Graduate
Management
Admission
Council®

Optimizing Item Exposure Control Algorithms for Polytomous Computerized Adaptive Tests With Restricted Item Banks

Michael Chajewski & Charles Lewis
Fordham University

Presented at the Item Exposure Paper Session, June 3, 2009

Abstract

In the face of modern-day developments in translating assessment instruments into computer-based adaptive measures, more and more concerns regarding test security and test validity have been raised. Further, as non-cognitive measures are being transformed into functioning CATs, item exposure control has become an important component of managing item banks. With the inception of evaluating exposure control mechanisms conditionally, many of the non-high stakes assessments are now seeking solutions to issues associated with constrained item banks and limited item availability. This simulation study compared three item exposure control mechanisms (Sympson-Hetter conditional, progressive restrictive maximum information, and total rate simplified exposure) against two unmanaged CAT administrations (simple maximum information and random item selection) for two restrictive item bank conditions. The competing methods were evaluated in their ability to provide conditional exposure control along with adequate bias, root mean square errors, and deviations of asymptotic valid standard errors from observed conditional standard deviations. Benefits and shortcomings are discussed.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Copyright © 2009 by the Authors

Citation

Chajewski, M. & Lewis, C. (2009). Optimizing item exposure control algorithms for polytomous computerized adaptive tests with restricted item banks. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Michael Chajewski, Fordham University, Department of Psychology, Psychometrics, 441 East Fordham Road, Bronx, NY 10458. Email: chajewski@fordham.edu

Optimizing Item Exposure Control Algorithms for Polytomous Computerized Adaptive Tests With Restricted Item Banks

Much of present day research focusing on computerized adaptive testing (CAT) focuses on its application in educational and cognitive assessments. Generally, this research uses item response theory (IRT) models for dichotomously scored items and relatively large item banks. More recently, developers and publishers in the testing industry have dedicated a portion of their resources to the evaluation of adaptive testing in the context of non-cognitive assessments, such as industrial-organizational inventories, low-stakes surveys, and a multitude of psychological evaluations.

Compared to operational cognitive educational testing programs, non-cognitive assessments are faced with similar testing issues but in a slightly different framework. For instance, large-scale assessments usually have access to relatively large item banks. An item bank with several hundred items from which the adaptive algorithm can choose to administer an item will provide flexibility and precision in the assessment of the examinee. Furthermore, it is possible to extend such item banks since the nature of most cognitive questions (e.g., algebra, reading comprehension) lends itself to the relatively easy development of new items. This is not the case with some non-cognitive assessments. There are many ways in which to test whether a person can add; however, there are not as many ways to inquire of an individual if he/she is depressed.

It is the explicit utility of adaptive tests to tailor assessments to examinees' particular level of ability (or proficiency). However, it is easily conceived that more than one individual with essentially the same level of ability will take the test. In the case of an adaptive measure this would suggest that these two individuals would likely be exposed to many of the same items. This creates an issue of test security in high-stakes tests where the overexposure of an item to a given group of individuals might compromise that item, making it not only useless but faulty in the assessment function of the measure. In the case of non-cognitive assessments, this is not so much a case of test security as one of validity. The public or non-public disclosure of an item would pose a threat to the assessment's validity in that if disclosed the item might no longer correctly assess the content for which it was developed. In order to avoid the dangers of overexposing an item, the selection algorithm must somehow control for the frequency with which an item is administered. This concept constitutes the logical basis behind item exposure control mechanisms (Weiss, 1983; Wainer, 2000)

Two additional sources of restriction arise in the case of controlling for item exposure in CAT measures that use small item banks. As compared to larger assessments, there are usually fewer examinees to whom the test is administered. Having fewer examinees usually (though not always) results in smaller revenues generated by the test's administration. With a significantly smaller budget, the instrument's research and evaluation are likely to be limited in addition to halting the expansion of the usually costly development of new items for the bank. In addition, as is the case with all assessments, there is the persistent issue of test-and-retest. Individuals who will have been assessed could (and in many instances will) be re-tested. In the case of paper-pencil instruments, where several test forms exist and new forms are constantly assembled, the examinee would never see the same form twice.

In the classical CAT environment, the dynamic test assembly is determined by the examinee's trait level. If that has not changed, the test would look largely the same as during the previous/original administration. If the item bank is sufficiently large the algorithm might have a choice between items of similar difficulty and discrimination. In that case the examinee might (but need not) be administered more of the items he/she has not seen previously because the algorithm would randomly choose between items that are equally suitable. The predicament arises when items differ in their psychometric properties and when the item bank from which items are taken is relatively small. In such a case an item exposure control algorithm would track the selection of an item across examinees (and potentially across global administrations) and constrain the administration procedure to choose only those items that have not been used too frequently. Subsequently, if an item that is particularly well suited for discriminating between two adjacent trait levels has been used too often, the algorithm would select the next best item. However, Stocking and Lord (1983) and Stocking and Swanson (1993) noted that there is a trade-off in this procedure. Greater item selection restriction will naturally result in the loss in precision of trait estimation since not all of the items administered are those that would provide the maximum amount of information about the particular examinee.

Presently there are many item exposure control algorithms, with varying utility and computational procedures. Georgiadou, Triantafillou and Economides (2007) reviewed the most common procedures developed from 1983 to 2005. Based on their review, they classified the published methods into five categories: randomization techniques, conditional selections, stratified strategies, combined methods, and multiple stage adaptive test designs (p. 7). Each class of methods provides benefits and shortcomings over the others. This is largely the case since most developers of item exposure control mechanisms have a specific goal in mind or attempt to address a specific program need. As mentioned above, larger testing programs will be faced with different practical issues than smaller endeavors (Kingsbury & Zara, 1991).

The specific aim of the research presented in this paper was to simulate conditions likely to be encountered with short non-cognitive CATs. Furthermore, the primary objective was to investigate and optimize the ability to control for item exposure control in assessments using relatively small item banks while preserving, as well as possible, the precision of latent trait estimation.

Method

The hypothetical test used for the assessment of the chosen algorithms consisted of a twenty-item inventory administered to every participant. As with most operational CATs, the test termination criterion was a fixed test length rather than reaching a pre-set asymptotic standard error of measurement. This choice was largely guided by the fact that most CATs will use fixed test length as a method to further provide fair assessment for all examinees. Moreover, in restricted item banks the danger persists that if the test termination criterion is left to the estimation of error components and a particular examinee is not responding systematically as expected of his/her latent trait level, the algorithm could potentially run out of items to administer. Particularly, if a given measure seeks to further subdivide its items into content specific areas, constraining the minimum number of items necessarily administered from any one subject group, a small item bank CAT could quickly get in trouble attempting to optimize error estimates conditional within each subject group or globally across the different content areas. The research presented here did not constrain items to specific question clusters, but aimed at providing a realistic scenario in which this might have been the case, providing additional support for the use of fixed test length in the simulated conditions.

While most educational and proficiency assessments are scored dichotomously (correct-incorrect), a polytomous IRT model was chosen for this study. Compared to the standard dichotomous models (1, 2, and 3-parameter logistic models), polytomous models have not been as widely and thoroughly researched in the context of CATs. Furthermore, most of the current psychological and industrial measures being adapted to CATs do not focus on the correct or incorrect endorsement of their items, but rather the degree of agreement with a given statement or a specific leaning/preference toward a specific extreme opinion.

Data Simulation

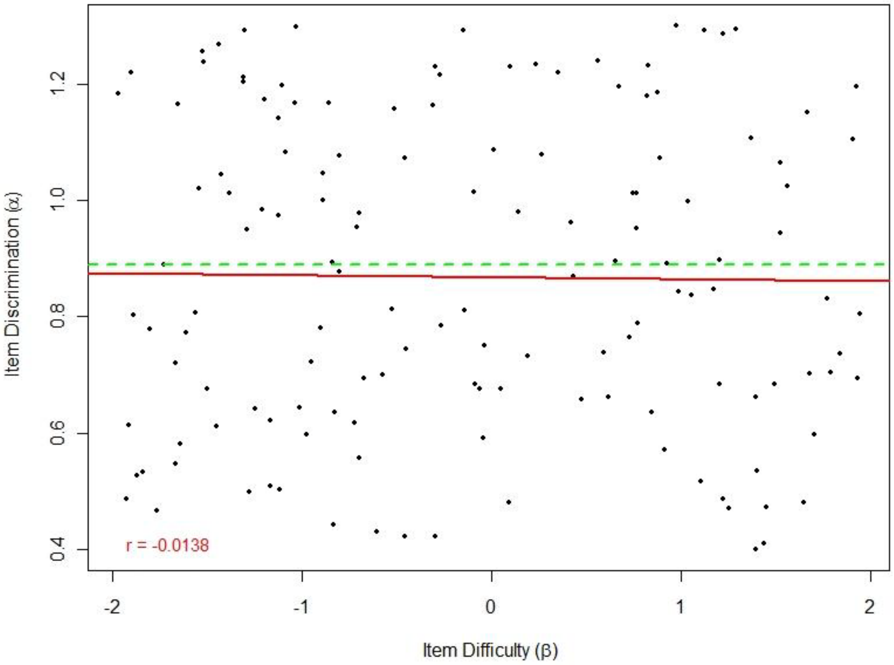
Two item bank conditions were simulated. The ratio of test length to total item bank size that constitutes restrictive item bank conditions is highly debatable. Several larger testing companies that utilize CAT procedures tend to maintain a reasonable 1:12 ratio of test length to item bank size. Compared to this standard, restricted item banks for this study (using a 20-item test length) had ratios of 1:7 (140 items) or 1:4 (80 items).

Since the simulated items were intended to emulate ordinal or gradient-like item responses, the modified graded response model (M-GRM) was used in simulating IRT item parameters. First proposed by Samejima (1969; 1996), the graded response model (GRM) incorporates information from an ordinal or Likert-type scale into the IRT framework by treating the ordered responses as a continuous response spectrum that is subdivided into several categories divided by thresholds. Later, Muraki (1990, 1992, 1993) extended Samejima's model to the modified form in which the discrimination parameters are allowed to vary across items and the "between category threshold parameters of the GRM are partitioned into two terms, namely, a location parameter for each item (here referred to as the item difficulty parameter), and a set of category threshold parameters for the entire scale" (Embertson & Reise, 2000, p.102).

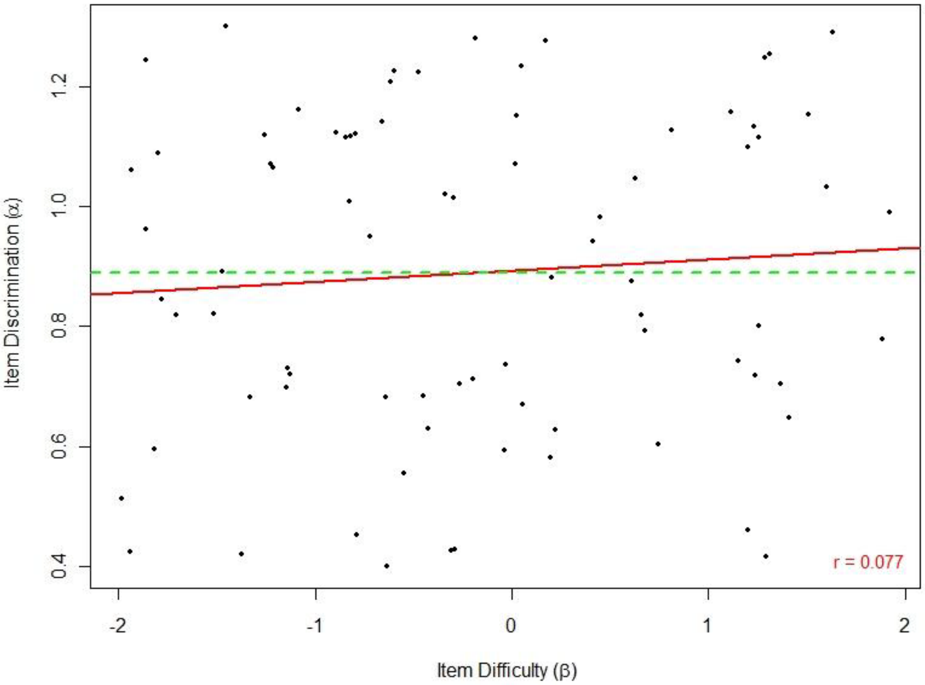
The simulated item difficulty or location parameters (β) were randomly drawn from a uniform distribution on the interval between –2.00 and 2.00. Item discrimination (or slope) parameters were randomly selected from a uniform distribution in the interval between .40 and 1.30. These item specifications were intended to cover a fairly broad range of items, ranging from easy to difficult and items with low discriminations to items with relatively high discrimination between individuals of different latent trait levels. Figure 1 shows the relationship between the items' discrimination and difficulty parameters for the two banks. As was intended, these parameters did not covary systematically in either bank. Furthermore, rather than specifying a particular latent trait (θ) distribution, θs were symmetrically chosen around zero at nineteen quadrature points, from –2.25 to 2.25 in steps of .25. This choice was made so as to simulate a set of examinees with varying trait levels. In the case of most non-cognitive assessment, the focus is not one of proficiency at a particular level (although certainly some psychological assessments stress the precision of their trait estimation at a particular cut score) but rather to use the measure as an investigative tool for a broad spectrum of possible examinees.

Figure 1. Simulated Item Difficulties and Item Discrimination Parameters (Regression Line in Solid Red and Constant Average Item Discrimination in Broken Green)

a. 140-Item Bank



b. 80-Item Bank

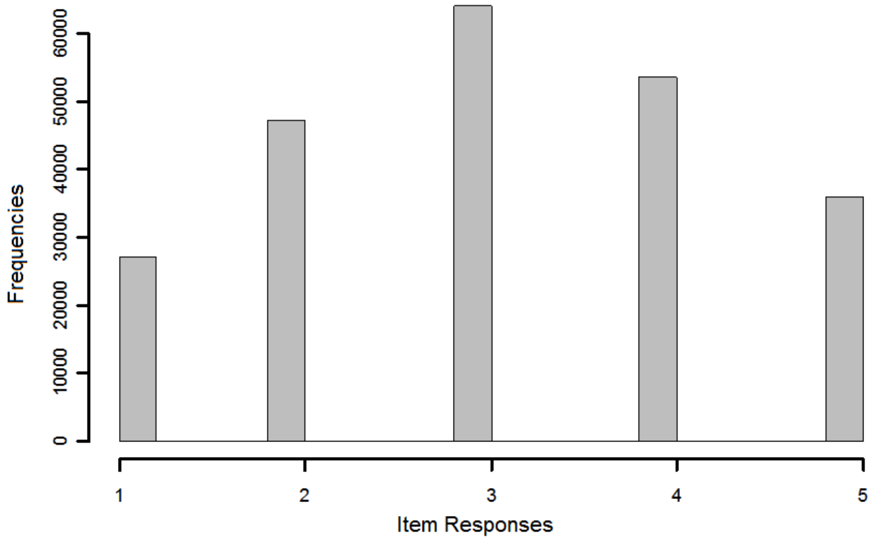


Consequently, the generated items and true θ values were created so as to reflect this particular scenario. Also, since most of the present day assessments of exposure control mechanisms are suggested to be conducted conditionally, equally represented θ values allowed for a relatively fair chance to each of the employed methods, in that if the algorithm failed to reasonably constrain the exposure rate at one θ level it still had the opportunity to outperform the other algorithms for other θ levels. For each true θ value, 150 cases (or unique response strings) were simulated, resulting in a data matrix of 2,850 cases for both the 140-item and 80-item bank conditions.

Modified Graded Response Model (M-GRM)

The structure of all the simulated items was set to five ordered response categories. With five (generally denoted as K) categories there are four specific thresholds identifying the particular response. Since the simulated true θ for all of the examinees were chosen to have a uniform distribution and the simulated 140- and 80-item CATs were intended to simulate agreement type questionnaires, the population distribution across all items and all examinees for the given responses were modeled to group fairly normally around the center response category (i.e., choice 3). Figure 2 shows the frequency distribution of the simulated item responses pooled across all examinees and all items. Since there was no one particular criterion response category for which the test was designed, nor a particular θ level at which the measures were to perform best in discriminating between adjacent levels, it was reasonable to model a “normal- like” distribution of item responses.

Figure 2. Item Responses Combined Over All Items and All θ Values



The threshold parameters used to achieve the distributional qualities delineated above were −2.40, −.80, .80, 2.40, following the convention of Samejima’s graded response model (Samejima, 1969; Muraki, 1990; Embertson & Reise, 2000), where the item response probability of a particular response given the person’s θ can be expressed as

$$P_{jk}\left(\theta_i\right)=\Pr\left(y_{ij}=k\left|\theta_i\right.\right),\text{ for } k=1,\cdots,K. \quad (1)$$

Equation 1 expresses the probability of being between two thresholds, where the actual probability of giving any one of the possible categorical responses is defined as the difference between two consecutive item category response functions

$$P_{jk}\left(\theta_i\right)=P^*_{jk}\left(\theta_i\right)-P^*_{j,k+1}\left(\theta_i\right) \quad (2)$$

that are defined as the IRT polytomous model version of the operating characteristic curve,

$$P^*_{jk}\left(\theta_i\right)=\frac{\exp\left[D\alpha_j\left(\theta_i-\beta_{jk}\right)\right]}{1+\exp\left[D\alpha_j\left(\theta_i-\beta_{jk}\right)\right]}=\Pr\left(y_{ij}\geq k\left|\theta_i\right.\right) \quad (3)$$

for $k=2,\cdots,K$, $P^*_{j1}\left(\theta_i\right)=1.0$, and $P^*_{jK+1}\left(\theta_i\right)=0.0$. A rating scale version of the graded response model decomposes the item category location parameter β_{jk} as $\beta_{jk}=\beta_j+\gamma_k$, with the constraint $\sum_{k=2}^{K}\gamma_k=0$ used to identify the model parameters. Since the responses are ordered, there is the further constraint that $\gamma_2\leq\cdots\leq\gamma_K$. As long as the constraint threshold values are ordered, $\gamma_2\leq\gamma_3\leq\gamma_4\leq\gamma_5$, all the consecutive differences (defining the item category response functions) will be non-negative. Moreover, the sum of the item category response functions will add to 1.00.

Design

The primary focus of most item exposure control algorithms is to establish and control the rate at which an item is administered to examinees. If the length of the CAT is n_t items, and the item bank has n_p items, then—regardless of the item selection algorithm used for the CAT—the average exposure rate for all items in the bank must be n_t / n_p. Therefore, the maximum item exposure rate must always be greater than or equal to this ratio. Ideally, the target exposure rate should be such that items are not overexposed, but at the same time not so restrictive so as to constrain all the items and prevent the CAT algorithm from reliably estimating the person's θ.

Each item bank condition was assessed using five CAT procedures. While there is a brief description of these five methods below, we direct the reader to the original publications for more detail regarding the computational details for each method. The research employed two unconstrained CAT procedures as controls and extreme scenario comparisons, and three algorithms managing the systematic administration of subsequent items.

Maximum information (MI). Maximum information as used here refers to a CAT with no item exposure control. This means that the examinee is either administered the first item at random or based on some baseline θ value (usually 0.0, as was done for this study). The person then responds to the item and is given another item based on the response given (Weiss, 1984; Weiss, 2008). For maximum likelihood estimation of θ_i, the likelihood function associated with a set of responses y_{ij}, for $j=1,\cdots,n$, is needed. As a practical matter, it is easier to work with the logarithm of the likelihood:

$$L\left(\theta \middle| y_1, \cdots y_n\right) = \sum_{j=1}^{n} \ln\left[P_{j,y_j}\left(\theta\right)\right]. \qquad [4]$$

This function was evaluated at a number of quadrature points for θ, and the value of θ corresponding to the largest value of *L* can be taken as an approximation to $\hat{\theta}$, the maximum likelihood estimate of θ, based on the responses $y_1, \cdots, y_n$. This approximation can, in turn, be refined by taking a new set of quadrature points in the neighborhood of the current maximum and repeating the process. This approach is less susceptible to the problem of local maxima than the alternative approach of taking the derivative of *L* with respect to θ, setting it equal to zero and solving the resulting equation for $\hat{\theta}$ (Wainer, 2000).

The next item to be administered is based on the item information functions of the remaining items in the item bank, evaluated using the hypothetical total category form of each item:

$$I_j\left(\theta\right) = \sum_{k=1}^{K} \frac{1}{P_{jk}\left(\theta\right)} \left[\frac{dP_{jk}\left(\theta\right)}{d\theta}\right]^2 . \qquad (5)$$

This selection leads to the administration of the next item that provides the most information at the present estimated θ level.

Since MI does not control for item exposure, it will always administer those items that provide the most information. This means that several items will almost always be used. In fact, if the original θ estimate is assumed to be 0.0 (before any items are administered) every examinee will be administered the same first item.

Random selection. Random selection is the polar opposite of MI. Though the procedural steps are the same, the random selection method ignores the computation of the maximum information and simply chooses the next item to be administered at random from the remaining items in the item bank. Here, no items ought to be overexposed because by random choice all items have the same probability of being chosen. The exposure rate for each item should be close to the ratio of test length to item bank size. However, a tradeoff arises in the estimation of θ. Since the administered items vary randomly in their discrimination and difficulty, the individual's θ will reflect this randomness. The set of administered items will provide limited information about the examinee's true θ, leading to imprecise estimation of his/her latent trait. The random selection algorithm, therefore, provides another control measure for the best item exposure control possible (but with the significant impairment that it will produce the largest errors in estimation).

Sympson-Hetter conditional method. There exists a class of item exposure control algorithms developed on the basis of the work of Sympson and Hetter (1985; 1997). The goal of these algorithms is to control the maximum item exposure rate (at a value *r*) for a bank of items. In the most basic form of the SH procedure, the item bank is submitted to a pre-simulation before it is administered to the target population. This pre-simulation tracks the number of times an item is administered and how frequently it is selected by the procedure.

Each item in the bank is assigned an initial item exposure parameter, $p_{a|s,i}^{(0)}$. One simple choice is ; $p_{a|s,i}^{(0)} = 1$ another is $p_{a|s,i}^{(0)} = r$ (for the specified target value of *r*). If an item is selected at a given stage for any of these simulated examinees, it is recorded. Next, the simulation should either allow the computer to administer this item, with probability $p_{a|s,i}^{(0)}$ for item *i*, or remove it from the list of available items in the bank for this CAT, just as though it had been administered. After the chosen number of pre-simulations, the relative frequency with which each item has been selected over the simulations is computed. This proportion, $p_{s,i}^{(1)}$ for item *i*, leads to the computation of the adjusted new exposure parameter value, $p_{a|s,i}^{(1)}$, for item *i*. In the following iterations of this procedure, an estimate of the proportion of times item *i* will be administered $\left(p_{a,i}^{(2)}\right)$ is given by

$$\hat{p}_{a,i}^{(2)} = p_{a|s,i}^{(1)} p_{s,i}^{(1)} . \qquad (6)$$

Since we want $p_{a,i}^{(2)} \leq r$ for all *i*, we can at least choose $p_{a|s,i}^{(1)}$ so that $\hat{p}_{a,i}^{(2)} \leq r$ if $p_{s,i}^{(1)} \leq r$. Then we may simply take $p_{a|s,i}^{(1)} = 1$ (i.e., always administer item *i* if the CAT algorithm selects it). If $p_{s,i}^{(1)} > r$, then we may take $p_{a|s,i}^{(1)} = r \big/ p_{s,i}^{(1)}$. This produces the estimate $\hat{p}_{a,i}^{(2)} = r$ for item *i*. The pre-simulations are repeated until the maximum relative frequency with which an item is administered $\left(\max\left\{p_{a,i}^{(k)}\right\}\right)$ stabilizes, hopefully in the neighborhood of the desired maximum, *r*. It should be noted that there is no reason to expect the individual $p_{a|s,i}^{(k)}$ values to stabilize. Therefore, the pre-simulation determination of $p_{a|s,i}^{(k)}$ values is considered successful when the maximum value across all items settles near the desired exposure level.

The conditional version of the SH algorithm (SHC) was considered by Stocking and Lewis (1995a, 1995b). Later Cheng and Twu (1998) extended the earlier concepts to a conditional version in which the maximum exposure rate is controlled for each of a set of θ values (in the case of our pre-simulations, controlled for each quadrature point).

Total rate simplified exposure (TRSE). This method was first proposed by Lewis (2007). The procedure is a simplification of earlier work by Stocking and Lewis (1995a; 1998) that considered the conditional Sympson-Hetter exposure control procedure.

The algorithm supposes that all items in a bank are assigned a single original *r* as their exposure control parameter. The exposure control parameter gives the probability that an item will actually be administered given that it has been selected by the CAT algorithm used to deliver the test (if the item is not administered at that point, it is dropped from the set of available items for that test). Suppose that, at each stage of the test, the CAT algorithm actually gives an ordered list of the remaining items in the bank, in terms of their “attractiveness” to be administered. The first (most attractive) item in the list would then have a probability *r* of actually being administered at this stage. The second item in the list has a probability $r(1-r)$ of being administered at this stage. The third item has probability $r(1-r)^2$ of being administered, and so on. In general, if we let *g* denote the position in the list of the item that is actually administered at this stage of testing, then its probability distribution is given by

$$p(g) = r(1-r)^{g-1}, \text{ for } g = 1, 2, \cdots . \qquad (7)$$

This is known as the Geometric distribution. The problem with this theoretical development is that it implicitly assumes that the item bank is infinitely large. Not only will the bank size $\left(n_p\right)$ be finite, but also we do not want to use all the items in the bank before administering the full set of n_t items. The way the procedure deals with both these issues is to truncate the list at a predetermined constant length. If we take the truncated list length to be $g_{\max} = \text{int}\left(n_p / n_t\right)$, then we are guaranteed to have enough items available to administer a test of length n_t.

The mean of the (untruncated) Geometric distribution is $1/r$ and, more importantly, the cumulative probability distribution function is given by

$$P\left(g\right) = 1 - \left(1 - r\right)^g . \tag{8}$$

This means that, if we truncate the distribution at $g = g_{\max}$, the probability distribution for g becomes

$$p\left(g \middle| g \le g_{\max}\right) = \frac{r\left(1-r\right)^{g-1}}{1-\left(1-r\right)^{g_{\max}}}, \text{ for } g = 1, 2, \cdots, g_{\max} . \tag{9}$$

Also, the truncated cumulative distribution function is given by

$$P\left(g \middle| g \le g_{\max}\right) = \frac{1-\left(1-r\right)^{g}}{1-\left(1-r\right)^{g_{\max}}} . \tag{10}$$

This is what we use when we're converting a uniform random number into a random value from our truncated Geometric distribution. Now the probability that the distribution given in Equation 9 assigns to $g = 1$ is given by

$$\Pr\left(g = 1 \middle| g \le g_{\max}\right) = \frac{r}{1-\left(1-r\right)^{g_{\max}}} > r . \tag{11}$$

Equation 11 says that, after truncation, the maximum item exposure probability must be greater than *r* since, for the first stage of testing, the first item selected by the CAT algorithm will have a probability greater than *r* of being administered. It turns out that the choice of *r* controls the degree of skewness for the truncated distribution. As *r* approaches 0, the distribution of *g* approaches a discrete uniform distribution, with each probability equal to $1/g_{\max}$. As *r* approaches 1, the probability that $g = 1$ approaches 1, and the probabilities for the remaining values of *g* approach 0.

A highly desirable aspect of the TRSE is that it is computationally simple to implement, as opposed to many other algorithms, and is itself adaptive in the sense that, if the originally specified exposure rate (in the case of this project always n_t / n_p.) is fairly low, the procedure will increase the minimum rate to fit the truncated distribution.

Progressive-restricted maximum information (PRMI). Revuelta & Ponsada (1998) proposed a hybrid algorithm that combined two essential components in item exposure control. The procedure has a random component that is weighted against the maximum information approach. Moreover, this is achieved creatively so that at the beginning of the test the items are chosen more at random (the first item being selected entirely at random), and as the test progresses and θ is being estimated the algorithm shifts to administering items based more heavily on the maximum information procedure. In addition, the algorithm restricts global item exposure to a predefined target exposure rate. As tests are being adaptively constructed, the algorithm tracks the administration of each item. If an item has been administered above the target exposure rate, it is removed from the item bank until its relative exposure rate has fallen below the criterion specification.

Analysis

All five methods within the two item bank conditions (140 and 80 items) were compared using three conditionally oriented procedures. The primary reason why a conditional approach was used in the assessment is that the global functioning of an exposure control algorithm only addresses the overall pooled administration rate issue. However, as discussed above, the reality for many constrained item bank (non-cognitive or other) assessments is that they are likely to be administered to groups of individuals with similar latent trait values. Therefore, the functionality of an algorithm needs to be evaluated not only at the global but also at the conditional level. If exposure for an item is held below some criterion globally, that might not inform the test administrators or developers about the fact that, for instance, the item is exposed almost always to the exact same group of people, hence justifying conditional inquiries.

First the amount of bias in the estimation of θ was computed for each method. Bias was defined as the average difference between the estimated θ and the true θ for each individual at each quadrature point:

$$\text{Bias}_i = \frac{\sum_{p=1}^{n} \hat{\theta}_{pi}}{n} - \theta_i = \frac{\sum_{p=1}^{n} \left(\hat{\theta}_{pi} - \theta_i \right)}{n}. \tag{12}$$

Second, the root mean square error (RMSE) of estimation for each method was computed. RMSE is defined as the square root of the average square of the difference between estimated and true θ. RMSE, therefore, combines information about bias and estimation variability:

$$\text{RMSE}_i = \sqrt{\frac{\sum_{p=1}^{n} \left(\hat{\theta}_{pi} - \theta_i \right)^2}{n}}. \tag{13}$$

Third, the averaged asymptotically valid standard error (ASE) estimates of the estimated θ at each quadrature point were compared against the standard deviations (SD) of the estimated θs. The ASE is derived for each examinee by computing the test information function for the test administered to the examinee at the estimated θ for that examinee, and then taking the inverse square root of the value:

$$ASE_p \left(\hat{\theta}_i \right) = \left[\sum_{j=1}^{n_t} I_j \left(\hat{\theta}_{ip} \right) \right]^{-.5}. \tag{14}$$

This provides the asymptotically valid standard error of the estimate (van der Linden & Glas, 2000; Wainer, 2000). The difference between the two statistics can be defined by

$$\mathrm{ASE}_i - \mathrm{SD}_i = \frac{\sum_{p=1}^{n} ASE_p\left(\hat{\theta}_{ip}\right)}{n} - \sqrt{\frac{\sum_{p=1}^{n}\left(\hat{\theta}_{pi} - \bar{\hat{\theta}}_{\cdot i}\right)^2}{n-1}}. \qquad (15)$$

where the first term is the averaged asymptotic standard error for the group of estimates corresponding to the same true θ value and the second term is the standard deviation of the estimated θs for that true value.

Results

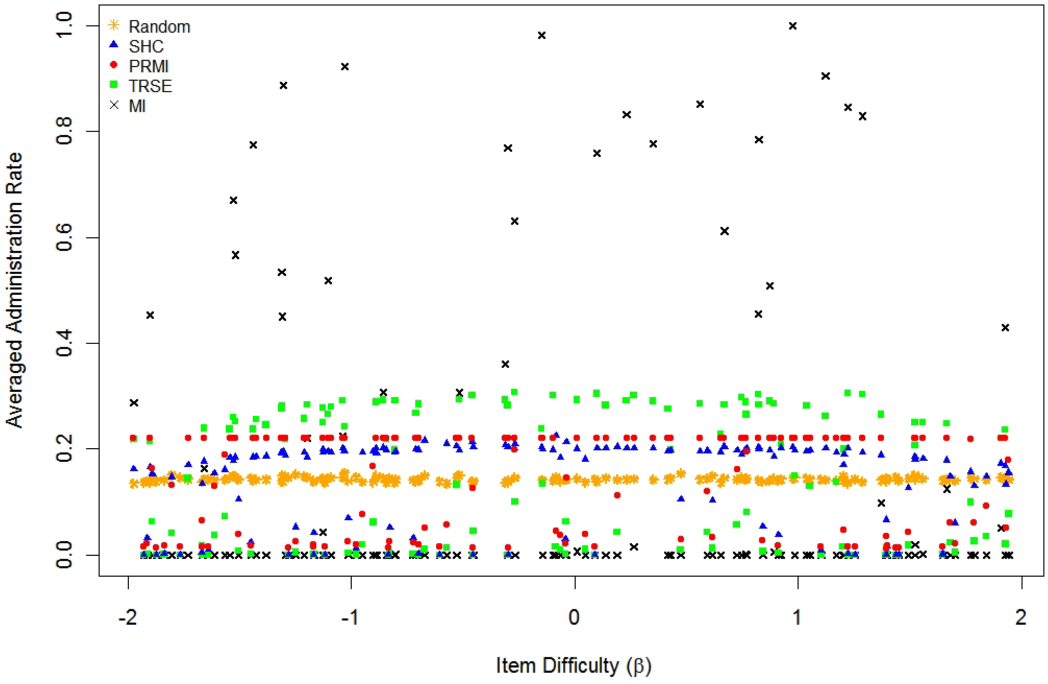
As a general convention in this paper, results for the less restrictive condition (i.e., the 140-item bank) will always be presented first. Figure 3 depicts the item difficulty against the pooled administration rate of all items in the bank, averaged across all true θ quadrature points.

For the random method, an exposure rate equal to the ratio of test length to item bank was expected. Therefore, an average exposure rate of .14 should be observed for items in the 140-item bank condition and .25 in the 80-item bank. Similarly, these two minimum values were used for the TRSE method as starting values for the target exposure rate for both conditions. For the PRMI and SHC, adjusted target exposure control rates were selected. The SHC target rates conditional on θ were controlled at an overall across-item maximum exposure of .22 (based on the S-H pre-simulations) for the 140-item bank condition. Since these results were intended to be comparable with the other methods, a target exposure of .20 was chosen for the PRMI. In the 80-item bank condition, the S-H pre-simulations settled around .40 which was in return used for both the SHC and PRMI target exposure rate in the more restrictive condition.

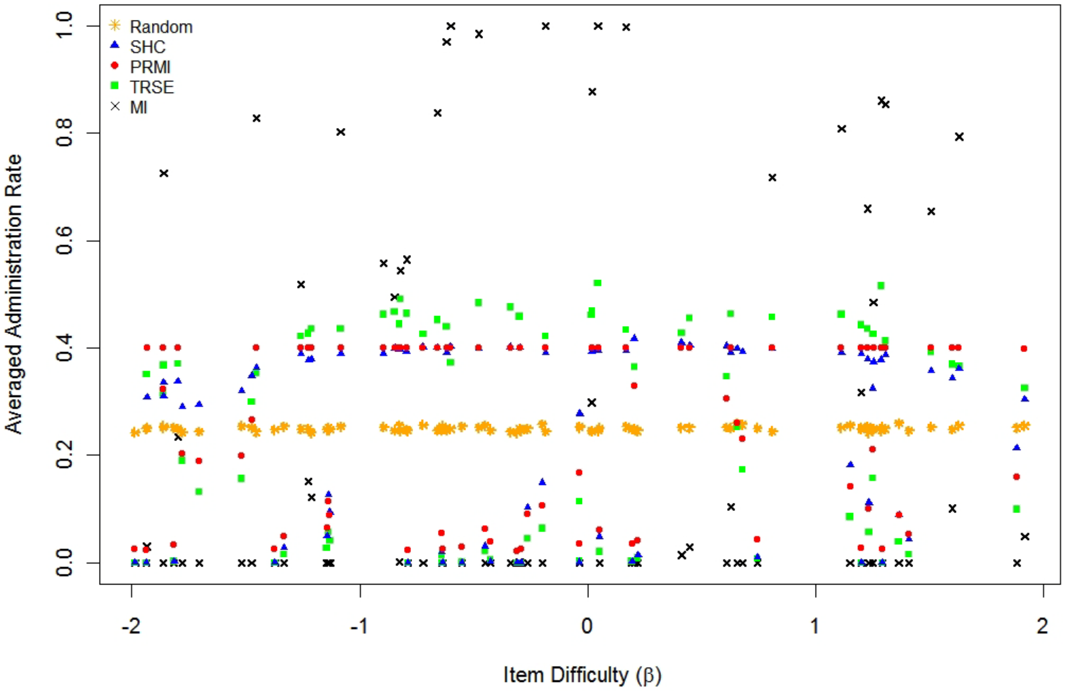
As anticipated, the MI method performed the worst for several items in both the 140- and 80-item banks, administering some items 100% of the time. As was remarked earlier, this is not surprising, since the algorithm will always select the same first item for all individuals. At the other extreme, the random procedure functioned the best in terms of exposure rate, because it randomly administers all items and so administers each item at close to its chance rate.

Figure 3. Administration Rates Pooled Across θ Levels

a. 140 Items



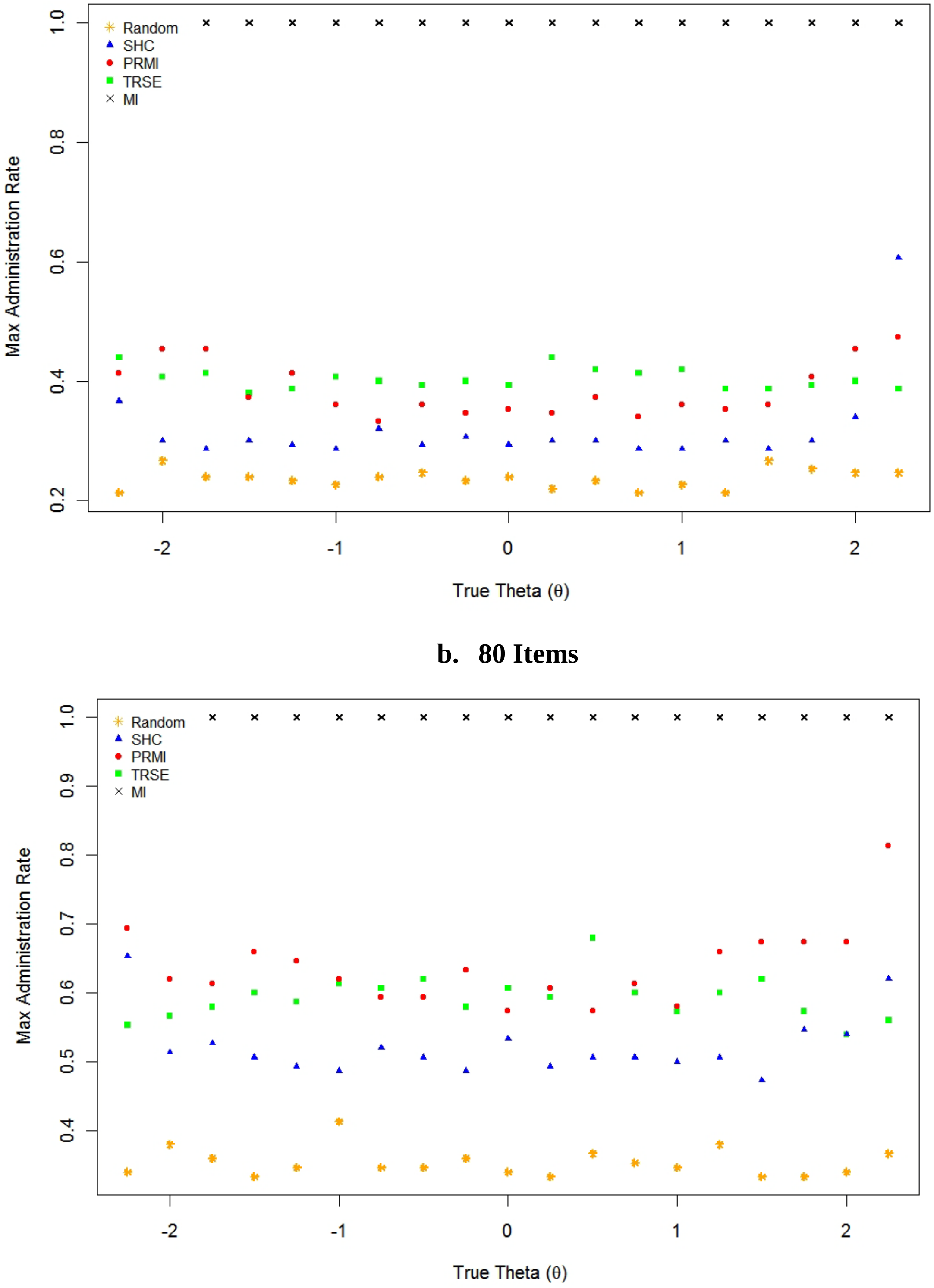
b. 80 Items



The three item exposure control algorithms of interest all performed globally appropriately in that SHC and PRMI did not exceed the target exposure rate to which they were set. TRSE, on the other hand, was not as uniform in its maximal administration resulting in a slightly higher exposure rate, given the nature of the item banks. In terms of item usage, both TRSE and PRMI used all available items in the 80-item bank condition, which is highly desirable when only a limited number of items are available. SHC, on the other hand, failed to utilize the full item bank in both conditions.

Figure 4 shows the maximum administration rate (for any one item from the corresponding item banks) as a function of true θ. Again, MI performed poorest where we can now see that, for each group of examinees, there was always at least one item that was exposed 100% of the time, whereas the random algorithm always performed best in terms of exposure (fluctuating within .10 around the test length-to-item bank size ratio). In the 140-item bank condition, SHC appeared to be controlling conditional exposure best, albeit possibly not as reliably as would be liked given that the maximal exposure for $\theta = 2.25$ was .61, meaningfully higher than its target exposure rate (see Table 1). With more items in the bank, PRMI came in second (after SHC), varying maximum exposure rate around its target rate for each θ group.

Figure 4. Maximum Conditional Exposure Rate



a. 140 Items

b. 80 Items

Table 1. Maximum Conditional Item Exposure Rate

Bank and	True θ										
Method	–2.25	–1.75	–1.25	–.75	–.25	0	.25	.75	1.25	1.75	2.25
140 Items											
MI	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Random	0.2133	0.2400	0.2333	0.2400	0.2333	0.2400	0.2200	0.2133	0.2133	0.2533	0.2467
SHC	0.3667	0.2867	0.2933	0.3200	0.3067	0.2933	0.3000	0.2867	0.3000	0.3000	0.6067
PRMI	0.4133	0.4533	0.4133	0.3333	0.3467	0.3533	0.3467	0.3400	0.3533	0.4067	0.4733
TRSE	0.4400	0.4133	0.3867	0.4000	0.4000	0.3933	0.4400	0.4133	0.3867	0.3933	0.3867
80 Items											
MI	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Random	0.3400	0.3600	0.3467	0.3467	0.3600	0.3400	0.3333	0.3533	0.3800	0.3333	0.3667
SHC	0.6533	0.5267	0.4933	0.5200	0.4867	0.5333	0.4933	0.5067	0.5067	0.5467	0.6200
PRMI	0.6933	0.6133	0.6467	0.5933	0.6333	0.5733	0.6067	0.6133	0.6600	0.6733	0.8133
TRSE	0.5533	0.5800	0.5867	0.6067	0.5800	0.6067	0.5933	0.6000	0.6000	0.5733	0.5600

As the item bank restriction became more severe, PRMI became less effective in its ability to control for item exposure. Since the algorithm does not track administration rates conditionally, it was not expected to perform as well in this evaluation as when the methods were considered globally. SHC still seemed to be providing the best control for exposure, even in the smaller item bank condition, whereas TRSE controlled the item exposure somewhere between SHC and PRMI.

Tables 2 and 3 provide the summary statistics for the methods' bias, RMSE and ASE-SD assessments, conditionally for each true θ. All estimates can be seen plotted in Figures 5 through 7 for all true θ values and the two bank conditions.

In terms of bias, all five methods performed surprisingly similarly, varying within a range of .07. All produced a positive bias in that the average final θ estimate was always higher than the true simulated θ. However, given the relatively small amount of bias this could be deemed negligible.

As for the RMSE, both MI and the random method performed as predicted. MI provided the smallest amount of error in both bank conditions, while the random selection of items produced the largest errors. The three competing algorithms clearly separated in the 140-item bank condition. TRSE performed best, with the lowest amount of RMSE for 18 of the 19 true θ quadrature points. Figure 6a shows that PRMI performed second best and SHC performed the poorest of the three. These relationships became less clear when the item bank became more constrained, as can be seen in Figure 6b, where the three algorithms' RMSEs are pulled closer together. Nonetheless, TRSE still outperformed the other two procedures, producing lowest RMSE values for 15 out of the 19 true θ values.

Table 2. Conditional Analysis Summary of Item Exposure Control Algorithms for the 140-Item Bank Condition

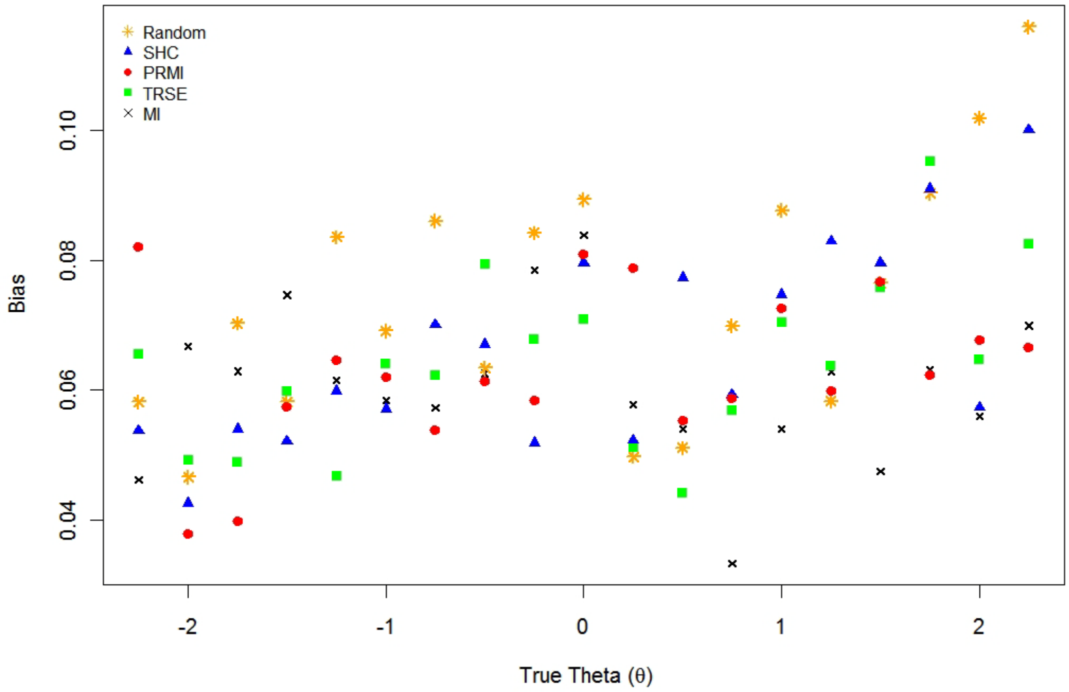
Statistic and	True θ										
Method	–2.25	–1.75	–1.25	–.75	–.25	0	.25	.75	1.25	1.75	2.25
Bias											
MI	0.0462	0.0629	0.0615	0.0573	0.0785	0.0839	0.0577	0.0333	0.0629	0.0631	0.0699
Random	0.0582	0.0703	0.0836	0.0860	0.0842	0.0894	0.0498	0.0699	0.0584	0.0904	0.1160
SHC	0.0538	0.0540	0.0599	0.0700	0.0519	0.0796	0.0523	0.0593	0.0830	0.0910	0.1001
PRMI	0.0821	0.0398	0.0646	0.0539	0.0585	0.0810	0.0788	0.0587	0.0599	0.0624	0.0666
TRSE	0.0656	0.0489	0.0468	0.0623	0.0678	0.0710	0.0512	0.0569	0.0638	0.0952	0.0825
RMSE											
MI	0.3619	0.3683	0.3338	0.3447	0.3745	0.3736	0.3432	0.3405	0.3437	0.3137	0.3743
Random	0.4964	0.4937	0.4538	0.4736	0.4951	0.4843	0.4626	0.5009	0.4672	0.5222	0.5338
SHC	0.4729	0.4479	0.4062	0.4411	0.4500	0.4494	0.4322	0.4418	0.4501	0.4575	0.4650
PRMI	0.4454	0.4257	0.4510	0.4491	0.4191	0.4169	0.4428	0.4312	0.4320	0.4312	0.4496
TRSE	0.4209	0.4104	0.3978	0.3861	0.4185	0.4021	0.4043	0.3879	0.4151	0.4215	0.4416
SD											
MI	0.3602	0.3641	0.3292	0.3411	0.3675	0.3653	0.3394	0.3400	0.3390	0.3083	0.3690
Random	0.5256	0.5176	0.4418	0.5054	0.4807	0.4687	0.4613	0.5152	0.4634	0.5116	0.5056
SHC	0.4799	0.4488	0.3990	0.4226	0.4658	0.4387	0.4559	0.4262	0.4296	0.4445	0.4538
PRMI	0.4384	0.4044	0.4458	0.4621	0.4322	0.4033	0.4483	0.4354	0.4073	0.4039	0.4364
TRSE	0.3986	0.3876	0.3640	0.3742	0.4151	0.3938	0.4054	0.3878	0.3812	0.4020	0.4289
ASE – SD											
MI	-0.1423	-0.1480	-0.1164	-0.1293	-0.1548	-0.1528	-0.1272	-0.1277	-0.1243	-0.0916	-0.1485
Random	-0.1823	-0.1905	-0.1601	-0.1825	-0.2066	-0.1946	-0.1770	-0.2104	-0.1735	-0.2062	-0.1876
SHC	-0.1943	-0.1811	-0.1444	-0.1801	-0.1915	-0.1882	-0.1753	-0.1831	-0.1795	-0.1744	-0.1652
PRMI	-0.1786	-0.1627	-0.1901	-0.1918	-0.1659	-0.1605	-0.1844	-0.1756	-0.1714	-0.1658	-0.1774
TRSE	-0.1628	-0.1644	-0.1568	-0.1466	-0.1787	-0.1608	-0.1660	-0.1482	-0.1687	-0.1585	-0.1695

Table 3. Conditional Analysis Summary of Item Exposure Control Algorithms for the 80-Item Bank Condition

Statistic and	True θ										
Method	–2.25	–1.75	–1.25	–.75	–.25	0	.25	.75	1.25	1.75	2.25
Bias											
MI	0.0204	0.0283	0.0332	0.0111	0.0243	0.0506	0.0589	0.0191	0.0327	0.0609	0.0365
Random	0.0026	0.0126	0.0072	0.0256	0.0242	0.0322	0.0389	0.0627	0.0479	0.0443	0.0697
SHC	-0.0073	0.0232	0.0040	0.0260	0.0135	0.0397	0.0410	0.0209	0.0227	0.0674	0.0559
PRMI	0.0153	0.0243	0.0240	0.0229	0.0457	0.0453	0.0401	0.0364	0.0430	0.0676	0.0474
TRSE	0.0149	0.0236	0.0226	0.0108	0.0290	0.0332	0.0177	0.0425	0.0231	0.0472	0.0338
RMSE											
MI	0.3127	0.3263	0.3212	0.3226	0.3418	0.3462	0.3359	0.3194	0.3601	0.3489	0.3221
Random	0.4424	0.4717	0.4190	0.4377	0.4217	0.4404	0.4188	0.4305	0.4449	0.4463	0.4562
SHC	0.3722	0.3813	0.3790	0.3694	0.3729	0.3740	0.3875	0.3651	0.3772	0.3956	0.4014
PRMI	0.3612	0.3898	0.3610	0.3634	0.3694	0.3767	0.3662	0.3557	0.3640	0.4017	0.3878
TRSE	0.3470	0.3771	0.3446	0.3578	0.3682	0.3620	0.3611	0.3443	0.3726	0.3797	0.3676
SD											
MI	0.3131	0.3262	0.3206	0.3235	0.3421	0.3437	0.3318	0.3199	0.3598	0.3447	0.3211
Random	0.4628	0.4832	0.4081	0.4437	0.4325	0.4180	0.4119	0.4281	0.4331	0.4431	0.4609
SHC	0.3863	0.3786	0.3764	0.3613	0.3749	0.3708	0.3868	0.3690	0.3896	0.3752	0.4042
PRMI	0.3833	0.3916	0.3640	0.3547	0.3750	0.3801	0.3687	0.3502	0.3589	0.4025	0.3699
TRSE	0.3371	0.3808	0.3530	0.3610	0.3671	0.3611	0.3523	0.3381	0.3650	0.3635	0.3646
ASE – SD											
MI	-0.0876	-0.1023	-0.1001	-0.1043	-0.1219	-0.1240	-0.1132	-0.1009	-0.1377	-0.1208	-0.0936
Random	-0.1360	-0.1794	-0.1383	-0.1600	-0.1454	-0.1635	-0.1405	-0.1464	-0.1560	-0.1458	-0.1304
SHC	-0.1103	-0.1282	-0.1344	-0.1271	-0.1311	-0.1304	-0.1441	-0.1227	-0.1291	-0.1336	-0.1280
PRMI	-0.1035	-0.1385	-0.1149	-0.1193	-0.1250	-0.1320	-0.1238	-0.1129	-0.1123	-0.1415	-0.1207
TRSE	-0.0926	-0.1311	-0.1061	-0.1228	-0.1328	-0.1254	-0.1264	-0.1058	-0.1311	-0.1283	-0.1062

Figure 5. Estimation Bias

a. 140-Item Bank



b. 80-Item Bank

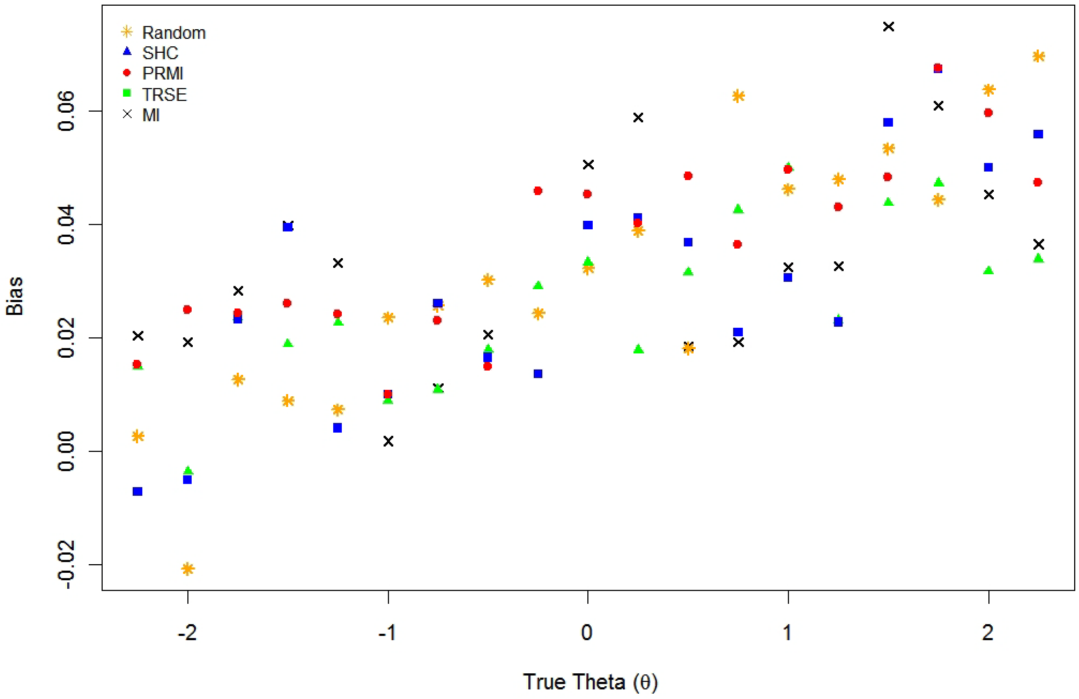
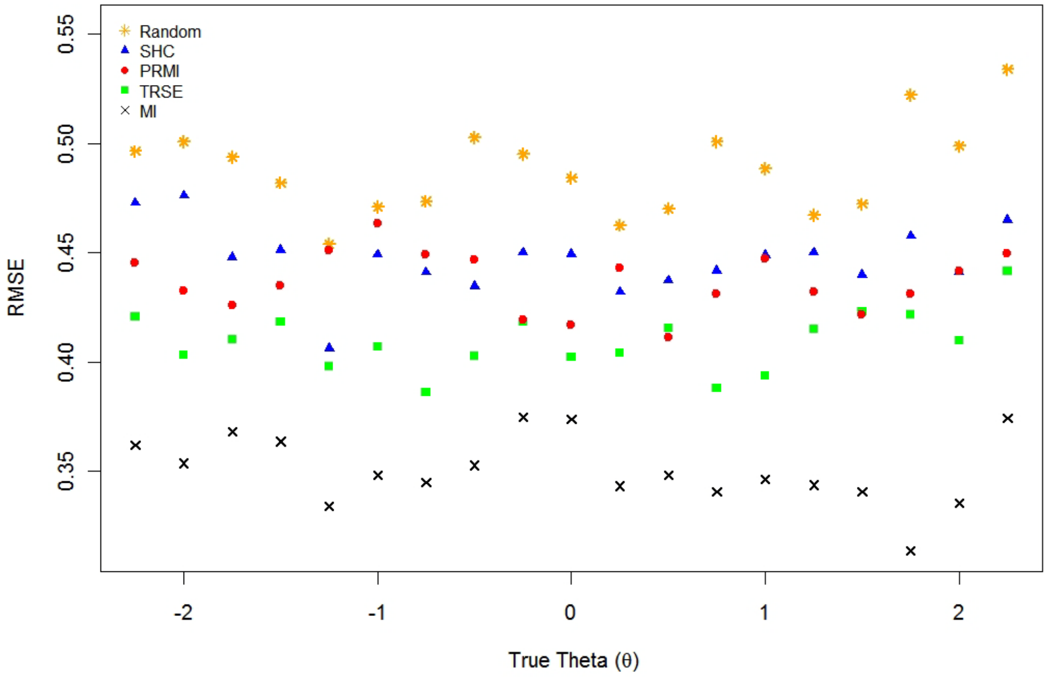


Figure 6. Root Mean Square Error

a. 140-Item Bank



c . 80-Item Bank

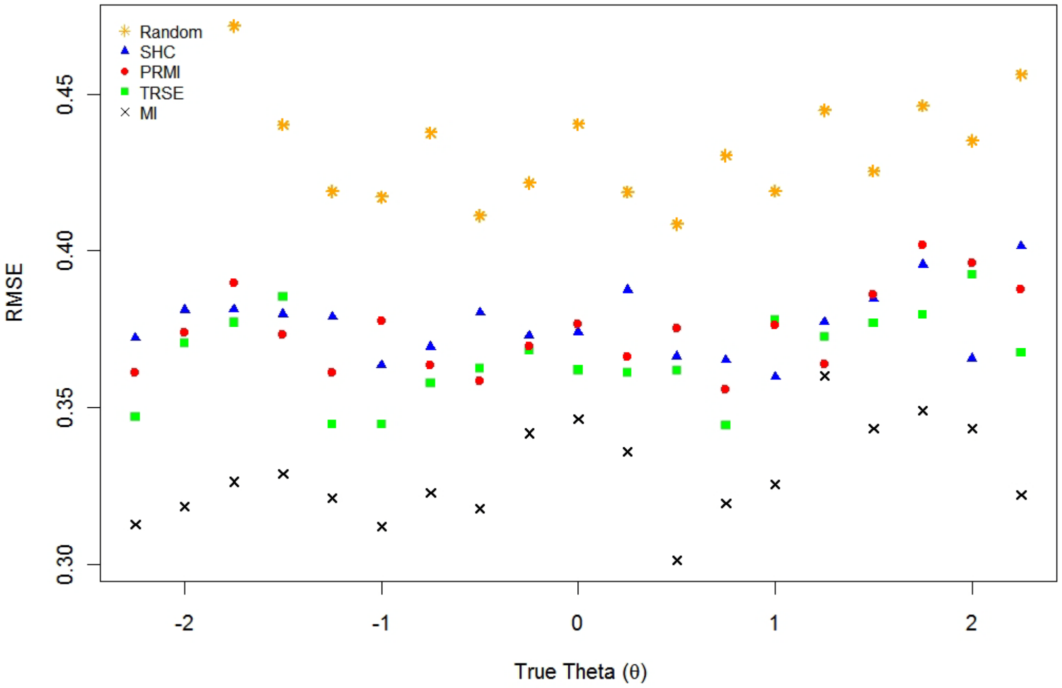
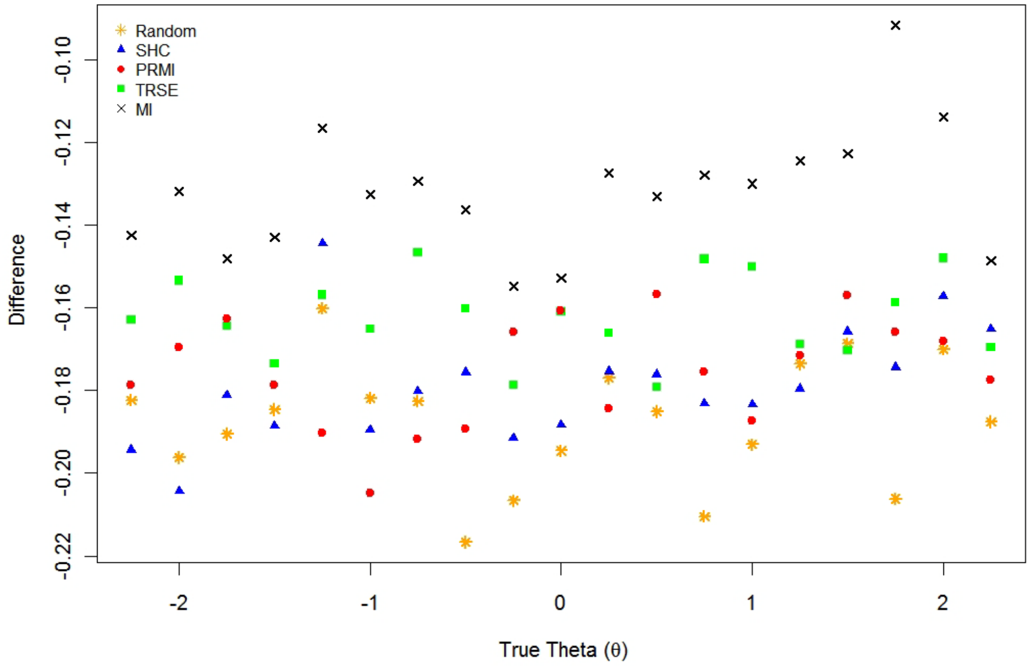
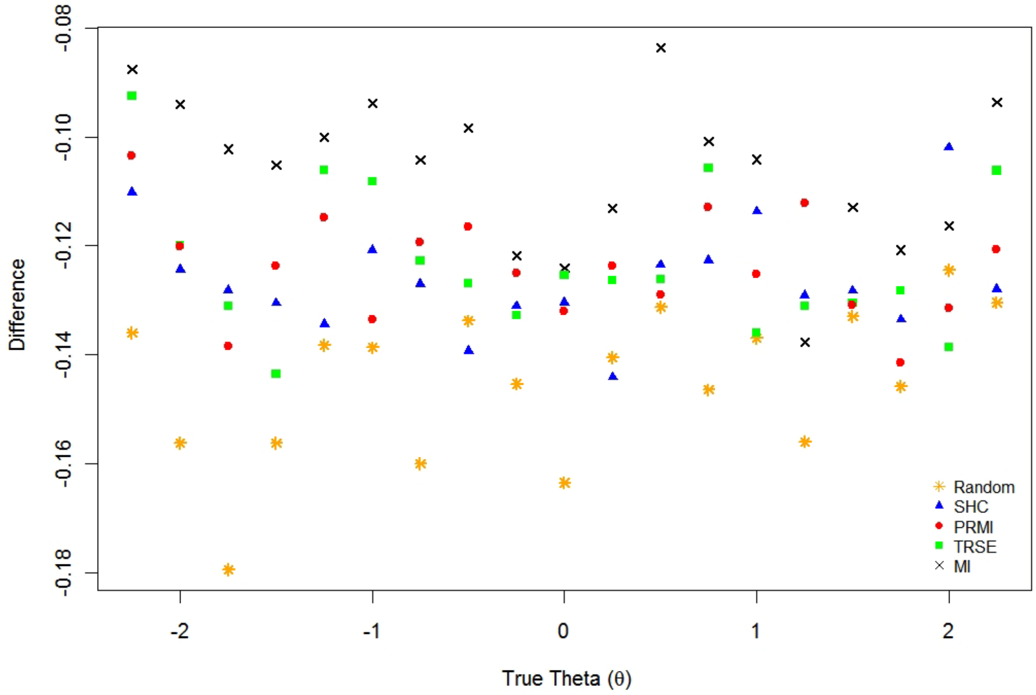


Figure 7. Difference Between Asymptotic Error and Standard Deviation

a. 140-Item Bank



b. 80-Item Bank



Finally, the conditional differences between ASE and θ SDs were compared. In both bank conditions, all algorithms produced a negatively biased estimate in that systematically all asymptotically derived standard errors of the latent trait estimates were smaller than their observed averaged standard deviations. The random method produced the most variable differences between the two statistics.

In the 140-item bank condition, the SHC procedure was the most consistent, albeit after the random selection procedure the method with the second highest discrepancy between estimates. TRSE produced the second least amount of discrepancy after MI. Figures 7a and 7b show that the variations in differences were within a small range of values. Again, as the item bank became more limited the methods no longer separated as clearly.

Conclusions

The three exposure control algorithms considered in this study, SHC, PRMI and TRSE, all demonstrated benefits and shortcomings resulting from their specific computational nature. PRMI did an excellent job in using the items in the bank while controlling the overall exposure, effectively constraining it to the target rate. Conditionally, however, PRMI was more unpredictable and allowed for significantly higher exposure rates for items within specific θ groups.

SHC was the best algorithm of those studied for controlling item exposure effectively while producing acceptable bias. However, SHC had some of the largest RMSEs in both conditions and, together with PRMI, the most deviation between ASE and SD. Further, SHC also appeared to be the most consistent method for controlling exposure rate across different θ values, but is without a doubt the most computationally intensive method. Since all SH procedures need pre-simulations in order to establish adequate exposure control rates, the SHC method is time and labor intensive, requiring large numbers of simulated examinee item responses.

TRSE is without a doubt the simplest method of the three. It also provided the smallest amount of RMSE and deviation between ASE and SD in the 140-item bank, but not the smallest bias across the two conditions or the lowest exposure rate for items overall.

References

Cheng, S., & Twu, B. (1998). *A comparative study of item exposure control methods in computerized adaptive testing*. ACT research Report Series 98-3.

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

Georgiadou, E., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning, and Assessment, 5,* 5-37.

Hetter, R. D. & Sympson, J. B. (1997). *Item exposure control in CAT-ASVAB*. In W. A. Sands, B. K. Waters & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington, DC: American Psychological Association.

Kingsbury, G. G. & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4*, 241-261.

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

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

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17,* 351-363.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, No. 17.*

Samejima, F. (1996). The graded response model. In W. J. van der Linden & Hambleton, R. K. (Eds.), *Handbook of modern item response theory*. New York, NY: Springer.

Stocking, M. L. & Lewis, C. (1995a). *A new method of item exposure control in computerized adaptive testing* (Research Report RR-95-25).Princeton, NJ: Educational Testing Service.

Stocking, M. L. & Lewis, C. (1995b). *Controlling item exposure conditional on ability in computerized adaptive testing* (Research Report RR-95-24). 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,* 57-75.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

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

Sympson, B.J. & Hetter, R.D. (1985). *Controlling item-exposure rates in computerized adaptive testing.* Paper presented at the annual conference of the Military Testing Association, San Diego.

van der Linden, W. J. & Glas, C.A.W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A Primer* (2nd Edition). Mahwah, NJ: Lawrence Erlbaum Associates.

Weiss, D. J. (ed.). (1983). *New horizons in testing: Latent trait theory and computerized adaptive testing*. New York: Academic Press.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21,* 361-375.

Weiss, D. J. (2008). *A global resource for computerized adaptive testing: Research and applications.* Retrieved November 15, 2008, from http://www.psych.umn.edu/psylabs/catcentral/