Running Head: A COMPARISON OF EXPOSURE CONTROL PROCEDURES IN CAT

A Comparison of Exposure Control Procedures in CAT Systems Based on Different Measurement Models for Testlets Using the Verbal Reasoning Section of the MCAT[1]

Aimee M. Boyd and Barbara G. Dodd

University of Texas at Austin

and

Steven J. Fitzpatrick

Pearson Educational Measurement

[1]This research was supported by the 2002 Graduate Student Research Program sponsored by the Association of American Medical Colleges.

A Comparison of Exposure Control Procedures in CAT Systems Based on Different Measurement Models for Testlets Using the Verbal Reasoning Section of the MCAT

Abstract

This study compared several item exposure control procedures for CAT systems based on a three-parameter logistic testlet response theory model (Wang, Bradlow, & Wainer, 2002) and Masters' (1982) partial credit model using real data from the Verbal Reasoning section of the MCAT. The exposure control procedures studied were the modified within .10 logits procedure (Davis & Dodd, 2001), the Sympson-Hetter procedure (Sympson & Hetter, 1985) with a maximum exposure rate restricted to 0.19, the Sympson-Hetter procedure (Sympson & Hetter, 1985) with a maximum exposure rate restricted to 0.29, and maximum information, a no exposure control condition, used as a baseline. The exposure control procedures were evaluated for measurement precision, utilization of the item pool, and item overlap across test administrations. For both measurement models, the modified within .10 logits procedure provided better pool utilization with little decrement in precision of measurement than either of the Sympson-Hetter procedures.

Keywords: computerized adaptive testing, exposure control, testlets, polytomous models, item exposure rate.

A Comparison of Exposure Control Procedures in CAT Systems Based on Different Measurement Models for Testlets Using the Verbal Reasoning Section of the MCAT

As test developers transform well established, reliable paper and pencil tests to computer adaptive testing (CAT) formats, various benefits are gained, including enhanced measurement precision, better test security, and shorter test lengths due to administration of more informative items (Wainer, 2000). In order to take advantage of these benefits, the psychometric properties of the test are based on item response theory (IRT), rather than traditional true score theory (Crocker & Algina, 1986). CAT tailors a test for each individual examinee by taking into account the examinee’s responses to previous items and selecting additional items that will most accurately discern and measure the examinee’s ability level.

Multiple-choice items are the most frequently used item format in CATs to date. This is due to the relative ease of developing and scoring multiple-choice items compared to other item formats (Haladyna, 1997). In addition, multiple-choice items tend to meet the assumptions of IRT, such as local independence, unidimensional latent trait, and non-speeded test administrations (Hambleton & Swaminathan, 1985). However, a set of multiple-choice items centered on a single stimulus, often referred to as a testlet, violates the assumption of local independence. This occurs because an examinee’s response to one item within the testlet is impacted by an examinee’s response to another item within the same testlet (Wainer & Kiely, 1987). The practice of using one stimulus for a group of items creates local dependence among the items.

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

Alternatively, one of the measurement models based on testlet response theory (TRT) (Wainer, Bradlow, and Du, 2000) can be used. In TRT, the item associated with a given testlet remains the unit of measurement. With TRT the most frequently used dichotomous IRT models (1PL, 2PL, and 3PL) have been modified to include a random effect parameter to account for the shared variance among items within a testlet, called the testlet effect. The b-, a-, and c-parameters of the TRT models retain the same interpretations and meanings as with the dichotomous IRT models. By incorporating local dependence of items within a testlet into the model, the issue is no longer being ignored or sidestepped.

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

Exposure Control Procedures

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

An initial randomization procedure, 5-4-3-2-1 procedure, was proposed by McBride & Martin (1983). This procedure selects the first item for administration randomly from the five most informative items. The second item is randomly selected from the four most informative items. This process is continued such that the third and fourth items are randomly selected from the three and two most informative items, respectively, until the fifth item. The remaining items administered are selected based on maximum information. The randomesque procedure (Kingsbury & Zara, 1989) is similar to the 5-4-3-2-1 procedure by randomly selecting an item from a group of optimal items for administration. The randomesque procedure differs in that it continues to employ this selection technique throughout testing rather than switching to maximum information selection.

Lunz and Stahl (1998) developed the within .10 logits procedure that randomly selects an item from all items within .10 logits of the desired difficulty level. Therefore all items within the specified range are available for selection rather than an arbitrary number of items. This procedure is continued throughout testing. Davis and Dodd (2001) developed the modified within .10 logits procedure for polytomous items. Polytomous items do not have a single difficulty level; therefore the selection procedure was modified to select the items that yield the most information for a range of ability levels around the examinee's current ability level. More specifically, a total of six items are selected, the two items that provide the most information at the desired ability level, the two most informative items at the ability level minus .10, and the two most informative items at the ability level plus .10. A single item is then randomly chosen from the six selected items for administration.

The most commonly used conditional selection procedure is the Sympson-Hetter procedure (Sympson & Hetter, 1985). The Sympson-Hetter procedure assigns an exposure control parameter value ranging from zero to one for each item based on the frequency of item administrations during an iterative CAT simulation program. Items with high administration frequencies will have smaller exposure control parameters to limit their administration in a live CAT test. This ensures a maximum item exposure rate. Parshall, Davey, and Nering (1998) developed the conditional Sympson-Hetter procedure in which the exposure control parameters are determined based on ability level.

The a-Stratified procedure (Chang & Ying, 1996) stratifies the item pool based on the discrimination parameter, a. During the beginning of the CAT when an examinee's ability is unknown, lower discriminating items are administered. As the examinee's ability is determined, higher discriminating items are administered.

Previous MCAT Research

Previous research conducted through the Medical College Admissions Test (MCAT) Graduate Student Research Program on the reading passages of the MCAT has indicated the presence of local item dependence on the Verbal Reasoning section and to a lesser extent on the Biological Sciences and Physical Sciences sections (Zenisky, Hambleton, & Sireci, 2000). On the Verbal Reasoning section, Smith, Plake, and De Ayala (2001) reported high levels of item overexposure and underexposure when selecting both the items and reading passages adaptively based on the difficulty parameter of the Rasch IRT model. They found selecting the reading passages adaptively and the items randomly resulted in improved measurement precision relative to selecting the passages randomly and the items adaptively.

Davis and Dodd (2001) applied polytomous scoring and Masters' (1982) partial credit model to the reading passages in the Verbal Reasoning section to account for local item dependence and investigated several item exposure constraint procedures. They investigated four exposure control procedures – a randomization method, a modification of the Lunz and Stahl within .10 logits randomization procedure, the Luecht and Nungester's (1998) computerized adaptive sequential testing (CAST) procedure, and a no exposure control method that served as a baseline measure. In terms of exposure control procedures, the Davis and Dodd variation of the Lunz and Stahl's (1998) randomization procedure and the CAST procedure provided the best balance of exposure control relative to loss in measurement precision. Unfortunately, the implementation of the polytomous IRT model resulted in the loss of twenty-seven reading passages and their respective items due to low category response frequencies or convergence problems fitting the polytomous IRT model to the data. In addition, items associated with a given passage cannot be added or deleted to create different forms without recalibrating the items. The success of the polytomous scoring in eliminating the issue of local dependence was off set by these other issues. Testlet response theory might be a viable option.

This study compares item exposure control procedures within the three-parameter logistic testlet response theory model (Wang, Bradlow, & Wainer, 2002) and within a polytomous IRT model, Masters' (1982) partial credit model, in the context of a CAT using real data from 22 forms of the Verbal Reasoning section of the Medical College Admissions Test. Each of the CAT systems includes an item selection procedure, content balancing, and an item exposure control procedure. Four item exposure control methods are investigated for each of the CAT systems. Two variations of the Sympson-Hetter (1985) procedure are compared with the Davis and Dodd (2001) modification of the Lunz and Stahl (1998) randomization procedure. Maximum information selection is used as a baseline, no exposure control condition, from which to compare the other three exposure control procedures. Measurement precision and exposure rates are examined under each condition. Content balancing is based on reading passage content area and number of multiple-choice items per passage.

Partial Credit Model

Masters' (1982) partial credit model is a polytomous IRT model that scores each item response into more than two categories to represent varying degrees of ability. When the partial credit model is applied to testlet data, each item within a given testlet is scored correct or incorrect and summed to create the polytomous score for the testlet. Thus for each testlet i, an examinee's testlet score will be categorized in one of $m_i + 1$ category scores, ranging from 0 to m_i. The probability that an examinee with an ability level, θ, will obtain a score of x on testlet i is denoted

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

where b_{ik} represents the step difficulty or threshold of transitioning from one category of m_i to the next category. The PC model assumes that testlets within a given test do not differ in their discrimination power.

Item information, $I_i(\theta)$, for the partial credit model conditional on theta is denoted

$$I_i(\theta) = \sum_{x_i=0}^{m_i} \frac{[P'_{x_i}(\theta)]^2}{P_{x_i}(\theta)} \quad , \qquad (2)$$

where P' is the first derivative of Equation 1 (Koch & Dodd, 1989). Item information during a CAT administration is used in the selection process in that the item with the maximum information for an examinee's current ability level is selected contingent on content balancing and exposure control procedures.

Testlet Response Theory

The three-parameter logistic testlet response theory model (Wang, Bradlow, & Wainer, 2002) is a dichotomous IRT model with three item parameters, difficulty (b), discrimination (a), and guessing (c) parameters; and two person-specific parameters, theta (θ) and the testlet effect ($\gamma_{id(j)}$). The probability that an examinee with an ability level, θ, will obtain a score of *y* on testlet *d(j)* is denoted

$$p(y_{ij} = 1) = c_j + (1 - c_j)\left[\frac{\exp(a_j(\theta_i - b_j - \gamma_{id(j)}))}{1 + \exp(a_j(\theta_i - b_j - \gamma_{id(j)}))}\right] \quad , \qquad (3)$$

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

Item information, $I(\theta_i)$, for the testlet response theory model conditional on theta for a single item response is denoted

$$I(\theta_i) = a_j^2 \left(\frac{\exp(t_{ij})}{1+\exp(t_{ij})} \right)^2 \frac{1-c_j}{c_j + \exp(t_{ij})} \quad , \tag{4}$$

where $t_{ij} = (a_j(\theta_i - b_j - \gamma_{id(j)}))$ (Wainer, Bradlow, and Du, 2000). Testlet information is the sum of the item informations within a testlet. During the CAT administration, the testlet with the most information for an examinee's current ability level is selected contingent on content balancing and exposure control procedures.

Method

Overview

Two measurement models appropriate for testlets were used to evaluate the relative merits of four item exposure control procedures in the context of CAT. The measurement models were the three-parameter logistic testlet response theory (TRT) model and the partial credit (PC) model. The four item exposure control procedures investigated were two levels of the Sympson-Hetter (Sympson & Hetter, 1985) procedure, a modification of the Lunz & Stahl (1998) randomization procedure (Davis & Dodd, 2001), and a no item exposure control method. Maximum item information selection was used for no exposure control condition in order to provide a baseline from which to compare the three other exposure control procedures for each measurement model. Measurement precision and exposure rates were examined to evaluate the effectiveness of the exposure control procedures for each measurement model.

Item Pool

The data consisted of examinee responses from 22 forms of the Verbal Reasoning section of the Medical College Admissions Test administered from April 1996 to April 2001. The average number of examinees per form was 7,234 examinees with a minimum of 2,510 and a maximum of 14,439 examinees. Each form contained 8 reading passages and 55 multiple-choice items. The reading passages differed by content (humanities, social science, or natural science) and the number of multiple-choice items associated with the passages (6, 7, 8, or 10 items).

For the partial credit model, the item pool contained 149 passages scored as polytomous items. The PC model item pool consisted of 40% humanities, 36% social science, and 24% natural science passages. In terms of the number of items per passage, the PC model item pool consisted of 68% six-item, 20% seven-item, 7% eight-item, and 5% ten-item passages. For the testlet response theory model, the item pool contained 176 passages with a total of 1,210 dichotomous items. The TRT model item pool consisted of 37.5% humanities, 37.5% social science, and 25% natural science passages. In terms of the number of items per passage, the TRT model item pool consisted of 60% six-item, 18% seven-item, 10% eight-item, and 12% ten-item passages. The discrepancy in the number of testlets for the PC and TRT models was due to the low category frequencies and convergence problems when estimating the item parameters for the PC model (Davis & Dodd, 2001).

Parameter Estimation

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

The estimated testlet parameters for the PC model were obtained from the Davis and Dodd (2001) study. In that research, the same data for the MCAT forms described above were calibrated using the PARSCALE software program (Muraki & Bock, 1993).

For the TRT model, the item parameters were estimated with the SCORIGHT software program (Wang, Bradlow, & Wainer, 2001). The three-parameter logistic model with the testlet effect, $\gamma_{id(j)}$, was used. The testlet effect was allowed to vary across testlets for all examinees. SCORIGHT employs a Markov Chain Monte Carlo technique with Gibbs sampling to draw inferences from the posterior distribution of the parameters to estimate the parameters of the model. The MCAT data were analyzed using 8000 iterations of which the first 7000 iterations were dropped. Every fifth-iteration of the remaining 1000 iterations was selected to create the posterior distribution of the parameters.

Data Generation

The PC model item response data was generated using the IRTGEN SAS macro (Whittaker, Fitzpatrick, Williams, & Dodd, in press). Response data was generated for 1,000 simulees. Each simulee was assigned a known theta value by randomly selecting theta from a normal distribution with mean zero and standard deviation equal to one. Based on the parameter estimates obtained from the calibration of the MCAT data and the simulee's theta value, the probability of responding in each category for a given testlet was calculated. The category probabilities for a given testlet were then summed to create cumulative subtotal probabilities for each response category. A random number was then selected from a uniform distribution that ranged from 0 to 1 and compared to the cumulative subtotal probabilities. If the random number was less than the subtotal probability for a given category, the simulee's response was that category score. This process was repeated for every testlet and every simulee. The resulting generated response data was used for each PC model CAT condition.

The TRT item response data was generated for 1,000 simulees. Each simulee was assigned a known theta value by randomly selecting theta from a normal distribution with mean zero and standard deviation equal to one. The probability of responding to an item was based on the simulee's theta value, the item parameter estimates obtained from SCORIGHT, and a generated person-specific testlet effect. The testlet effect parameter was determined by selecting a random variable from a normal distribution with mean zero and standard deviation equal to the square root of the variance of the testlet effect for a given testlet. The selected random number was used as the testlet effect parameter in the probability model for all items in a testlet for the simulee. In order to introduce random error, the simulee's response was compared to a randomly selected number from a uniform distribution that ranged from 0 to 1. The simulee received a correct response (1) if the random number was less than the simulee's response and an incorrect response (0) otherwise. This process was repeated for every item and every person. The same generated response data was used for each CAT condition based on the TRT model.

CAT Simulations

The CAT simulations were based on a SAS program created by Chen, Hou, & Dodd (1998) and modified by Davis & Dodd (2001). The initial theta estimate was set to 0.0 representing the mean of the population. Each CAT consisted of item selection based on maximum information contingent on content balancing and exposure control procedures. The ability and the person-specific testlet effects were estimated using expected a posteriori (EAP) procedures after each testlet was administered. The stopping rule for test administration was seven reading passages resulting in the administration of 50 items.

For administration of the first reading passage, the content and the number of items per passage were randomly selected for each examinee. The remaining reading passages were selected using the Kingsbury & Zara (1989) procedure, which compares the target proportions for content balancing to the actual proportions during test administrations and selects the next item from the content with the largest discrepancy between the target and actual proportions. Therefore, each simulated test consisted of 40% humanities, 36% social science, and 23% natural science reading passages. Concurrently, the Kingsbury & Zara (1989) procedure controlled the number of items per passage such that each simulated test consisted of 42% six-item, 28% seven-item, 14% eight-item and 14% ten-item reading passages.

The exposure control procedures were the modified within .10 logits (Davis & Dodd, 2001) and the Sympson-Hetter procedure (Sympson & Hetter, 1985). The Sympson-Hetter procedure was examined at two levels: a maximum exposure rate equal to .19 and a maximum exposure rate equal to .29. Maximum information with no exposure control served as the baseline condition.

Data Analyses

Assessment of the CAT systems was based on retrieval of simulees' known theta values and the effectiveness of the exposure control procedures. The degree to which the CAT systems recovered the known theta values was evaluated through descriptive statistics, the Pearson product-moment correlation, bias, standardized difference between means (SDM), root mean squared error (RMSE), standardized root mean squared difference (SRMSD), and average absolute difference (AAD). The following equations illustrate the computation of bias, RMSE, SDM, SRMSD, and AAD:

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

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

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

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

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

where $\hat{\theta}_k$ is the estimated ability level obtained from the CAT and θ_k is the known ability level used to generate the response data for person *k*.

Evaluation of the exposure control procedures was based on descriptive statistics of the item exposure rates including frequency, mean, standard deviation, and maximum exposure rate. Simulees' audit trails were examined to determine the frequency with which an item was administered in each CAT condition. The item exposure rate represented the number of times an item was administered to simulees divided by the total number of simulees. The percentage of items not administered during any of the CAT administrations represented pool utilization. In addition, each item was evaluated for test overlap across all simulees, simulees with similar abilities and simulees with different abilities. Simulees' audit trails were compared to determine the test overlap. Similar abilities were defined as simulees having theta values within 2 logits and different abilities were simulees with discrepancy in theta values larger than 2 logits.

Results

Descriptive Statistics

The degree of dependency present in the testlets used in the current research was examined for the TRT model. The mean of the variance of the testlet effect was 0.49 with a standard deviation of 0.35. The minimum was 0.01 and the maximum was 1.67. Since the testlet effect was allowed to vary across testlets, testlets were examined for differences in the testlet effects. Specifically, the variances of the testlet effect parameters were compared across content of the reading passages and number of items per reading passage. A significant difference ($F(2,173) = 6.25$, $p = 0.0024$) in the estimates of the testlet effect parameter variances estimates was found between the reading passages (humanities, social science, and natural science). A post-hoc Tukey's test indicated a significant difference between the mean testlet effect variances between humanities (mean = 0.588) and natural science reading passages (mean = 0.358). The mean for social science was 0.489. Analysis of variance yielded no significant differences in the means for the number of items per reading passage (6, 7, 8, and 10); F = .22, df = 3, 172, and p-value = 0.8806.

The mean and standard deviation for the estimated theta and standard error for each of the exposure control procedures are listed in Table 1 for the partial credit model and Table 2 for the testlet response theory model. The known theta for the PC model was -0.04 with a standard deviation equal to 1.02. The known theta for the TRT model was -0.01 with a standard deviation equal to 1.04. The estimated theta and standard deviation for each condition under each model approximated a normal distribution with mean zero and standard deviation equal to one. The mean standard error for the PC model ranged from 0.28 to 0.31. The mean standard error for the TRT model ranged from 0.31 to 0.35.

Insert Tables 1 and 2 about here

The correlation between the known theta and the estimated theta for each condition and the measurement statistics, bias, standardized difference between means (SDM), root mean squared error (RMSE), standardized root mean squared difference (SRMSD), and average absolute difference (AAD), are reported in Table 3 for the PC model. The correlation coefficients obtained for the PC model under both of the Sympson-Hetter conditions and the maximum information condition were 0.96. The correlation coefficient was slightly lower (0.95) for the modified within .10 logits condition. The bias and SDM statistics were functionally zero when rounded to the second decimal place for each condition except the Sympson-Hetter with a maximum exposure control rate set to .29 yielding 0.01 for bias and SDM. The RMSE ranged from 0.29 to 0.31, the SRMSD ranged from 0.53 to 0.57 and the AAD statistic ranged from 0.22 to 0.24. For each statistic, the modified within .10 logits procedure reported the highest value although only slightly higher.

Insert Table 3 about here

The same measurement statistics that were presented for the PC model in Table 3 are shown in Table 4 for each of the exposure control conditions under the TRT model. The correlation coefficients obtained for all of the exposure control conditions under the TRT model were 0.92. The bias statistic ranged from -0.02 to 0.01 and the SDM statistics ranged from -0.01 to 0.02. For each statistic the largest difference was reported for the maximum information procedure. The RMSE ranged from 0.40 to 0.41, the SRMSD ranged from 0.64 to 0.66 and the AAD statistic ranged from 0.32 to 0.33. For each statistic, the Sympson-Hetter (.29) procedure reported the lowest value although only slightly lower.

Insert Table 4 about here

Pool Utilization and Exposure Rates

For the PC model, the frequency of items by exposure rate, the number of items not administered, the mean, standard deviation, and maximum exposure rate, and the percent of items in the item pool not administered are presented in Table 5 for each of the exposure control procedures. Maximum information yielded a high number of items never administered (92) resulting in 62% of the item pool not administered. The Sympson-Hetter (.19) and Sympson-Hetter (.29) procedures reported 51% and 56% of the item pool not administered, respectively. The modified within .10 logits procedure administered most of the items in the item pool, reporting 27% of the items not administered. The maximum exposure rate for the exposure control procedures were 0.29 for Sympson-Hetter (.29), 0.21 for Sympson-Hetter (.19), and 0.18 for the modified within .10 logits procedure. Although, the maximum exposure rate for the maximum information procedures is expected to be 1.0, due to the first reading passage being randomly selected in terms of content balancing, the maximum exposure rate was 0.61 for maximum information.

Insert Table 5 about here

For the TRT model, the frequency of items by exposure rate, the number of items not administered, the mean, standard deviation, and maximum exposure rate, and the percent of items in the item pool not administered are reported in Table 6 for each of the exposure control procedures. Maximum information yielded the highest number of items never administered (125) resulting in 70% of the item pool not administered. The Sympson-Hetter (.19) and Sympson-Hetter (.29) procedures reported 58% and 64% of the item pool not administered, respectively. The modified within .10 logits procedure administered most of the items in the item pool, reporting 31% of items not administered. The maximum exposure rate for the exposure control procedures were 0.70 for maximum information, 0.31 for Sympson-Hetter (.29), 0.22 for Sympson-Hetter (.19), and 0.24 for the modified within .10 logits procedure.

Insert Table 6 about here

Item Overlap

Table 7 presents the item overlap results for the CATs based on the PC model. The mean item overlap values were highest for the maximum information procedure across all three conditions (overall average overlap (1.91), different abilities average overlap (0.47), and similar abilities average overlap (2.20)), when compared to the other exposure control procedures. The modified within .10 logits procedure yielded the lowest mean item overlap across all ability levels (0.71) and for similar abilities (0.77) in contrast to the Sympson-Hetter procedures and maximum information method. On the other hand, both the Sympson-Hetter (.19) and Sympson-Hetter (.29) yielded a lower mean item overlap than the modified within .10 logits procedure for the different ability levels. For all three overlap calculations, the Sympson-Hetter (.19) yielded lower mean item overlap than the Sympson-Hetter (.29) procedure.

Insert Table 7 about here

The item overlap results for the 3 PL TRT model are listed in Table 8. The exposure control procedures yielded the same pattern of findings as the PC model, but with slightly higher mean overlap values. For maximum information, the mean overlap across all ability levels was 2.47 items. As expected, the mean overlap for similar abilities was higher at 2.74 items and the mean overlap for different abilities was lower at 1.17 than the mean overlap calculated on all ability levels. The Sympson-Hetter (.29) procedure yielded the second highest item overlap means. For similar abilities, the mean overlap was highest at 1.70 items. The mean for the overall overlap was 1.51 items and the mean overlap for different abilities was 0.59. The Sympson-Hetter (.19) procedure yielded a mean overall item overlap of 1.09 items, a mean overlap of 0.52 items for different ability levels, and a mean overlap equal to 1.22 items for similar ability levels. Across the exposure control procedures, the modified within .10 logits procedure yielded the lowest mean overall overlap (0.80) and the lowest mean overlap (0.85) for similar ability levels. For different ability levels, the Sympson-Hetter (.19) procedure yielded a lower mean overlap than the other exposure control conditions.

Insert Table 8 about here

Discussion

Item Pool

The item pool for the CAT conditions under the partial credit model contained 149 of 176 reading passages from 22 forms of the Verbal Reasoning section of the MCAT. The omission of 27 reading passages and their respective items was due to sparse data restricting the ability to estimate the item parameters. The polytomous scoring of the reading passages led to small frequencies of responses in the lower category scores. The item pool for the CAT conditions under the testlet response theory model contained all 176 reading passages and their respective items. The items rather than the testlet are the unit of analysis with TRT therefore the calibration was not impacted by the sparse data found with the PC model. The decision to use different item pools for the CAT system based on the PC and TRT models, respectively, stemmed from wanting to use all available items for each model. Also the item parameters for the measurement models were not equated thereby making direct comparisons across the models inappropriate since most of the outcome measures are scale dependent.

Exposure Control Procedures

The precision of measurement outcome measures were very similar across the exposure control conditions within the partial credit model and the testlet response theory model. Although there were slight variations in the correlation coefficients and measurement statistics, the exposure control procedures yielded high levels of measurement precision as evidenced by the small bias and standardized difference between means statistics. The correlation coefficients for the PC model reflect those found in other CAT research using the PC model and the modified within .10 logits procedure (Davis & Dodd, 2001; Davis, 2002) and the Sympson-Hetter procedure (Davis, 2002). The correlation coefficients for the TRT model mirror results from the Wang, Bradlow, and Wainer (2002) study in which the variance of the testlet effect equaled 0.50.

Although the modified within .10 logits procedure did not permit specification of a restricted maximum exposure rate, it yielded the lowest maximum exposure rate for the PC model (.18) and the second lowest for the TRT model (.24). The lowest maximum exposure rate for the TRT model was .22 for the Sympson-Hetter (.19) procedure. For both the PC and TRT models, the modified within .10 logits procedure administered considerably more of the item pool than the other exposure control procedures. The percent of pool not administered for the PC model was 27% and for the TRT model was 31%. This was a sharp contrast to the other exposure control procedures that utilized less than 50% of the item pool. The percent of pool not administered by the maximum information procedure was expected to be high due to the fact that it was a no exposure control condition and therefore the administration of the most informative item at each item selection stage during the CATs. Both the Sympson-Hetter procedures restricted the maximum exposure rate, but did not administer much of the item pool for either the PC model or the TRT model.

The modified within .10 logits exposure control procedure yielded the lowest mean item overlap across all ability levels for both the PC and TRT models. More importantly for examinees with similar abilities (abilities within 2.0 logits), the mean overlap was less than one, indicating that examinees with similar abilities will most likely not receive the same reading passages using the modified within .10 logits procedure. This reduces the opportunity for examinees to share knowledge of the test content.

Conclusion

Of the exposure control procedures investigated in the current research, the modified within .10 logits procedure (Davis & Dodd, 2001) yielded the best balance between measurement precision and test security for both the partial credit model and the testlet response theory model. In addition, the modified within .10 logits procedure is easier to implement than the Sympson-Hetter procedures. For the MCAT data, a CAT system based upon either the PC model or the TRT model with the modified within .10 logits item exposure control procedure and content balancing using the Kingsbury and Zara (1989) procedure performed very well. In order to determine which of the two measurement models would be preferred for an operational CAT for MCAT, additional research needs to be conducted.

The three-parameter logistic testlet response theory model (Wang, Bradlow, & Wainer, 2002) offers an advantage over the partial credit model by keeping the item as the unit of measurement, rather than the testlet being the unit of measurement. As evidenced in the present study, the pool of available items was larger for the TRT model than the PC model. The CAT system based on the 3PL TRT model used in the current research, adapted the test at the testlet level rather than at the item within the testlet level. CATs based on one of the TRT models that allow selecting items adaptively *within* a testlet might further expand the functional item pool size and possibly allow for content balancing based on the cognitive level of the item. The MCAT Verbal Reasoning section items are categorized according to cognitive level and yet this information has not been used in the CAT applications. Future research is needed to determine if content balancing on the basis of the cognitive level of the item would enhance a CAT version of the MCAT. The use of other exposure control procedures such as the conditional Sympson-Hetter (Parshall, Davey, & Nering, 1998), randomesque procedures (Kingsbury & Zara, 1989) and the a-Stratified procedure (Chang and Ying, 1996) with the TRT models also need to be explored before recommendations for an operational CAT of the MCAT can be made.

References

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

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

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

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

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

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

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

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

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),* 335-357.

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

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

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

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

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.

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

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

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

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

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

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and application. *Applied Psychological Measurement, 26* (1), 109-128.

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

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

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

Zenisky, A. L., Hambleton, 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
Mean (and Standard Deviation) of the Estimated Thetas and Standard Errors yielded by the Partial Credit Model (N = 1000)

Exposure Control Procedure	Theta Estimate*	Standard Error
Maximum Information	-0.04 (0.97)	0.28 (0.05)
Modified within .10 logits	-0.04 (0.95)	0.31 (0.06)
Sympson-Hetter (.19)	-0.04 (0.98)	0.29 (0.05)
Sympson-Hetter (.29)	-0.05 (0.98)	0.29 (0.05)

* Known Thetas: Mean = -0.0416 and SD = 1.0211

TABLE 2
Mean (and Standard Deviation) of the Estimated Thetas and Standard Errors yielded by the Testlet Response Theory Model (N = 1000)

Exposure Control Procedure	Theta Estimate*	Standard Error
Maximum Information	0.01 (0.90)	0.31 (0.03)
Modified within .10 logits	-0.02 (0.90)	0.35 (0.04)
Sympson-Hetter (.19)	0.00 (0.92)	0.33 (0.03)
Sympson-Hetter (.29)	0.00 (0.93)	0.32 (0.03)

* Known Thetas: Mean = -0.0086 and SD = 1.0355

TABLE 3
Partial Credit Model 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)
(N = 1,000)

Exposure Control Procedure	Correlation	Bias	SDM	RMSE	SRMSD	AAD
Maximum Information	0.96	0.00	0.00	0.30	0.54	0.22
Modified within .10 logits	0.95	0.00	0.00	0.31	0.57	0.24
Sympson-Hetter (.19)	0.96	0.00	0.00	0.29	0.53	0.22
Sympson-Hetter (.29)	0.96	0.01	0.01	0.29	0.54	0.23

TABLE 4
Testlet Response Theory Model 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)
(N = 1,000)

Exposure Control Procedure	Correlation	Bias	SDM	RMSE	SRMSD	AAD
Maximum Information	0.92	-0.02	0.02	0.41	0.66	0.33
Modified within .10 logits	0.92	0.01	-0.01	0.41	0.66	0.33
Sympson-Hetter (.19)	0.92	-0.01	0.01	0.42	0.66	0.33
Sympson-Hetter (.29)	0.92	0.00	0.00	0.40	0.64	0.32

TABLE 5
Pool Utilization and Exposure Rates for the CATs Based on the Partial Credit Model
(N = 1000)

	Exposure Control Procedure			
Exposure Rate	Maximum Information	Modified within .10 logits	Sympson-Hetter (.19)	Sympson-Hetter (.29)
1.0	0	0	0	0
0.91 – 0.99	0	0	0	0
0.81 – 0.90	0	0	0	0
0.71 – 0.80	0	0	0	0
0.61 – 0.70	1	0	0	0
0.51 – 0.60	0	0	0	0
0.41 – 0.50	2	0	0	0
0.36 – 0.40	4	0	0	0
0.31 – 0.35	0	0	0	0
0.26 – 0.30	0	0	0	7
0.21 – 0.25	5	0	3	7
0.16 – 0.20	4	9	24	6
0.11 – 0.15	6	14	6	8
0.06 – 0.10	14	35	13	13
0.01 – 0.05	21	51	27	24
Not Administered	92	40	76	84
Mean Exposure Rate	0.05	0.05	0.05	0.05
SD Exposure Rate	0.10	0.05	0.07	0.08
Max Exposure Rate	0.61	0.18	0.21	0.29
% of Pool Not Administered	62%	27%	51%	56%

TABLE 6
Pool Utilization and Exposure Rates for the CATs Based on the Testlet Response Theory 3PL Model
(N = 1000)

	Exposure Control Procedure			
Exposure Rate	Maximum Information	Modified within .10 logits	Sympson-Hetter (.19)	Sympson-Hetter (.29)
1.0	0	0	0	0
0.91 – 0.99	0	0	0	0
0.81 – 0.90	0	0	0	0
0.71 – 0.80	0	0	0	0
0.61 – 0.70	1	0	0	0
0.51 – 0.60	2	0	0	0
0.41 – 0.50	1	0	0	0
0.36 – 0.40	1	0	0	0
0.31 – 0.35	2	0	0	3
0.26 – 0.30	2	0	0	10
0.21 – 0.25	3	6	3	4
0.16 – 0.20	3	6	25	3
0.11 – 0.15	6	12	4	2
0.06 – 0.10	10	22	12	15
0.01 – 0.05	20	75	28	27
Not Administered	125	55	104	112
Mean Exposure Rate	0.04	0.04	0.04	0.04
SD Exposure Rate	0.11	0.06	0.07	0.08
Max Exposure Rate	0.70	0.24	0.22	0.31
% of Pool Not Administered	71%	31%	58%	64%

TABLE 7
Mean and (Standard Deviation) of Overall Average Overlap, Different Abilities Average Overlap, and Similar Abilities Average Overlap for the Partial Credit Model.

Exposure Control Procedure	Overall Average Overlap (N = 499,500)	Different Abilities Average Overlap (N = 82,329)	Similar Abilities Average Overlap (N = 417,171)
Maximum Information	1.91 (1.61)	0.47 (0.77)	2.20 (1.58)
Modified within .10 logits	0.71 (0.82)	0.41 (0.63)	0.77 (0.84)
Sympson-Hetter (.19)	1.06 (1.18)	0.31 (0.59)	1.21 (1.21)
Sympson-Hetter (.29)	1.32 (1.32)	0.38 (0.66)	1.50 (1.34)

TABLE 8
Mean and (Standard Deviation) of Overall Average Overlap, Different Abilities Average Overlap, and Similar Abilities Average Overlap for the Testlet Response Theory model.

Exposure Control Procedure	Overall Average Overlap (N = 499,500)	Different Abilities Average Overlap (N = 87,315)	Similar Abilities Average Overlap (N = 412,185)
Maximum Information	2.47 (1.56)	1.17 (1.12)	2.74 (1.50)
Modified within .10 logits	0.80 (0.83)	0.57 (0.72)	0.85 (0.85)
Sympson-Hetter (.19)	1.09 (1.09)	0.52 (0.75)	1.22 (1.11)
Sympson-Hetter (.29)	1.51 (1.33)	0.59 (0.83)	1.70 (1.33)