paragraphs) | 3% | 3% | 3% |
| ***Within .10 Logits*** | 0.42 | 0.15 | 0.47 |
| (131 Paragraphs) | 6% | 2% | 7% |
| ***CAST*** | 0.36 | 0.18 | 0.40 |
| (104 Paragraphs) | 5% | 3% | 6% |

TABLE 5
*Frequency of CAST Panel and Module Usage*
*(N=834)*

| | **Stage2 Difficulty** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Easy | Easy | Average | Average | Average | Hard | Hard | |
| | **Stage3 Difficulty** | | | | | | | |
| **Panel** | Easy | Average | Easy | Average | Hard | Average | Hard | Totals |
| 1 | 13 | 10 | 1 | 12 | 0 | 8 | 61 | 105 |
| 2 | 23 | 0 | 9 | 0 | 16 | 7 | 60 | 115 |
| 3 | 19 | 2 | 7 | 15 | 4 | 3 | 60 | 110 |
| 4 | 21 | 6 | 1 | 16 | 19 | 10 | 36 | 109 |
| 5 | 19 | 10 | 0 | 0 | 2 | 3 | 45 | 79 |
| 6 | 17 | 4 | 1 | 14 | 1 | 16 | 54 | 107 |
| 7 | 33 | 4 | 1 | 1 | 1 | 4 | 57 | 101 |
| 8 | 16 | 2 | 0 | 1 | 19 | 21 | 49 | 108 |
| Totals | 161 | 38 | 20 | 59 | 62 | 72 | 422 | 834 |

FIGURE 1
*Three Possible Panel Structures Using CAST*



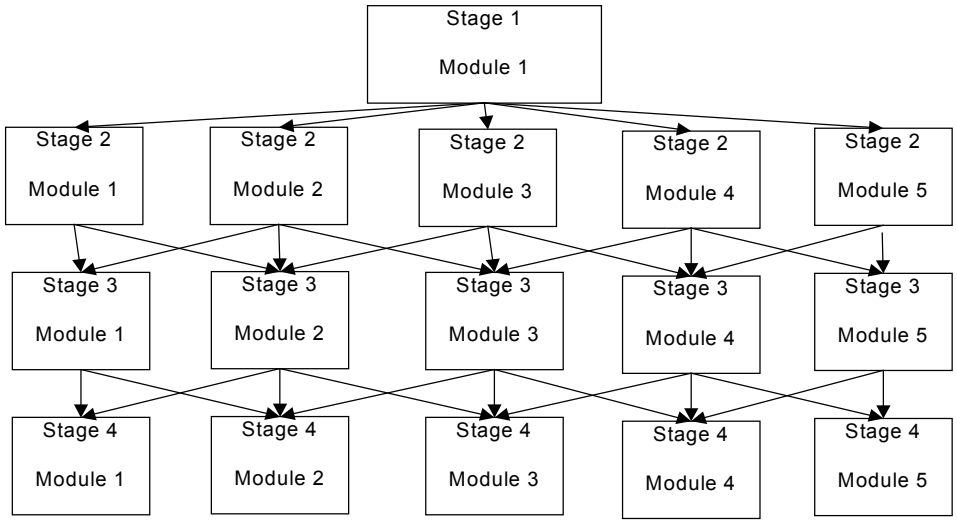



FIGURE 2

*Test Information Function for the MCAT Verbal Reasoning Item Pool for 149 Passages Calibrated According to the Partial Credit Model*

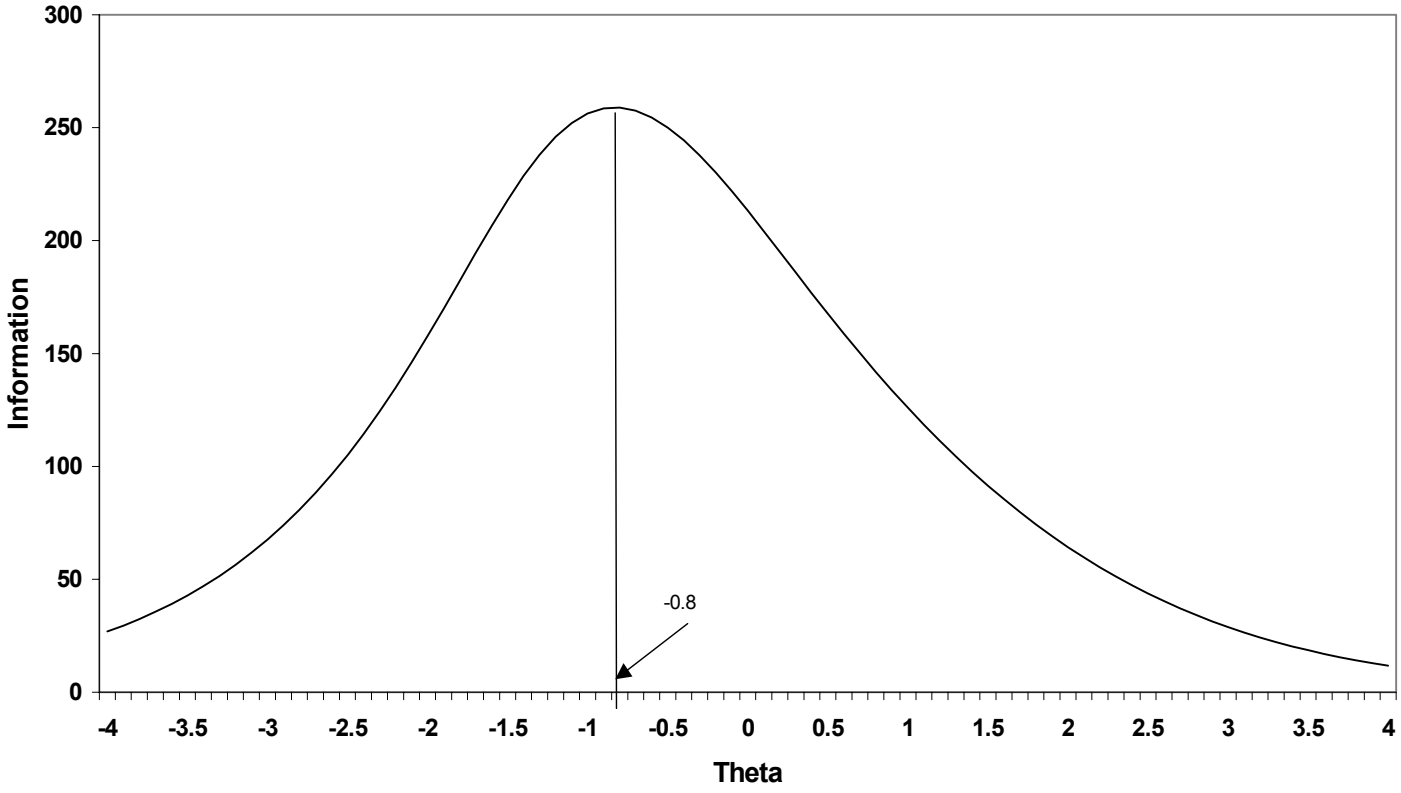


FIGURE 3
*Panel Information Functions*

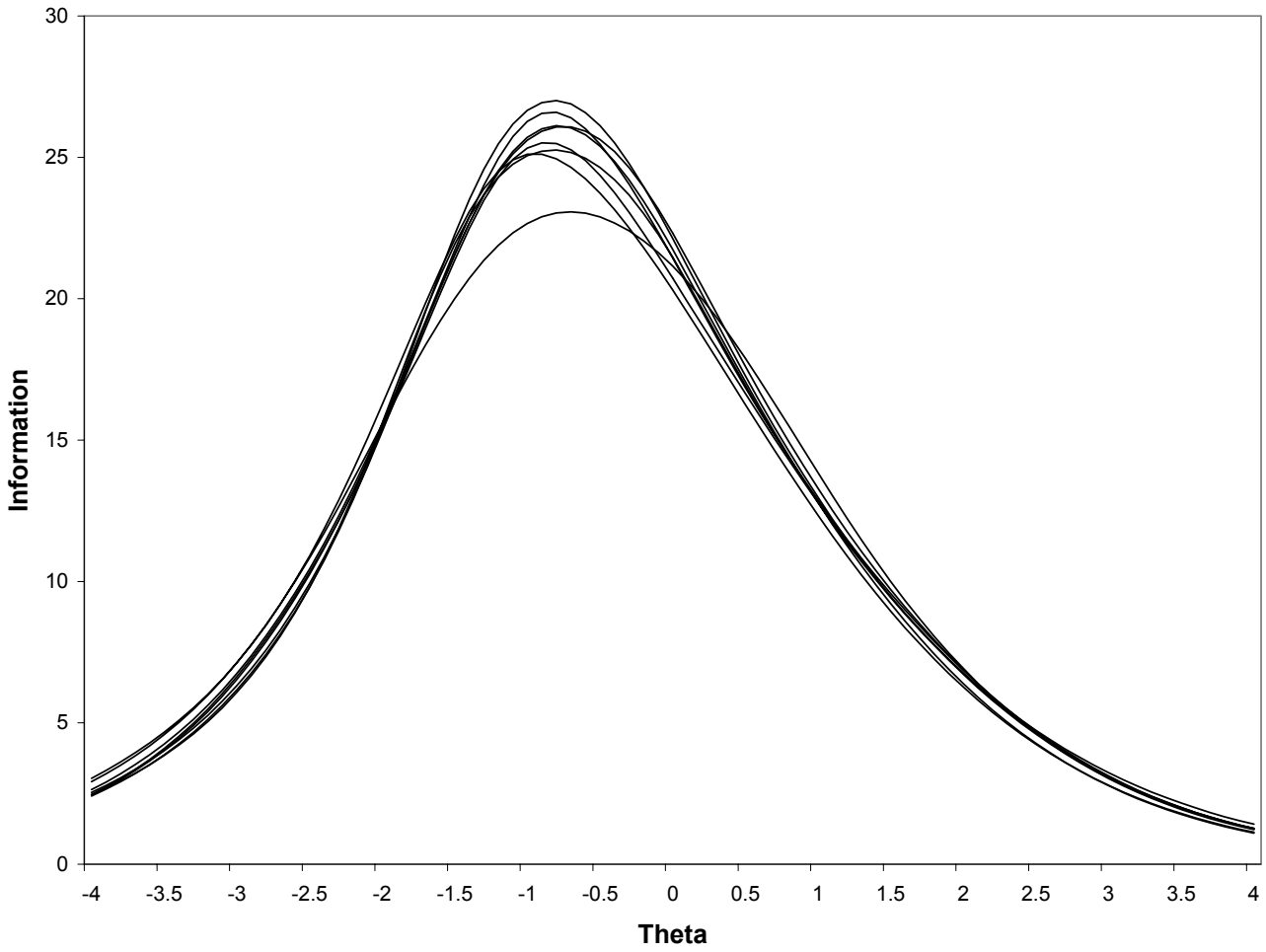
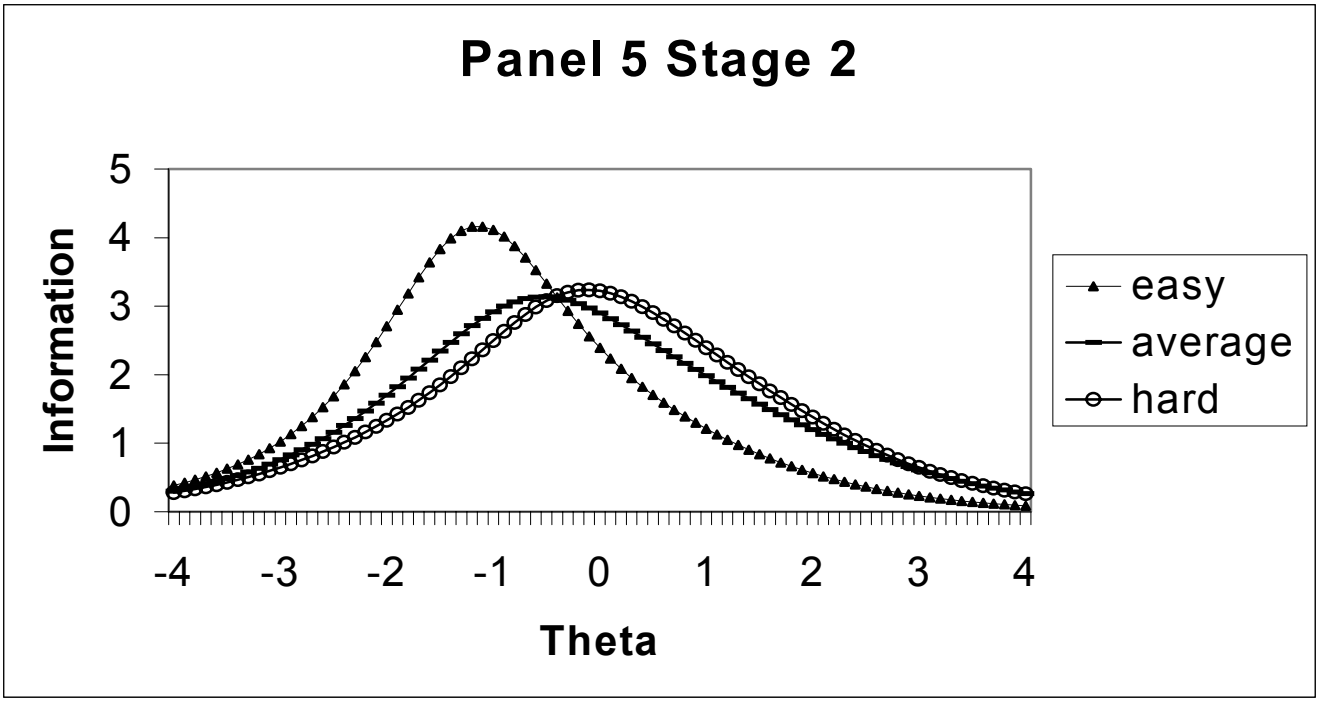


FIGURE 4
*Module Level Information Plots for Panel 5*



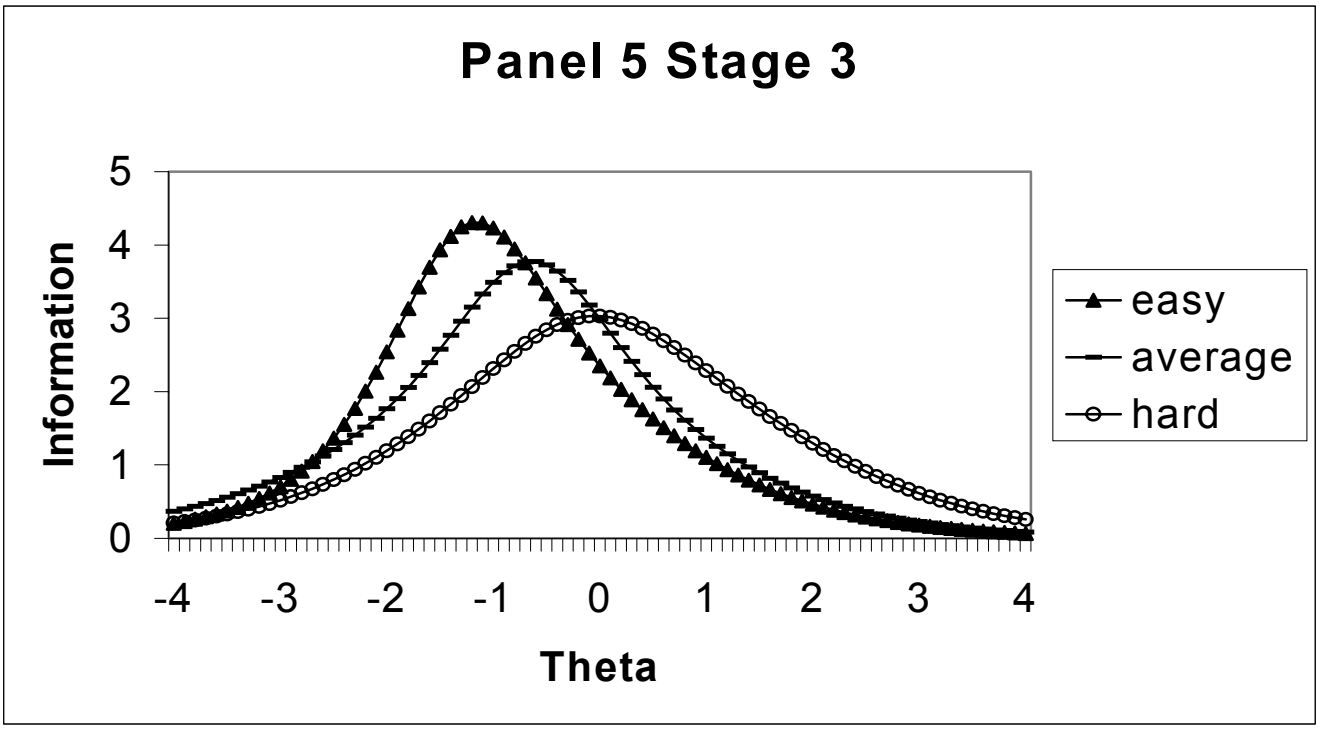



FIGURE 5
*Module Level Information Plots for Panel 8*