α -Stratified Computerized Adaptive Testing in the Presence of Calibration Error

Educational and Psychological
Measurement

2015, Vol. 75(2) 260–283

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414530719

epm.sagepub.com

Ying Cheng¹, Jeffrey M. Patton¹, and
Can Shao¹

Abstract

α -Stratified computerized adaptive testing with b -blocking (AST), as an alternative to the widely used maximum Fisher information (MFI) item selection method, can effectively balance item pool usage while providing accurate latent trait estimates in computerized adaptive testing (CAT). However, previous comparisons of these methods have treated item parameter estimates as if they are the true population parameter values. Consequently, capitalization on chance may occur. In this article, we examined the performance of the AST method under more realistic conditions where item parameter *estimates* instead of true parameter values are used in the CAT. Its performance was compared against that of the MFI method when the latter is used in conjunction with Sympson–Hetter or randomesque exposure control. Results indicate that the MFI method, even when combined with exposure control, is susceptible to capitalization on chance. This is particularly true when the calibration sample size is small. On the other hand, AST is more robust to capitalization on chance. Consistent with previous investigations using true item parameter values, AST yields much more balanced item pool usage, with a small loss in the precision of latent trait estimates. The loss is negligible when the test is as long as 40 items.

Keywords

computerized adaptive testing, α -stratification with b -blocking, maximum Fisher information, exposure control, capitalization on chance, item pool usage, measurement precision

¹University of Notre Dame, Notre Dame, IN, USA

Corresponding Author:

Ying Cheng, Department of Psychology, University of Notre Dame, 118 Haggar Hall, Notre Dame, IN 46556, USA.

Email: ycheng4@nd.edu

Among the greatest advantages of computerized adaptive testing (CAT) is efficiency: CAT is often able to achieve measurement goals much more quickly than paper-and-pencil (P&P) tests. Simply by matching item difficulty with examinee ability (θ), computerized adaptive tests can produce accurate ability estimates using fewer items than nonadaptive or linear tests (Weiss, 1982). Widely used CAT item selection algorithms are often more complicated than just matching item difficulty with examinee ability, however. A popular method is to choose items that maximize Fisher information (MFI) evaluated at the current estimate of ability (denoted by $\hat{\theta}$). Because test information (i.e., the sum of item information) is inversely related to the asymptotic variance of the maximum likelihood estimator of θ , this procedure quickly leads to a $\hat{\theta}_{ML}$ with small standard error.

However, the MFI method has some disadvantages. By selecting the “best” items in a narrow statistical sense, some items, usually items with high discrimination parameters, are chosen frequently, whereas others are used rarely or not at all. This can lead to several serious problems. For example, the distributions of item exposure rates can be highly skewed, which has important implications for test security and the costs of item development. Moreover, Chang and Ying (2008) demonstrated that when highly discriminating items are administered early in the test, the final θ estimate can be unduly influenced by the initial responses. In addition, $\hat{\theta}$ is likely to be far from its true value early in the test, so the true value of item information may be far from that evaluated at $\hat{\theta}$ (Chang & Ying, 1999). This means that maximizing item information at $\hat{\theta}$ may be off target. If so, we would be wasting highly discriminating items early in the test. Research has shown that when early misfit occurs (e.g., when able examinees happen to answer the first couple of items incorrectly), CAT may not be able to recover the true θ (Guyer, 2008).

A novel solution to these issues is to employ an *a*-stratified item pool in CAT, first proposed by Chang and Ying (1999) and later modified in Chang, Qian, and Ying (2001). This latter method is called *a*-stratified CAT with *b*-blocking (AST), where *a* and *b* are the item discrimination and difficulty parameters, respectively. The method first rank orders items in the pool according to their *b* parameters. Items with similar *b* parameters are grouped to form a “*b*-block,” and the items within each block are rank-ordered based on their *a* parameters. The low-*a* stratum is formed by pooling the low-*a* items from all *b*-blocks, and the high-*a* stratum is formed by pooling the high-*a* items. This way the item pool can be divided into *K* strata. The strata are ordered by increasing levels of average item discrimination, but the distribution of difficulty is roughly equivalent across strata. Then from the *k*th stratum ($k = 1, 2, \dots, K$), m_k items are administered by minimizing the distance between $\hat{\theta}$ and item difficulty. This way examinees receive low-*a* items at the beginning of the test, when our knowledge of the true θ is minimal, and higher-*a* items are reserved for the later stage of the test. Chang and Ying (1999) demonstrated that in contrast with the MFI method with Sympton-Hetter (S-H) item exposure control (Hetter & Sympton, 1997; Sympton & Hetter, 1985), the prototype *a*-stratification method led to more even item exposure rates and comparable θ recovery. By

forming b -blocks in addition to a -stratification, the AST method successfully takes into account the positive correlation between the discrimination and difficulty parameters in operational item banks, and consequently further improves θ recovery (Chang et al., 2001). Over the years, the method has been acclaimed as a novel method that balances item pool usage and helps counteract the uncertainty of $\hat{\theta}$ in the beginning of the test (Chang & Ying, 2008; Hau & Chang, 2001).

But there is another layer of complexity to the issue: In practice, the true values of item parameters are unknown, and only estimates of the item parameters are available. The optimization scheme in MFI was found to result in an overrepresentation of large calibration errors among the selected items (van der Linden & Glas, 2000). Or as Hambleton, Jones and Rogers (1993) put it, "Among items with generally high discrimination parameters (meaning in terms of the true values), those that are overestimated (i.e., with positive estimation errors) are often the ones selected for a test" (p. 145). This phenomenon is referred to as "capitalization on item calibration error" (van der Linden & Glas, 2000) or "capitalization on chance" (Cheng & Verkuilen, 2008; Olea, Barrada, Abad, Ponsoda, & Cuevas, 2012; Patton, Cheng, Yuan, & Diaio, 2013; Veldkamp, 2012). In particular, the MFI method tends to select items with spuriously high discrimination estimates, resulting in spuriously low standard errors (SEs) for $\hat{\theta}$. van der Linden and Glas (2000) demonstrated this phenomenon in a number of fixed-length CAT simulations where items were selected on the basis of their parameter estimates. As the size of the calibration sample (i.e., the sample used to estimate the item parameters) decreased, the distribution of item exposure rates became more skewed and θ estimation suffered. These effects were exacerbated when the ratio of test length to conditional bank size (i.e., the number of items with difficulty parameters close to the current ability estimate) was small and item exposure control was not employed. Patton et al. (2013) also illustrated how capitalization on chance occurred in variable-length CAT. Among the selected items, test information based on item parameter estimates from a small calibration sample was spuriously high, which translated into spuriously low SEs for $\hat{\theta}$. When the stopping rule of the variable-length CAT was a sufficiently small SE, the test terminated prematurely.

In this study, we would like to examine the performance of AST in the presence of calibration error for the first time. AST uses low- a items in the beginning of the test, and within each stratum, item selection is made with respect to item difficulty alone, so one might expect the effects of capitalization on chance to be mitigated. Additionally, θ recovery and the SE of $\hat{\theta}$ may be robust to the magnitude of item calibration errors.

Thus, the goal of our study is to compare the performance of AST against that based on MFI. In practice, the MFI method is usually coupled with exposure control techniques, for example, the S-H method. Also, the performance of the pure MFI method in the presence of calibration error has been well-studied (van der Linden & Glas, 2000). Therefore, in this study we compare the performance of the AST against two hybrid item selection methods: MFI in conjunction with S-H exposure control and MFI with randomesque exposure control. Various magnitudes of item calibration

error are introduced by manipulating the calibration sample size. The dependent measures of interest are θ recovery and item pool usage. Note that this study is concerned with calibration error that arises from random measurement error instead of systematic error induced by, for example, item position effects (Bowles, Wise, & Kingsbury, 2008; Liu, Zhu, & Guo, 2006), or item parameter drift (Yi & Fan, 2012) due to practice or curriculum change (Han & Guo, 2011) or a test security breach (Guo, 2009). Thus, even though we introduce error into item calibration by manipulating the sample size, the item parameter estimates remain asymptotically unbiased because the error is random. In contrast, item parameter estimates based on the aforementioned systematic errors would (probably) no longer be asymptotically unbiased, and the bias for finite sample sizes would be inflated.

Exposure Control Methods

A simple method of exposure control is called the randomesque method (Kingsbury & Zara, 1989). It is a method that is intuitively easy to understand and straightforward to implement. Rather than selecting the “best” item according to the item selection criterion (such as MFI), this method identifies the d most informative items at the current ability estimate and randomly selects one among them for administration. This way the exposure of the most informative items is reduced, particularly if d is large. Because the most informative items also tend to be those with spuriously large discrimination estimates, this method can be expected to reduce capitalization on chance.

An alternative is the S–H method of exposure control (Simpson & Hetter, 1985). It is “the most popular method of item-exposure control in computerized adaptive testing” (van der Linden, 2003, p. 249). The S–H method conducts a probability experiment to determine whether to administer an item once it is selected. If S_j represents the selection of item j for a randomly sampled examinee and A_j represents the administration of the item, the S–H method forces $A_j \subset S_j$. This implies that $P(A_j) = P(A_j | S_j)P(S_j)$, where $P(A_j | S_j)$ is called the exposure control parameter. For example, if MFI is the item selection criterion, all eligible items in the bank are rank-ordered with respect to information at the current ability estimate. The most informative item is selected, and a number is drawn from a uniform (0, 1) distribution. If this number is smaller than $P(A_j | S_j)$, the item is administered; otherwise, the item is removed from the bank for the remainder of the test, and the next most informative item is considered. If r_j is the desired maximum exposure rate for item j , then setting $P(A_j | S_j) \leq r_j/P(S_j)$ ensures that $P(A_j) \leq r_j$. Estimating the exposure control parameters requires estimates of $P(S_j)$, which can be obtained via simulation; usually, several iterations of the simulation are required before the observed $P(A_j)$ are sufficiently small. In contrast with the randomesque method, the goal of S–H is to keep the maximum exposure rates below a certain threshold. This indirectly forces the administration of items with smaller values of information. Because we expect the most informative items to have spuriously large discrimination estimates, adding

S–H exposure control to the MFI method is expected to substantially reduce capitalization on chance. The randomesque method also forces the administration of items with smaller values of information, though not as effectively as the S–H method. However, neither approach actively increases the exposure of the least informative items in the bank; thus, regardless of method, large numbers of items with exposure rates close to zero are expected. Also, neither method is able to *ensure* that the maximum exposure rate is below a threshold.

Method

We used 520 items from the retired item bank of a large-scale achievement test. The goal of this study is to manipulate the amount of calibration error and examine the performance of the AST method in the presence of varying amounts of calibration error. Two IRT models are considered, the two parameter logistic (2PL) model and the three parameter logistic (3PL) model. Descriptive statistics of the parameter values are shown in Table 1; the distribution of item difficulty is roughly symmetric with a mean of 0.16, and the distribution of discrimination has a slight positive skew. Also, difficulty and discrimination have a medium positive correlation ($r = .45$), which is commonly found in the literature (e.g., Chang et al., 2001). These item parameters served as *true* parameter values and were used to generate item responses. For the 2PL model, the c parameters are set to 0. The true items are denoted as γ , and the item parameter *estimates* as $\hat{\gamma}$.

The 520 items were calibrated with each of three different sample sizes ($N = 2,500, 1,000$, or 500) using BILOG-MG-3 (Zimowski, Muraki, Mislevy, & Bock, 2003). Item calibration was performed with both the 2PL and 3PL models given each sample size. So in total there are six conditions ($2 \text{ models} \times 3 \text{ sample sizes}$), and for each condition we obtained 50 sets of converged item parameter estimates. For each replication, dichotomous item responses under the 2PL or 3PL model were generated with each of three different sample size with true θ s sampled from $N(0, 1)$. For calibration of the 2PL model, no prior is specified. Calibration of the 3PL model typically requires a large sample size (de Ayala, 2009; Jones, Smith, & Talley, 2006); with small sample sizes the estimation may not converge or may lead to unstable estimates of the pseudo-guessing parameter. For calibration of the 3PL, therefore, the default prior (beta(5, 17), with mean of .2) for the c parameters is used. Even with the default prior for the c parameters, not all replications converged for calibration of the 3PL model. It took 55 replications for one condition ($N = 500$) to reach 50 converged set of item parameter estimates. All others conditions needed only 50 replications.

The average root mean squared error (RMSE) of parameter estimates for each calibration sample size under each IRT model (across replications), computed as the square root of the average mean square error, is shown in Table 2. For a given value of N , a given item parameter (discrimination or difficulty), and a given IRT model, the RMSE was computed (see Equation 3 for the computation of RMSE but replace $\hat{\theta}$

Table 1. Descriptive Statistics of True Item Parameter Values.

Item parameter	Mean	Standard deviation	Skewness
<i>a</i>	1.76	0.54	0.58
<i>b</i>	0.16	0.92	−0.15
<i>c</i>	0.19	0.08	0.72

Table 2. Root Mean Squared Error of Item Parameter Estimates (Averaged Across 50 Replications).

<i>N</i>	2PLM		3PLM		
	<i>a</i>	<i>b</i>	<i>a</i>	<i>b</i>	<i>c</i>
2,500	.11	.07	.24	.23	.06
1,000	.17	.12	.37	.28	.07
500	.24	.17	.54	.34	.07

with $\hat{\gamma}$) for each set of parameter estimates. As expected, smaller values of *N* yielded larger calibration errors and thus larger RMSEs.

Each calibrated bank was then employed in three different CAT scenarios. In each scenario, item selection and examinee scoring were conducted with respect to item parameter estimates, whereas responses were generated using the true parameters of the corresponding items. We also included a baseline condition in which the true item parameters were used for item selection, response generation, and scoring. This corresponds to the hypothetical situation in which the examinee sample size is infinite (i.e., $N = \infty$) and the item parameter values are known exactly.

CAT Scenarios

In all CAT simulations in this study, item responses were generated based on the true item parameters, but item selection and ability estimation were performed based on the item parameter estimates. The simulated CATs had a fixed length of 15 or 40 items and employed one of three item selection methods: the AST method, the MFI_SH method, and the MFI_R method. For the AST method, we rank-ordered the bank with respect to estimated difficulty and partitioned the bank into 130 *b* blocks of four items each. The items within each block were then assigned to one of four *a*-strata, from lowest to highest. In this way, each *a*-stratum exhibited a similar distribution of item difficulty. Examinees were administered roughly equal numbers of items from each *a*-stratum, beginning with the *a*-stratum of lowest discrimination and proceeding to the *a*-stratum of highest discrimination. So for the 15-item test,

4 items were administered from each of the first three strata, and 3 items were administered from the last. For the 40-item test, 10 items were administered from each stratum. Within a given a -stratum, an item was selected by minimizing the discrepancy between estimated difficulty and the current $\hat{\theta}$.

For the MFI_SH method, the MFI criterion was paired with S-H exposure control. The maximum exposure rate for every item in the bank was set at $r = 0.2$. For each of the 50 calibrated banks, exposure control parameters were estimated by administering the CAT to 1,000 simulated examinees with θ values drawn from a standard normal density; the final exposure control parameters were obtained after eight iterations of the simulation. For the MFI_R method, the MFI criterion was paired with random-esque exposure control. The 10 most informative items at the current ability estimate were identified, and 1 item of the 10 items was randomly selected for administration.

Regardless of the item selection method, maximum likelihood was used to estimate ability. An initial ability estimate was obtained as follows: the item bank (or the first a -stratum for AST) was ordered with respect to estimated difficulty; the 20 easiest items, 20 most difficult items, and 20 items of medium difficulty were identified; and one item was randomly selected for administration from each group. For those examinees who responded incorrectly or correctly to all three items, a provisional ability estimate of -4 or 4 was used, respectively, until a finite $\hat{\theta}$ could be obtained. After the first three items, items were selected according to the AST, MFI_SH, or MFI_R method.

Dependent Measures

Latent Trait Recovery. To obtain results conditional on specific values of θ , one scoring sample of 200 examinees at each of nine equally spaced θ values between -2.0 and 2.0 in increments of $.50$ was generated. Recall that for each finite value of N , a given bank was calibrated 50 times. To capture the joint effects of measurement error and calibration error at each θ value, each of the 50 calibrated banks was “administered” to a separate scoring sample. The bias and standard error of $\hat{\theta}$ was computed within each group of 200 examinees and then averaged across the 50 replications. That is, for a given replication, bias was estimated for the 200 examinees with a given true θ value using the following equation:

$$\text{Bias}(\hat{\theta}) = \frac{1}{200} \sum_{i=1}^{200} (\hat{\theta}_i - \theta_i), \quad (1)$$

and the 50 bias estimates were averaged. To obtain the empirical standard error, we first computed the sample variance of ability estimates for the 200 examinees with a given true θ value using the following equation:

$$\text{var}(\hat{\theta}) = \frac{1}{199} \sum_{i=1}^{200} (\hat{\theta}_i - \bar{\theta})^2, \quad (2)$$

where $\bar{\theta}$ is the sample mean of the 200 ability estimates. Then the square root of the average of the 50 variance estimates yielded the standard error.

To obtain results that reflect a more realistic population of examinees, an additional scoring sample consisted of 1,000 examinees with θ values drawn from a standard normal density. Again, to capture the joint effects of measurement error and calibration error, each of the 50 calibrated banks was administered to a separate scoring sample. Marginal bias and SE were computed in a way analogous to that described above. We also computed the RMSE. For a given replication, the MSE was computed as follows:

$$MSE(\hat{\theta}) = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{\theta}_i - \theta_i)^2, \quad (3)$$

and the square root of the average of the 50 MSEs was computed. Pearson correlation coefficients between $\hat{\theta}$ and θ were also computed.

Capitalization on Chance. We introduce relative test efficiency as a measure of the degree to which capitalization on chance occurred. Test efficiency (TE) is defined as the average value of (estimated) test information:

$$TE = \frac{1}{n} \sum_{i=1}^n I(\hat{\theta}_i, \hat{\gamma}_i), \quad (4)$$

where $I(\cdot)$ is the test information function, n is the number of examinees, $\hat{\theta}_i$ is the final ability estimate for examinee i , and $\hat{\gamma}_i$ is the vector of item parameter estimates for those items administered to examinee i . For example, if the test for examinee i consists of Items 1 to 15 in the item pool, then $\hat{\gamma}_i$ is a vector with all the discrimination and difficulty parameter estimates of those 15 items. In practice, the inverse of $I(\hat{\theta}_i, \hat{\gamma}_i)$ is often used to approximate the variance of maximum-likelihood-based $\hat{\theta}_i$. Because test information is extremely sensitive to item discrimination, if the item selection method capitalizes on spuriously large discrimination estimates, the value of test information based on item parameter estimates will generally be larger than that based on the corresponding true item parameter values. Thus, we also evaluated Equation 1 with respect to γ_i , the vector of true parameter values for those items administered to examinee i . We then took the ratio of “estimated” to “true” test efficiency, called relative test efficiency (RTE):

$$RTE = \frac{\sum_{i=1}^n I(\hat{\theta}_i, \hat{\gamma}_i)}{\sum_{i=1}^n I(\hat{\theta}_i, \gamma_i)}. \quad (5)$$

If no capitalization on chance occurs, then we expect RTE to be close to 1.0. When RTE deviates substantially from 1, it suggests that capitalization on chance occurs. In the context of CAT, it usually means capitalizing on spuriously

large discrimination parameter estimates, which in turn leads to an RTE that is substantially larger than 1. This was computed using results from the “conditional” scoring sample. That is, we computed RTE within each group of 200 examinees at a given true θ value, and the RTE was averaged across the 50 replications.

Item Pool Usage. To examine item pool usage, data from the “marginal” scoring sample were used. To quantify item pool usage, we used a χ^2 statistic proposed by Chang and Ying (1999), which measures the variability of item exposure rates:

$$\chi^2 = \sum_{j=1}^M \frac{(r_j - \bar{r})^2}{\bar{r}}, \quad (6)$$

where M is the total number of items in the pool (here $M = 520$), r_j is the observed exposure rate of item j , and $\bar{r}$ is the desired (uniform) rate for each item, which is equal to the average item exposure rate (computed as the ratio of test length over the size of the item pool). This statistic was computed for each scoring sample and then averaged across the 50 replications. For each replication, we also computed the proportion of underexposed items (i.e., items with exposure rates less than .02) as well as the proportion of overexposed items (i.e., items with exposure rates greater than .20). These proportions were then averaged across the 50 replications.

Results

Latent Trait Recovery

Conditional Results. Figure 1 displays the conditional bias of $\hat{\theta}$ for each combination of test length and item selection method under the 2PL model. Note that the conditional results of MFI_R are so close to those of MFI_SH that only the latter are presented in Figure 1. This is true for all the figures. And within each panel, there is a separate line for each calibration sample size ($N = \infty, 2,500, 1,000$, or 500). In the top row of Figure 1, where test length = 15, bias exhibits a systematic relationship with N under MFI, regardless of the exposure control method: relative to the $N = \infty$ condition, the average $\hat{\theta}$ is pulled “inward” toward zero as N decreases. The reverse trend is true for the AST method, where $\hat{\theta}$ is pulled “outward” as N decreases. This makes the conditional bias of $\hat{\theta}$ from the AST method noticeably larger in absolute magnitude than that from the MFI_SH method, when the test is short and the sample size is small. The top row (i.e., test length = 15) shows that with both methods, when the true θ are low or high there is clear separation between the sample sizes: generally the larger the calibration sample, the smaller the bias.

In the bottom row of Figure 1, where test length = 40, however, the AST method catches up quickly. The bias under the AST method does not change much with the

Figure 1. Conditional bias of ability estimates for each test length and item selection method (2PLM, conditional results).

calibration sample size N , showing that when the test is long enough, the AST method is robust against small calibration sample size. For the MFI-based methods, on the other hand, the reduction in bias from lengthening the test is not as substantial. With the MFI method, even with a test of 40 items, the separation among different sample sizes is still visible at the low and high ends of θ .

Next, Figure 2 displays the standard error of $\hat{\theta}$ for each combination of test length and item selection method under the 2PL model. The top row of Figure 2 displays the standard error of $\hat{\theta}$ for the test length of 15, while the bottom row displays the same results for the test length of 40. For the shorter test, there is a clear separation of SE curves under the AST method at low and high θ , but not as much for the MFI methods, suggesting again that the AST is less robust against item calibration error when the test is short. Additionally, the SE is in general larger under the AST method

Figure 2. Empirical standard error of ability estimates for each test length and item selection method (2PLM, conditional results).

regardless of the calibration sample size than under MFI-based methods. For the longer test, however, we find that the AST method almost catches up, and there is not any clear separation of the SE curves in the AST panel with respect to different calibration sample size.

Similar to Figures 1 and 2, Figures 3 and 4 presents the bias and standard error of $\hat{\theta}$ for each combination of test length and item selection method under the 3PL model, respectively. In the top row of Figure 3, where test length = 15, the separation between the bias from different sample sizes is more evident than in Figure 1, and the magnitude of bias in each condition seems inflated by the existence of the pseudo-guessing parameter. For example, in the top right of Figure 1, all the curves hug the horizontal axis closely; in the top right of Figure 3, the lines are spread out and the bias from small sample size is noticeably larger than that in Figure 1. When

Figure 3. Conditional bias of ability estimates for each test length and item selection method (3PLM, conditional results).

the test length reaches 40, the bias curves for the AST method and the MFI-based methods look much more similar than when the test length is 15. On the other hand, most of the biases, even though larger than in Figure 1, are still small. The largest bias in absolute magnitude (around .2) occurs at the low end of θ when the test is short under the AST method.

The SE curves in Figure 4 compared to those in Figure 2 are all higher. The separation among the curves from different sample sizes is more conspicuous with the 3PL model. In Figure 2, there is no clear separation when the test length is 40. In Figure 4, the separation is still evident at the low range of θ . When the test is short, the SE of the AST method is higher than that of the MFI-based methods; but when the test length reaches 40, there is hardly any difference.

Figure 4. Empirical standard error of ability estimates for each test length and item selection method (3PLM, conditional results).

Marginal Results. For the results based on the “marginal” scoring sample, the RMSE of $\hat{\theta}$ and correlation between θ and $\hat{\theta}$ under the 2PL model are displayed in Table 3. It is clear that when the test is short, using the AST method leads to lower measurement precision than the MFI-based methods, regardless of the calibration sample size. But the AST method performs almost as well when the test lengthens. With the 40-item test, the correlation between θ and $\hat{\theta}$ is identical up to the second decimal places. Furthermore, regardless of the item selection method and test length, the calibration sample size influences measurement precision: the larger the calibration sample, the higher correlation between the θ and $\hat{\theta}$, and the smaller the RMSE, though the difference is small. The difference is practically negligible when the test is short, and even smaller when the test is longer.

Table 3. Ability Recovery for Each Test Length and Item Selection Method (2PLM, Marginal Results).

Item selection method	N	15 Items		40 Items	
		RMSE	Correlation	RMSE	Correlation
AST	∞	.354	.943	.200	.981
	2,500	.361	.942	.202	.981
	1,000	.366	.942	.208	.980
	500	.388	.939	.216	.980
MFI_SH	∞	.272	.965	.173	.985
	2,500	.275	.964	.176	.985
	1,000	.280	.963	.182	.984
	500	.286	.961	.189	.984
MFI_R	∞	.275	.964	.164	.987
	2,500	.277	.964	.166	.987
	1,000	.282	.962	.171	.986
	500	.286	.960	.178	.985

Table 4. Ability Recovery for Each Test Length and Item Selection Method (3PLM, Marginal Results).

Item selection method	N	15 Items		40 Items	
		RMSE	Correlation	RMSE	Correlation
AST	∞	.470	.900	.256	.969
	2,500	.496	.894	.267	.967
	1,000	.514	.889	.277	.965
	500	.548	.882	.293	.962
MFI_SH	∞	.360	.939	.222	.976
	2,500	.381	.932	.234	.974
	1,000	.391	.928	.245	.972
	500	.406	.921	.254	.970
MFI_R	∞	.365	.937	.208	.979
	2,500	.379	.932	.223	.976
	1,000	.395	.926	.235	.974
	500	.409	.920	.245	.971

Similarly, the RMSE of $\hat{\theta}$ and correlation between θ and $\hat{\theta}$ under the 3PL model are summarized in Table 4. Compared to Table 3, ability recovery is worse under the 3PL model: the RMSE is higher, and the correlation between θ and $\hat{\theta}$ is lower in every condition/cell. But the patterns discussed above about Table 3 still apply here: the AST leads to inferior performance in ability recovery, especially when the test is short. Smaller calibration sample sizes yield larger RMSEs and smaller correlations

between θ and $\hat{\theta}$, regardless of the item selection method and the test length. In addition, between the two MFI-based methods, the difference in ability recovery is small; however, regardless of the calibration sample size, the MFI_SH method does slightly better when the test is short. The trend is reversed with a 40-item test: the MFI_R method performs slightly better. Such reversing is observed in both Tables 3 and 4. This has to do with the trend in item pool usage; we will address this point later.

Capitalization on Chance

Figures 5 and 6 display relative test efficiency for each combination of test length and item selection method under the 2PL and 3PL models, respectively. Again, relative efficiency is the ratio of test information computed from estimated item parameters over that from the true item parameter values. Under MFI_SH (as well as under MFI_R, though these results are not shown as they are essentially very similar to results of MFI_SH), Figure 5 shows that, as expected, test information based on item parameter estimates is spuriously high relative to that based on the true item parameter values—this confirms that capitalization on chance indeed occurred. Furthermore, this overestimation of test information increases as the size of the calibration sample decreases, suggesting that capitalization on chance is more severe when the calibration sample is smaller. This makes sense because when the calibration sample size is small, the resulting item parameter estimates will contain larger errors. In particular, the most informative items (when evaluated with item parameter estimates) have large discrimination estimates that tend to contain positive calibration errors; the size of these errors tends to be larger for smaller samples. As the inverse of test information is used to approximate the variance of $\hat{\theta}_{ML}$, an important implication of this result is that the standard errors of ability estimates tend to be spuriously low under the MFI-based methods, resulting in overconfidence in the precision of test scores. Under AST, in contrast, there is very little capitalization on chance and the effect of N is very small. This is because item selection does not rely as heavily on the value of item discrimination as do the MFI-based methods, and thus we do not expect test information evaluated with respect to item parameter estimates to differ systematically from that based on the true parameter values. Figure 6 shows similar but more conspicuous patterns. The effect of capitalization on chance is more severe under the 3PL model than under the 2PL model, and the effect of sample size is more evident as the curves from different sample sizes are further separated. In summary, smaller calibration sample sizes result in larger effects of capitalization on chance, especially under the 3PL model. The AST method, compared against the MFI-based methods, successfully reduces the effect of capitalization on chance.

Item Pool Usage

Last, item pool usage results under the 2PL model are shown in Table 5 for each combination of test length and item selection method. As expected, for a given test

Figure 5. Relative efficiency for each test length and item selection method (2PLM, conditional results).

length, the AST method achieves more uniform item exposure rates than both MFI-based methods. The advantage is substantial. The effect of calibration sample size on item pool usage, on the other hand, is negligible. Between the two MFI-based methods, the difference is small but an interesting pattern emerges. When the test is short, item pool usage is the worse with the MFI_SH method, indicated by the larger χ^2 values, and higher percentages of overexposed and underexposed items. When the test is long, the MFI-R method leads to the worse item pool usage. So for the 15-item test, the MFI_SH method did not have to reach very far down the list of items (i.e., ordered from most to least informative), whereas the MFI_R method randomly chooses an item from the 10 most informative. For the 40-item test, the MFI_SH method will not consider an item for selection if it has been considered and rejected previously (for a given examinee). Thus, the MFI_SH method ends up reaching far

Figure 6. Relative efficiency for each test length and item selection method (3PLM, conditional results).

down the list of items, whereas the MFI_R method still confines itself to the 10 most informative. Hence, when the test lengthens, the MFI_R method results in worse exposure control due to its inflexibility. This “reversed” trend in terms of exposure control is exactly why a reversed trend is observed in measurement precision. When the test is short, the MFI_SH method does slightly worse than the MFI_R method in terms of exposure control, and therefore slightly better in recovering the ability parameter. When the test lengthens, the MFI_R method does slightly worse in terms of exposure control and slightly better in ability recovery. Equally interesting, as the test gets longer, both the AST and MFI_SH methods have smaller χ^2 values, suggesting that some unused items in the pool are used when the test lengthens. This is an indication that the exposure control is working properly. For the MFI_R, however, the trend is reversed; as test lengthens, the χ^2 increases dramatically, indicating that

the group of highly desirable items are used again and again, and exposure is not spread to less discriminating or less statistically desirable items.

By breaking down the unbalanced item pool usage into overexposure and underexposure, we can easily see that for the MFI-based methods, underexposure is a big issue. The advantage of the AST method is more evident when the test is longer. When the test length is 40, only about 1% of the items are overexposed and only about 1% of the items are underexposed under AST, while under the MFI-based methods, about 14% to 15% of items are overexposed and a staggering 50% to 58% of items are underexposed. This means that over half of the item pool is largely dormant when the MFI-based methods are used. These findings concerning item pool usage are consistent with previous studies (e.g., Yi, 2002) that used true item parameters instead of item parameter estimates.

Similarly, results of item pool usage under the 3PL model are displayed in Table 6 for each combination of test length and item selection method. The same trends as discussed above in Table 5 are observed.

Discussion

Though our statistical training tells us that larger samples are always better, for the calibration of items in some operational testing programs this is simply not an option. Indeed, because the costs associated with adding new items to an existing bank can be quite high, many test developers cut costs by limiting the size of the calibration sample (van der Linden & Pashley, 2010). Thus, it is important to study the effects of calibration sample size on measurement outcomes and to identify test designs that are robust to calibration error. In particular, our goal was to contrast the *a*-stratification method against item selection methods based on the principle of maximizing Fisher information. It has been shown that the MFI-based methods are prone to capitalization on chance (van der Linden & Glas, 2000), meaning that items with high discrimination parameter estimates, which tend to contain positive errors, are favored. Our simulations demonstrate that MFI, even when combined with exposure control techniques such as Symptom–Hetter or randomesque, still tends to capitalize on spuriously large discrimination estimates, leading to spuriously high test information. This effect is exacerbated when the calibration sample size decreases. In contrast, the *a*-stratification method leads to “observed” test information that is much closer to the “true” test information, and the method is quite robust to changes in the calibration sample size. This is one of the most important findings here. Additionally, item pool usage under the *a*-stratification method is much better than that under the MFI-based methods across all conditions.

Previous studies have demonstrated that with the *a*-stratification method, ability recovery tends to suffer a bit relative to the MFI method when item parameters are known. The results here demonstrate that this is also true when the item parameters are estimated, especially when the test is short. When the test gets longer, the *a*-stratification method quickly catches up. With a test of 40 items, the *a*-stratification

Table 5. Item Pool Usage for Each Test Length and Item Selection Method (2PLM, Marginal Results).

Item selection method	N	15 Items			40 Items		
		χ^2	% Underexposed	% Overexposed	χ^2	% Underexposed	% Overexposed
AST	∞	9.0	40.1	0.1	8.0	1.3	0.9
	2,500	9.5	39.0	0.1	8.2	1.1	1.2
	1,000	9.6	39.2	0.1	8.4	1.1	1.3
	500	10.1	40.0	0.1	8.9	1.2	1.5
MFI_SH	∞	64.2	74.8	3.8	52.7	50.4	13.8
	2,500	64.4	75.2	3.8	52.7	50.1	13.9
	1,000	64.6	75.1	4.0	52.7	50.0	14.1
	500	64.4	74.9	3.8	52.3	49.6	14.0
MFI_R	∞	60.8	73.0	3.0	100.0	57.7	15.3
	2,500	61.6	73.1	3.2	100.6	57.4	15.1
	1,000	62.6	73.1	3.3	101.0	57.1	15.1
	500	63.1	73.1	3.5	101.9	56.8	15.0

Note. Underexposed items have exposure rates $<.02$, and over-exposed items have exposure rates $>.20$.

Table 6. Item Pool Usage for Each Test Length and Item Selection Method (3PLM, Marginal Results).

Item selection method	N	15 Items			40 Items		
		χ^2	% Underexposed	% Overexposed	χ^2	% Underexposed	% Overexposed
AST	∞	10.7	39.9	0.3	9.5	1.8	1.2
	2,500	10.7	39.9	0.2	9.4	1.6	1.4
	1,000	11.3	40.6	0.2	9.6	1.7	1.6
MFL_SH	500	11.9	41.4	0.2	10.3	1.8	1.8
	∞	66.0	75.6	4.2	52.7	50.5	13.7
	2,500	65.4	75.3	4.1	51.9	49.7	13.8
MFL_R	1,000	64.9	75.2	4.0	51.8	49.5	13.6
	500	64.0	74.8	3.8	51.5	48.9	13.5
	∞	62.5	74.0	3.5	101.6	57.7	14.6
	2,500	60.5	73.3	3.1	97.2	56.1	14.4
	1,000	58.5	73.0	2.8	94.8	55.7	14.7
	500	56.9	72.5	2.6	92.2	55.0	14.9

Note. Underexposed items have exposure rates $< .02$, and overexposed items have exposure rates $> .20$.

method is only slightly worse in terms of measurement precision (on multiple indices the difference is in the second decimal place) but leads to much better item pool usage. Taken together, we find that the a -stratification method is largely robust to capitalization on chance, and the effect of calibration sample size on its measurement precision and exposure control is very small. With a 40-item test, the loss of measurement precision from the a -stratification method is negligible but its gain in balancing item pool usage is substantial.

The discussion above applies to both the 2PL and 3PL models. Between these two models, the latter is more susceptible to capitalization on chance, and the effect of calibration sample size is more conspicuous. When the test is long, regardless of the IRT model, the a -stratification method manages to achieve much better item pool usage and reduce susceptibility to small calibration sample size, while reaching essentially equivalent level of measurement precision.

It is important to note that the generalizability of our results is somewhat limited. Operational CAT programs often impose nonstatistical constraints on item selection such as content balancing. Content balancing may reduce the risk of capitalization on chance by considering only a subset of the item bank (i.e., those items that correspond to content areas not yet covered by the test). Note that the a -stratification method has been extended successfully to incorporate content balancing (e.g., Cheng, Chang, & Yi, 2007; Yi & Chang, 2003). Therefore, it is definitely possible and worthwhile to compare the a -stratification method when combined with content balancing against the conventional MFI-based algorithms with content control. In addition, there are also many other exposure control techniques available, including the restrictive maximum information method which freezes items that are overexposed (Revuelta & Ponsoda, 1998), the progressive method which adds a random component that receives progressively less weight than the usual information component (Revuelta & Ponsoda, 1998), and the flexible a -stratified method where increasing numbers of items are selected from strata with higher a parameters (Deng, Ansley, & Chang, 2010). Many hybrid strategies have also been proposed, such as the progressive restricted strategy (Revuelta & Ponsoda, 1998), the combination of a -stratification and S-H strategy (Leung, Chang, & Hau, 2002), and a -stratification with freezing (Parshall, Harmes, & Kromrey, 2000). For a review of exposure control strategies developed up to the year 2005, please see Georgiadou, Triantafyllou, and Economides (2007). Since 2005 many new methods have been proposed, for example, the online S-H procedure, where exposure parameters are updated on the fly rather than being obtained through iterative simulations beforehand (Chen, Lei, & Liao, 2008), and the random component method (Barrada, Olea, Ponsoda, & Abad, 2008), which subsumes the progressive method and the proportional method (Segall, 2004) as special cases. It is beyond the scope of the current study to review and compare all available exposure control strategies. The AST method was chosen as the focus of this study because of its pre-alignment and stratification of items based on the error-prone discrimination parameter estimates. It is of interest in future studies

to follow-up with newly developed exposure control strategies and evaluate their performance in the presence of calibration error.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The project is supported by the 2010 Innovation R&D Grant from CTB/McGraw-Hill.

References

- Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness in the Fisher information for improving item-exposure control in CATs. *British Journal of Mathematical and Statistical Psychology*, 61, 493-513.
- Bowles, R., Wise, S., & Kingsbury, G. (2008). *A report on position effects in the NCLEX RN Examination* (Report for the Joint Research Council of the National Council of State Boards of Nursing). Retrieved from www.ncsbn.org/Position_Effects_NCLEX_RN.pdf
- Chang, H.-H., Qian, J., & Ying, Z. (2001). a-Stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, 25, 333-341.
- Chang, H.-H., & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-222.
- Chang, H.-H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika*, 73, 441-450.
- Chen, S. Y., Lei, P. W., & Liao, W. H. (2008). Controlling item exposure and test overlap on the fly in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 61, 471-292.
- Cheng, Y., Chang, H., & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement*, 31, 467-482.
- Cheng, Y., & Verkuilen, J. (2008, March). *Probabilistic item selection in computerized adaptive testing using the maximum entropy criterion*. Paper presented at the annual meeting of the National Council of Educational Measurement, New York City, NY.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.
- Deng, H., Ansley, T., & Chang, H.-H. (2010). Stratified and maximum information item selection procedures in computer adaptive testing. *Journal of Educational Measurement*, 47, 202-226.
- Georgiadou, E., Triantafyllou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment*, 5(8). Retrieved from <http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1647/1482>
- Guyer, R. D. (2008). *Effect of early misfit in computerized adaptive testing on the recovery of theta* (Unpublished doctoral dissertation). University of Minnesota-Twin Cities.

- Guo, F. (2009). *Quantifying impact of compromised items in CAT*. Paper presented at the 2009 meeting of the National Council on Measurement in Education, San Diego, CA.
- Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement*, 30, 143-155.
- Han, K. T., & Guo, F. (2011). *Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing* (GMAC® Research Reports, RR-11-02). Retrieved from <http://www.gmac.com/market-intelligence-and-research/research-library/validity-and-testing/research-reports-validity-related/potential-impact-of-item-parameter-drift-due-to-practice-and-curriculum-change-on-item-calibration.aspx>
- Hau, K.-T., & Chang, H.-H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement*, 38, 249-266.
- Hetter, R. R., & Sympon, J. B. (1997). Item-exposure in CAT-ASVAB. In W. A. Sands, J. R. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington, DC: American Psychological Association.
- Jones, P., Smith, R. W., & Talley, D. (2006). Developing test forms for small-scale achievement testing systems. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 487-525). Mahwah, NJ: Lawrence Erlbaum.
- Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2, 359-375.
- Leung, C. K., Chang, H.-H., & Hau, K.-T. (2002). Item selection in computerized adaptive testing: Improving the a-Stratified design with the Sympon-Hetter algorithm. *Applied Psychological Measurement*, 26, 376-392.
- Liu, M., Zhu, R., & Guo, F. (2006). *The impact of item position effects on ability estimation in CAT: A simulation study* (Law School Admission Council Computerized Testing Report 01-06). Newton, PA: Law School Admission Council.
- Olea, J., Barrada, J. R., Abad, F. J., Ponsoda, V., & Cuevas, L. (2012). Computerized adaptive testing: The capitalization on chance problem. *Spanish Journal of Psychology*, 15, 424-441.
- Parshall, C., Harnes, J. C., & Kromrey, J. D. (2000). Item exposure control in computer-adaptive testing: The use of freezing to augment stratification. *Florida Journal of Educational Research*, 40(1), 28-52.
- Patton, J. M., Cheng, Y., Yuan, K.-H., & Diao, Q. (2013). The influence of item calibration error on variable-length computerized adaptive testing. *Applied Psychological Measurement*, 37, 13-22.
- Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, 35, 311-327.
- Segall, D. O. (2004). A sharing item response theory model for computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 29, 439-460.
- Sympon, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.
- van der Linden, W. J. (2003). Some alternatives to Sympon-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 28, 249-265.
- van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education*, 13, 35-53.

- van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3-30). New York, NY: Springer.
- Veldkamp, B. P. (2012). Ensuring the future of computerized adaptive testing. In *Psychometrics in practice at RCEC* (pp. 39-50). RCEC, Cito/University of Twente, Enschede, Netherlands. Retrieved from <http://doc.utwente.nl/80200/>
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473-492.
- Yi, Q. (2002). *Incorporating the Simpson-Hetter exposure control method into the a-stratified method with content blocking*. Paper presented at the 2002 meeting of the American Educational Research Association, New Orleans, LA. Retrieved from <http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/yi02-01.pdf>
- Yi, Q., & Chang, H. (2003). a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, 56, 359-378.
- Yi, Q., & Fan, M. (2012, April). *Impact of item parameter shifts on score estimates in computerized adaptive testing*. Paper presented at the annual meeting of the National Council of Educational Measurement, Vancouver, Canada.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Skokie, IL: Scientific Software International, Inc.