Optimizing the use of response times for item selection in CAT

Article *in* Journal of Educational and Behavioral Statistics · April 2018

DOI: 10.3102/1076998617723642

CITATIONS
12

READS
67

3 authors:

Edison M. Choe
Graduate Management Admission Council
4 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Justin L. Kern
University of Illinois, Urbana-Champaign
15 PUBLICATIONS **68** CITATIONS
SEE PROFILE

hua-hua Chang
University of Illinois, Urbana-Champaign
93 PUBLICATIONS **2,294** CITATIONS
SEE PROFILE

Article

Journal of Educational and Behavioral Statistics
2018, Vol. 43, No. 2, pp. 135–158
DOI: 10.3102/1076998617723642

Optimizing the Use of Response Times for Item Selection in Computerized Adaptive Testing

Edison M. Choe
Graduate Management Admission Council

Justin L. Kern
University of California, Merced

Hua-Hua Chang
University of Illinois at Urbana-Champaign

Despite common operationalization, measurement efficiency of computerized adaptive testing should not only be assessed in terms of the number of items administered but also the time it takes to complete the test. To this end, a recent study introduced a novel item selection criterion that maximizes Fisher information per unit of expected response time (RT), which was shown to effectively reduce the average completion time for a fixed-length test with minimal decrease in the accuracy of ability estimation. As this method also resulted in extremely unbalanced exposure of items, however, a-*stratification with* b-*blocking was recommended as a means for counterbalancing. Although exceptionally effective in this regard, it comes at substantial costs of attenuating the reduction in average testing time, increasing the variance of testing times, and further decreasing estimation accuracy. Therefore, this article investigated several alternative methods for item exposure control, of which the most promising was a simple modification of maximizing Fisher information per unit of* centered *expected RT. The key advantage of the proposed method is the flexibility in choosing a centering value according to a desired distribution of testing times and level of exposure control. Moreover, the centered expected RT can be exponentially weighted to calibrate the degree of measurement precision. The results of extensive simulations, with item pools and examinees that are both simulated and real, demonstrate that optimally chosen centering and weighting values can markedly reduce the mean and variance of both testing times and test overlap, all without much compromise in estimation accuracy.*

Keywords: *computerized adaptive testing; response time; item selection; item exposure; test overlap*

The primary objective of computerized adaptive testing (CAT) is to efficiently measure an examinee's ability (or any latent trait), where efficiency is by and large conceptualized as the degree of measurement

precision for a given number of items administered. This is generally accomplished by an algorithm that sequentially selects items according to an information-based optimality criterion. Among the various criteria propagated in literature over the decades, the classic maximum Fisher information criterion (maximum information [MI]; Lord, 1980) remains dominant in current practice due to its straightforward implementation and direct link to measurement precision. Specifically, the asymptotic standard error of the maximum likelihood estimate (MLE) of ability is the inverse square root of the cumulative Fisher information of scored items. Therefore, in theory, measurement is most precise when selecting items purely based on maximizing Fisher information.

Nevertheless, despite common operationalization, measurement efficiency of CAT should not only be assessed in terms of the number of items administered but also the time it takes to complete the test. To this end, Fan, Wang, Chang, and Douglas (2012) proposed a novel item selection criterion that maximizes the ratio of Fisher information to expected response time (RT; MI with time [MIT]), which can also be interpreted as information per unit of time. In other words, the MIT algorithm selects the next item in the pool with the highest rate of information for the examinee, thus greatly reducing the average completion time for a fixed-length test with only a marginal decrease in the accuracy of ability estimation. In fact, a recent study found that this simple method results in shorter average test times and fewer RT constraint violations compared to imposing explicit constraints or implementing more complex optimization approaches (Veldkamp, 2016). However, perhaps unsurprisingly, MIT also results in extremely skewed selection of items, since items with both high discrimination and low time intensity are strongly favored. Given that *a*-stratification with *b*-blocking (ASB; Chang, Qian, & Ying, 2001) is a powerful technique for balancing item exposure, a time-weighted version of it (ASB with time [ASBT]) was recommended as a better balanced alternative to MIT. According to results presented later, however, ASBT comes at substantial costs of increasing both the mean and variance of testing times and estimation error relative to MIT.

Therefore, this article investigated the following three alternative techniques for leveraging RTs in item selection: (1) partitioning the item pool into multiple stages according to time intensities and utilizing MIT within each stage, (2) maximizing the ratio of Fisher information to the absolute difference between item time intensity and examinee latent speed, and (3) maximizing the ratio of Fisher information to an optimally centered and weighted expected RT. Extensive simulations, with item pools and examinees that are both simulated and real, were conducted to evaluate the performances of these methods in controlling both item exposure and testing time distribution while maintaining an adequate level of measurement precision.

CAT Framework

The CAT mechanism is predominantly enabled by a class of models within the item response theory (IRT) framework. One of the most frequently employed IRT models in CAT applications measuring a single latent construct with dichotomous items is the univariate three-parameter logistic model (3PLM; Lord & Novick, 1968), generally parameterized as:

$$P(X_{ij} = 1|\theta) = P_j(\theta_i) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta_i - b_j)}}. \quad (1)$$

Note that X_{ij} is a binary random variable mapping the ith examinee's response to the jth item as either 1 for correct or 0 for incorrect, and θ is a latent variable representing ability. Hence, function $P_j(\theta_i)$ computes the probability of correctly responding to item j given the examinee's ability θ_i, where a_j, b_j, and c_j represent the item discrimination, difficulty, and pseudo-guessing parameters, respectively. Also, θ_i and b_j are fixed to be on the same scale.

The conventional MI method of item selection is based on the Fisher information, which for a 3PLM item is given as:

$$I_j(\theta_i) = a_j^2 \left(\frac{1 - P_j(\theta_i)}{P_j(\theta_i)}\right) \left(\frac{P_j(\theta_i) - c_j}{1 - c_j}\right)^2. \quad (2)$$

With the ultimate goal of maximizing cumulative information, the algorithm selects the next item with the largest $I_j(\hat{\theta}_i)$ in the pool, where $\hat{\theta}_i$ is the interim MLE of θ_i based on the examinee's responses to items thus far (Lord, 1980). Specifically, given the set of k observed responses, $\mathbf{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{ik}\}$, the MLE of θ_i is obtained as

$$\hat{\theta}_i^{\mathrm{ML}} = \arg\max_{\theta_i} L(\theta_i|\mathbf{x}_i) = \arg\max_{\theta_i} \prod_{j=1}^{k} P_j(\theta_i)^{x_{ij}} [1 - P_j(\theta_i)]^{1-x_{ij}}, \quad (3)$$

in which $L(\theta_i|x_{ij})$ is the likelihood function of θ_i given observed response x_{ij}:

$$L(\theta_i|x_{ij}) = P_j(\theta_i)^{x_{ij}} [1 - P_j(\theta_i)]^{1-x_{ij}}. \quad (4)$$

However, a major caveat of MLE in practice is that the estimate of θ_i can be highly unstable at the beginning of a test when only a small number of items have been administered. Furthermore, MLE requires at least one correct and one incorrect response to calculate a proper estimate. When all responses are correct, θ_i is estimated to be ∞; likewise, when all responses are incorrect, θ_i is estimated to be $-\infty$. Therefore, a popular alternative that shares none of these particular limitations is a Bayes estimator called *expected a posteriori* (EAP; Bock & Mislevy, 1982), which takes the expected value of the posterior distribution of θ_i given $\mathbf{x}_i$ as follows:

$$\hat{\theta}_i^{\text{EAP}} = E_{\theta_i} f(\theta_i|\mathbf{x}_i) = \int_{\Theta} \theta_i \frac{L(\theta_i|\mathbf{x}_i)g(\theta_i)}{\int_{\Theta} L(\theta_i|\mathbf{x}_i)g(\theta_i)d\theta_i} d\theta_i = \frac{\int_{\Theta} \theta_i L(\theta_i|\mathbf{x}_i)g(\theta_i)d\theta_i}{\int_{\Theta} L(\theta_i|\mathbf{x}_i)g(\theta_i)d\theta_i}. \quad (5)$$

Here, Θ is the latent parameter space of θ_i (i.e., $\theta_i \in \Theta$) and $g(\theta_i)$ is a prior density function of θ_i, typically assumed to be uniform or normal when lacking an empirical prior. Using numerical integration, the EAP estimate of θ_i can be approximated by

$$\hat{\theta}_i^{\text{EAP}} \approx \frac{\sum_{Q} \theta_q L(\theta_q|\mathbf{x}_i)g(\theta_q)}{\sum_{Q} L(\theta_q|\mathbf{x}_i)g(\theta_q)}, \quad (6)$$

where Q is a finite set of quadrature nodes θ_q (i.e., $\theta_q \in Q$).

Numerous simulation studies that have compared these two estimators (e.g., van der Linden & Pashley, 2010; T. Wang & Vispoel, 1998; Weiss, 1982) generally confirm the classic bias–variance trade-off between maximum likelihood and Bayesian estimation: MLE tends to have lower bias but higher standard error, while EAP tends to have higher bias (toward the prior mean) but lower standard error. Nevertheless, differences are practically negligible for moderate test lengths or at least 30 items according to T. Wang and Vispoel (1998). In addition, a fairly common practice is to use a combination of MLE and EAP (van der Linden & Pashley, 2010). For example, EAP could act as a provisional fail-safe if an infeasibility occurs with MLE.

Regardless of the choice between estimators or combinations thereof, MI is generally a well-substantiated item selection criterion in terms of measurement efficiency. In its pure form, however, MI is also notoriously prone to selecting items with high *a* parameters, simply because they have high information (Chang et al., 2001; Chang & Ying, 1999; Hau & Chang, 2001). Although clearly optimal from an efficiency standpoint, this inevitably results in exceedingly unbalanced exposure of items, which is highly undesirable from both resource management and test security perspectives. Consequently, a mechanism for item exposure control is typically implemented when using MI. Georgiadou, Triantafillou, and Economides (2007) provide a fairly comprehensive review of various strategies, of which a few prominent ones include the Sympson–Hetter (SH) method (Hetter & Sympson, 1997; Sympson & Hetter, 1985), modifications of SH (Stocking & Lewis, 1998; van der Linden, 2003), and the so-called randomesque method (Kingsbury & Zara, 1989).

As an alternative to MI, a noteworthy item selection method is the ASB procedure, which is illustrated with a small-scale example in Figure 1. The general setup is as follows: (1) partition the item bank into several even blocks according to the magnitude of *b* values, (2) sort each block according to the magnitude of *a* values, and (3) form strata by grouping items with the same rank order of *a* across the blocks. The rationale behind *b*-blocking is to ensure a

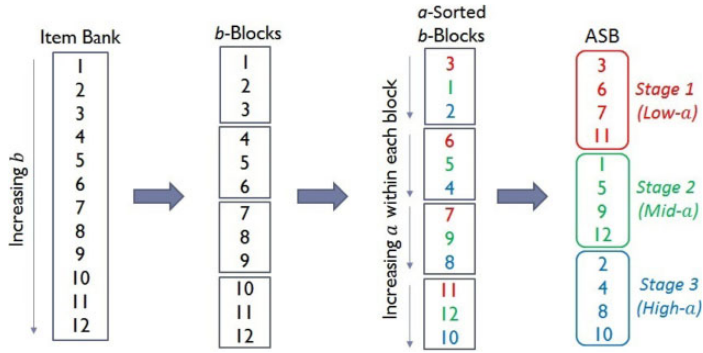

FIGURE 1. *An illustration of the a-stratification with b-blocking process.*

balanced distribution of difficulties in each stratum for item pools that exhibit a correlation between a and b, which should be examined by practitioners (Chang et al., 2001; Chang & van der Linden, 2003; Wingersky & Lord, 1984). The CAT administration is then divided into successive stages, with the best performance yielded by progressing from the lowest a stratum to the highest a stratum (Hau & Chang, 2001). At any given stage during a testing session, the next item chosen is the one that maximizes the b-matching criterion defined as

$$B_j(\hat{\theta}_i) = \frac{1}{|\hat{\theta}_i - b_j|}. \tag{7}$$

In other words, the item with the b parameter closest to $\hat{\theta}_i$ (using estimator of choice) from the current stratum is selected next. Although not as efficient as MI overall, ASB drastically improves exposure balance by drawing items more evenly across the pool. Furthermore, advancing from low to high discrimination items has been shown to curtail the underestimation of examinees who make inadvertent mistakes at the beginning, particularly for short-length tests (Chang & Ying, 2008).

RT Framework

In recent years, there has been a growing interest in using RTs in testing. The immense potential of RTs as a rich source of information is certainly not news, but their practical utility could not be realized until the advent of modern computerized test delivery. These days, test delivery software can now store virtually all examinee by task interactions, including RTs for every item, thus greatly facilitating endeavors to harness them via modeling. Some of the more popular models include the lognormal model (van der Linden, 2006), a generalization of the lognormal called the Box–Cox normal model (Klein Entink, van der Linden,

& Fox, 2009), a flexible semiparametric approach called the Cox proportional hazards model (C. Wang, Fan, Chang, & Douglas, 2013), and a further generalization called the linear transformation model that subsumes the previous three as special cases (C. Wang, Chang, & Douglas, 2013). Each of these RT models was primarily developed as a component in van der Linden's (2007) two-level hierarchical framework for modeling speed and accuracy. The first level consists of separate measurement models for latent speed and accuracy (e.g., lognormal and 3PLM, respectively), and the second level specifies the population and item domain models (i.e., joint distributions of person and item parameters, respectively). Note that the population model relates speed and accuracy across examinees using a covariance parameter. On the other hand, this modeling framework disregards the within-person speed–accuracy trade-off, a particularly robust cognitive phenomenon in reaction time tasks. Unless a test is unduly speeded, a reasonable assumption is made that an examinee operates steadily at his or her innate speed, thereby precluding any speed-induced fluctuations in accuracy (van der Linden, Breithaupt, Chuah, & Zhang, 2007).

Among a variety of RT models, the lognormal is perhaps the most recognized due to its relative simplicity and practicability for typical RT data. While it lacks the flexibility of more complex and general models, it is one of the most straightforward to conceptualize and implement, particularly within the hierarchical framework. Specifically, the lognormal model defines the density function of RT for examinee i on item j (T_{ij}), given the latent speed parameter for the examinee (τ_i), as

$$f(t_{ij}|\tau_i) = \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} e^{-[\alpha_j(\log t_{ij} - \beta_j + \tau_i)]^2/2}, \tag{8}$$

where α_j and β_j are the time discrimination and time intensity parameters for item j, respectively, and β_j and τ_i are fixed to be on the same scale. Rewriting the density function in standard form,

$$f(t_{ij}|\tau_i) = \frac{1}{t_{ij}\sqrt{2\pi(1/\alpha_j)^2}} e^{-[\log t_{ij} - (\beta_j - \tau_i)]^2/[2(1/\alpha_j)^2]}, \tag{9}$$

it becomes clear that $\mu = \beta_j - \tau_i$ and $\sigma^2 = (1/\alpha_j)^2$. Thus, the marginal model can be written as

$$T_{ij}|\tau_i \sim \text{Lognormal}\,[\beta_j - \tau_i,\ 1/\alpha_j^2]. \tag{10}$$

Finally, given that the expected value of a lognormal random variable with log mean μ and log variance σ^2 is $e^{\mu+\sigma^2/2}$, an examinee's expected RT for an item is

$$E(T_{ij}|\tau_i) = e^{\beta_j - \tau_i + 1/(2\alpha_j^2)}. \tag{11}$$

Note that items with low β_j and high α_j have low $E(T_{ij}|\tau_i)$.

Motivation

In efforts to increase measurement efficiency in terms of time, Fan et al. (2012) demonstrated the integration of RT into MI by inversely weighting the Fisher information by the examinee's expected RT for each item. The next item chosen is the one that maximizes the MIT criterion, now formally defined as

$$\mathrm{IT}_j(\hat{\theta}_i, \hat{\tau}_i) = \frac{I_j(\hat{\theta}_i)}{E(T_{ij}|\hat{\tau}_i)}. \tag{12}$$

Here, $\hat{\tau}_i$ is the MLE of τ_i, which is conveniently computed by the closed form expression,

$$\hat{\tau}_i = \frac{\sum_{j=1}^{k} \alpha_j^2(\beta_j - \log t_{ij})}{\sum_{j=1}^{k} \alpha_j^2}, \tag{13}$$

given an examinee's RTs $(t_{i1}, \ldots, t_{ik})$ for the k items administered thus far (van der Linden, 2006). Clearly, MIT favors items with high information and low expected RTs, thus attempting to accomplish the two (possibly competing) tasks of accurately estimating ability while reducing the testing time as much as possible. Although quite successful in this regard, Fan et al. (2012) showed that MIT also results in item exposure that is even more skewed than MI. Hence, they introduced ASBT as a compromise that stratifies the item pool as in ASB and inversely weights the *b*-matching criterion by the expected RT. Specifically, this method selects the next item in the present stratum that maximizes the following criterion:

$$\mathrm{BT}_j(\hat{\theta}_i, \hat{\tau}_i) = \frac{B_j(\hat{\theta}_i)}{E(T_{ij}|\hat{\tau}_i)}, \tag{14}$$

which was shown to balance item exposure rather well, but as shown later, largely by heavily sacrificing the benefits of time weighting in the first place.

General Method

In search of alternatives to MIT or ASBT due to their aforementioned drawbacks, this article investigated the performance of three new RT-informed criteria for item selection in CAT, all under the hierarchical framework with 3PLM and lognormal as the measurement models. In the simulation studies that follow, each of these new methods was directly compared to MIT and ASBT, along with MI as the performance baseline and random (i.e., completely random item selection) as a reference for ideal item pool usage but worst accuracy.

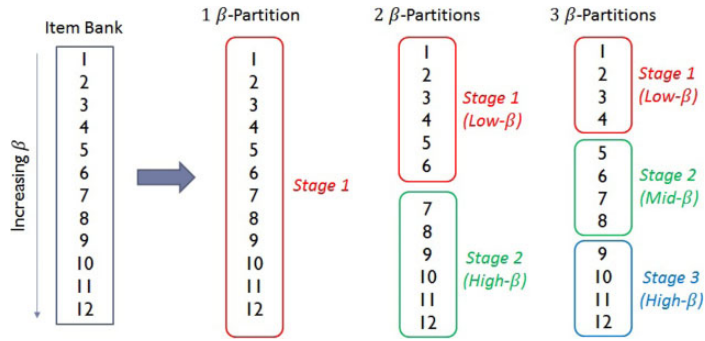

FIGURE 2. *An illustration of the β-partitioning process.*

Proposed Item Selection Procedures

The first method is called β-partitioned MIT (BMIT), in which β-partitioning works analogously to the *b*-blocking procedure in ASB. For a given item pool, the items are sorted according to increasing β values and evenly partitioned into a specified number of stages as illustrated in Figure 2. Items are then selected from each successive stage using MIT, proceeding from the lowest to highest β-partitions. In this way, BMIT forces a more balanced selection of items across the entire range of time intensities as opposed to a normally very biased selection of low-β items.

The second method is called MI with β-matching (MIB), which inversely weights Fisher information by the absolute difference between β_j and $\hat{\tau}_i$ in lieu of $E(T_{ij}|\hat{\tau}_i)$:

$$\mathrm{IB}_j(\hat{\theta}_i, \hat{\tau}_i) = \frac{I_j(\hat{\theta}_i)}{|\beta_j - \hat{\tau}_i|}. \tag{15}$$

This method primarily stems from the hypothesis that, compared to MIT, the item exposure skew could be greatly reduced when examinees are administered items in accordance with their latent speed. Provided that the distributions of β_j and τ_i are similar, matching them as closely as possible would be far less restrictive than perpetually selecting items with the lowest β_j and highest α_j values. Moreover, MIB would have the additional benefit of lower RT variability across examinees compared to MIT. This is because MIB strives to achieve $\beta_j = \tau_i$, in which case the expected RT for item j is reduced to

$$E(T_{ij}) = \mathrm{e}^{1/(2\alpha_j^2)} \tag{16}$$

for any examinee regardless of latent speed.

The third method is Generalized MIT (GMIT), which appends a centering value v and weighting exponent w to the expected RT term:

$$\mathrm{IT}_j^G(\theta_i, \tau_i) = \frac{I_j(\theta_i)}{|E(T_{ij}|\tau_i) - v|^w}, \quad \{v, w\} \in \mathbb{R}^2_{\geq 0}. \tag{17}$$

Note that GMIT reduces to MI for $w = 0$ and MIT for $\{v, w\} = \{0, 1\}$. The rationale of the generalization is as follows: First, maximizing IT_j^G is in part achieved by minimizing $|E(T_{ij}|\tau_i) - v|$, which occurs when $E(T_{ij}|\tau_i) = v$. In the case of MIT where $v = 0$, expected RT of 0 is the unattainable lower bound regardless of τ_i, so the effective item pool is severely confined to a handful of the least time intensive items. This also results in substantial variability of testing times, since much of the same items are being administered to all examinees of varying speeds. In contrast, for a reasonable value of $v > 0$, the RT-optimal items would vary from person to person depending on τ_i, consequently improving item pool usage. This would also stabilize testing times, because every examinee would generally be administered items that take on average v time units to answer. Second, w allows for varying the influence of the centered expected RT in item selection. Presumably, decreasing w would decrease the influence of $|E(T_{ij}|\tau_i) - v|$, thereby improving item exposure balance at the expense of increasing overall testing time. Third, the absolute value of the centered expected RT is taken since it is of no consequence whether the expected RT is lower or higher than v (and taking a noninteger exponent of a negative value may result in a complex number). For the simulation studies presented shortly, the sets of v and w values were limited to $V = \{0.0, 0.1, ..., 3.0\}$ and $W = \{0.50, 0.75, 1.00\}$, respectively, and every $\{v, w\} \in V \times W$ was run (for a total of $|V \times W| = 93$ scenarios).

Evaluation Criteria

The following criteria were used to evaluate the performance of each item selection method given n examinees:

- root mean squared error (RMSE) for estimation accuracy of examinee parameters,

$$\mathrm{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}, \tag{18}$$

$$\mathrm{RMSE}(\hat{\tau}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}_i - \tau_i)^2}; \tag{19}$$

- mean and standard deviation of testing times (tt_i) across examinees as measures of time efficiency and stability,

$$\overline{\text{tt}} = \frac{1}{n}\sum_{i=1}^{n}\text{tt}_i = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\in R_i} t_{ij}, \tag{20}$$

$$s_{\text{tt}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\text{tt}_i - \overline{\text{tt}})^2}, \tag{21}$$

  where R_i is the set of all items administered to examinee i;

- mean and standard deviation of test overlap rates ($\text{tor}_{ii'}$) between all possible pairs of examinees i and i' as measures of test security,

$$\overline{\text{tor}} = \binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{i'=i+1}^{n}\text{tor}_{ii'} = \frac{n}{L(n-1)}\sum_{j=1}^{m}\text{er}_j^2 - \frac{1}{n-1}, \tag{22}$$

$$s_{\text{tor}} = \sqrt{\left[\binom{n}{2} - 1\right]^{-1}\sum_{i=1}^{n-1}\sum_{i'=i+1}^{n}(\text{tor}_{ii'} - \overline{\text{tor}})^2}, \tag{23}$$

  where m is the size of the item pool, L is the fixed test length, $\text{tor}_{ii'}$ is computed as the number of common items between a pair of examinees divided by L, and er_j is the observed exposure rate for item j calculated as the number of times the item was used divided by n.

The less computationally intensive formula for $\overline{\text{tor}}$ using er_j was derived by S.-Y. Chen, Ankenmann, and Spray (2003). Also, C. Wang, Zheng, and Chang (2014) advocated the use of s_{tor} in addition to the traditional $\overline{\text{tor}}$, since it is entirely possible to have low $\overline{\text{tor}}$ overall but very high $\text{tor}_{ii'}$ among a subgroup of examinees. From this perspective, a relatively constant $\text{tor}_{ii'}$ but slightly higher $\overline{\text{tor}}$ is generally preferable to a widely varying $\text{tor}_{ii'}$ but lower $\overline{\text{tor}}$. As lower bound comparisons, when items are selected completely at random, the expected value and standard deviation of $\text{tor}_{ii'}$ are, respectively,

$$\mu_{\text{tor}} = \mu_{\text{er}} = \frac{L}{m}, \quad \sigma_{\text{tor}} = \frac{m-L}{m\sqrt{m-1}}. \tag{24}$$

It is worth noting that Fan et al. (2012) used an alternative indicator of test security,

$$\chi^2 = \sum_{j=1}^{m}\frac{(\text{er}_j - \mu_{\text{er}})^2}{\mu_{\text{er}}}, \tag{25}$$

which measures the skewness of item exposure rates (Chang & Ying, 1999). Although χ^2 and $\overline{\text{tor}}$ are sometimes reported as two distinct statistics that capture

TABLE 1.
Summary of Item Selection Methods and Evaluation Criteria

Item Selection Methods		Evaluation Criteria	
MI	Maximum information	$\text{RMSE}(\hat{\theta})$	Root mean squared error of $\hat{\theta}$
MIT	MI with time	$\text{RMSE}(\hat{\tau})$	Root mean squared error of $\hat{\tau}$
ASB	*a*-stratification with *b*-blocking	$\overline{\text{tt}}$	Mean test time
ASBT	ASB with time	s_{tt}	Standard deviation of test time
MIB	MI with β-matching	$\overline{\text{tor}}$	Mean test overlap rate
BMIT	β-partitioned MIT	s_{tor}	Standard deviation of test overlap rate
GMIT	Generalized MIT		

different aspects of item pool usage (e.g., Y. Cheng, Chang, & Yi, 2007; Deng, Ansley, & Chang, 2010), it can be shown that one is simply a linear transformation of the other as follows:

$$\chi^2 = \frac{m(n-1)}{n}\overline{\text{tor}} + \frac{m}{n} - L. \tag{26}$$

The derivation and implications of this result will be presented in a separate paper currently in preparation. The present article opted to report $\overline{\text{tor}}$, instead of χ^2, for its more intuitive interpretation and wider familiarity.

For easy reference, all item selection methods and evaluation criteria are summarized in Table 1.

Study 1: Simulated Item Pools and Examinees

Method

For this initial study, hundreds of simulations were conducted with a broad range of parameter values in efforts to ensure that the findings are not limited to idiosyncratic data. In the interest of brevity and clarity, just two representative sets of simulated item pools and examinees are presented here to evaluate the item selection criteria under disparate conditions. The first set of data was specified as

Set 1

$$(a_j^*, b_j, \beta_j) \sim \mathcal{N}_2[\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1], \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 0.3 \\ 0.0 \\ 0.0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.10 & 0.15 & 0.00 \\ 0.15 & 1.00 & 0.25 \\ 0.00 & 0.25 & 0.25 \end{bmatrix},$$

where $a_j^* = \log a_j$, meaning a_j has a lognormal distribution;

$$c_j \sim \beta[2,\ 10];$$

$$\alpha_j \sim U[2,\ 4];$$

$$(\theta_i,\ \tau_i) \sim \mathcal{N}_2[\boldsymbol{\mu}_2,\ \boldsymbol{\Sigma}_2],\ \boldsymbol{\mu}_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix},\ \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.00 & 0.25 \\ 0.25 & 0.25 \end{bmatrix},$$

and the second set of data was specified as

Set 2

$$(a_j^*,\ b_j,\ \beta_j) \sim \mathcal{N}_2[\boldsymbol{\mu}_1,\ \boldsymbol{\Sigma}_1],\ \boldsymbol{\mu}_1 = \begin{bmatrix} 0.30 \\ 0.00 \\ -0.25 \end{bmatrix},\ \boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.10 & 0.15 & 0.00 \\ 0.15 & 1.00 & 0.20 \\ 0.00 & 0.20 & 0.16 \end{bmatrix};$$

$$c_j \sim \beta[2, 10];$$

$$\alpha_j \sim U[0.5,\ 2.5];$$

$$(\theta_i,\ \tau_i) \sim \mathcal{N}_2[\boldsymbol{\mu}_2,\ \boldsymbol{\Sigma}_2],\ \boldsymbol{\mu}_2 = \begin{bmatrix} 0.00 \\ 0.25 \end{bmatrix},\ \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.00 & 0.20 \\ 0.20 & 0.16 \end{bmatrix}.$$

Note that there are two key differences between the sets: (1) The marginal distributions of β_j and τ_i are identical in Set 1, whereas they are narrower and shifted apart in Set 2, and (2) the mean of α_j is greater in Set 1. Otherwise, the parameter specifications are equivalent.

For each set, $m = 500$ items and $n = 1{,}000$ examinees were randomly generated, then each examinee's response and RT were simulated for every item. The test length was fixed at $L = 50$ items, with the first item chosen randomly in order to calculate initial estimates of θ_i and τ_i. Estimation was performed with a combination approach, in which EAP was used as an interim substitute whenever MLE failed. For ASBT, the item pool was divided into five strata of 100 items each, then 10 items were selected in each successive stage. For BMIT, the following β-partitions were implemented:

- One β-partition: equivalent to no β-partitioning (i.e., single partition of 500 items).
- Two β-partitions: item pool divided into low and high β-partitions of 250 items each. The first 25 items were selected in the low β stage, then the next 25 items were selected in the high β stage.
- Three β-partitions: item pool divided into low, mid, and high β-partitions with 167, 167, and 166 items, respectively. The first 17 items were selected in the low β stage, the next 17 items were selected in the mid β stage, and then the final 16 items were selected in the high β stage.

Results

Figures 3 and 4 show the results of BMIT and MIB with Sets 1 and 2, respectively. Each of the evaluation criteria is plotted as a function of the number of β-partitions, which only applies to BMIT marked as ○'s. Note that MIT, marked separately as ×, is equivalent to BMIT with one β-partition. All of the other methods are plotted as horizontal lines representing a single value. The following observations can be made on each set of criteria:

(1) *Estimation accuracy*: In terms of RMSE($\hat{\theta}$), BMIT was very close to MI regardless of the number of β-partitions, while MIB was very close to ASBT but still well below Random. RMSE($\hat{\tau}$) (shown in the shaded plot area) was extremely low and essentially equivalent for all methods. There were no discernable differences in relative performance between Sets 1 and 2.

(2) *Mean and standard deviation of testing times*: $\overline{\text{tt}}$ and s_{tt} generally increased for BMIT with more β-partitions, the effect being greater with Set 1 than Set 2. For Set 1, in particular, the distribution of testing times for BMIT became worse than that of ASBT and MI beyond two β-partitions. MIB performed exceptionally well with Set 1, which was second only to MIT in terms of $\overline{\text{tt}}$ and even better than MIT in terms of s_{tt}; on the contrary, MIB performed terribly with Set 2, where both $\overline{\text{tt}}$ and s_{tt} were the worst out of all methods.

(3) *Mean and standard deviation of test overlap rates*: $\overline{\text{tor}}$ generally decreased for BMIT with more β-partitions, while s_{tor} generally remained the same regardless. Even with five β-partitions, BMIT had higher $\overline{\text{tor}}$ and only slightly lower s_{tor} than MI. MIB performed more similarly to ASBT, especially with Set 1 where MIB was nearly identical to ASBT and very close to Random in terms of $\overline{\text{tor}}$.

In summary, BMIT is almost as accurate as MI in terms of estimation but using as few as two or three β-partitions may inordinately increase the mean and standard deviations of testing times while hardly improving the balance of item exposure compared to MIT. On the other hand, MIB is generally similar to ASBT in terms of estimation and better at controlling test overlap rates than BMIT and MI; however, it may counterproductively increase the mean and variance of testing times if the distributions of β and τ are significantly non-overlapping. Therefore, neither BMIT nor MIB proves to be practicable techniques in broader contexts.

Figure 5 shows the results of GMIT with Set 1. The corresponding results with Set 2 were very similar in all respects, so they are not presented here. Each of the evaluation criteria is plotted as a function of v, which only applies to GMIT. Note that MIT, explicitly marked with ×, is equivalent to GMIT at $v = 0$ and $w = 1$. Also note that MIT, ASBT, MI, and Random are all exactly the same as in Figures 3 and 4. This time, the following observations about GMIT can be made on each set of criteria:

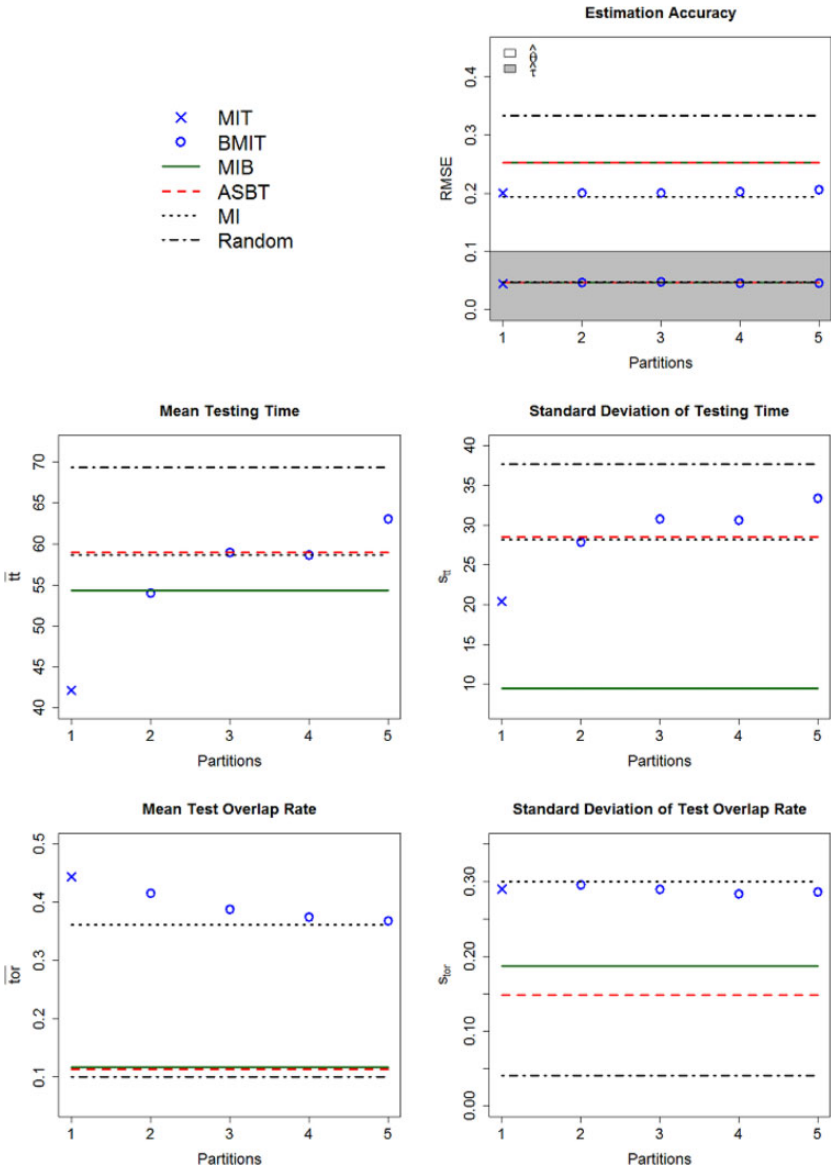

FIGURE 3. *Performances of BMIT and MIB for simulated data: Set 1. The β-partitions only apply to BMIT. BMIT = β-partitioned MIT; MIT = MI with time; MI = maximum information; MIB = MI with β-matching.*

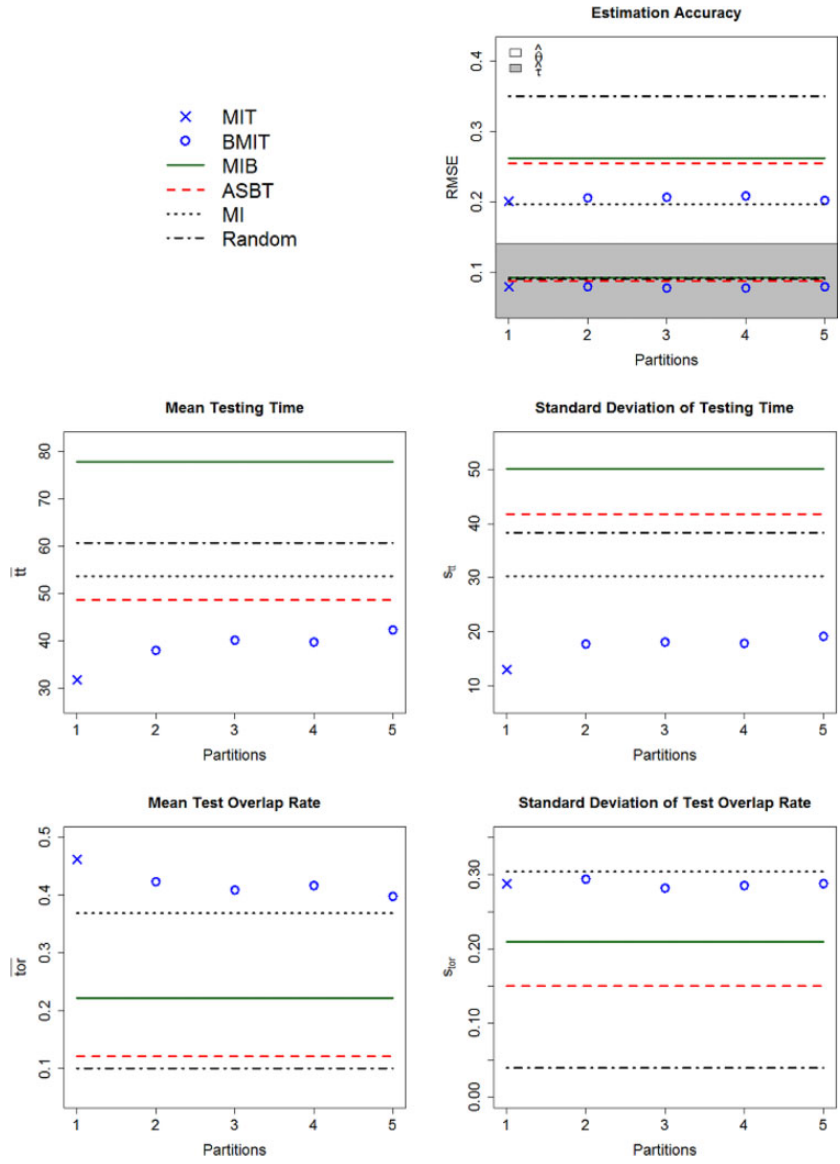

FIGURE 4. *Performances of BMIT and MIB for simulated data: Set 2. The β-partitions only apply to BMIT. BMIT = β-partitioned MIT; MIT = MI with time; MI = maximum information; MIB = MI with β-matching.*

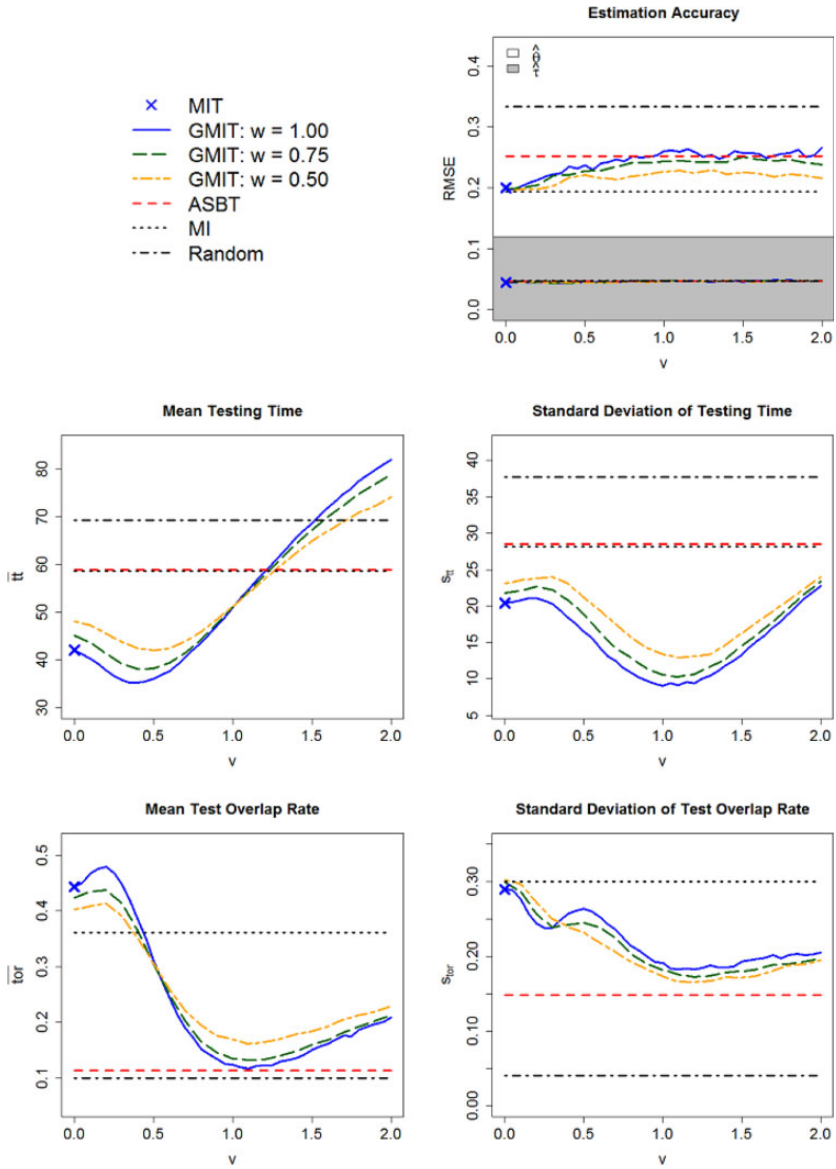

FIGURE 5. *Performances of GMIT for simulated data: Set 1. The centering values* v *only apply to GMIT. GMIT = Generalized MIT; MIT = MI with time; MI = maximum information.*

(1) *Estimation accuracy*: RMSE($\hat{\theta}$) slowly climbed, then leveled out as v increased. For $w = 1$, RMSE($\hat{\theta}$) plateaued around the level of ASBT. At any given v, RMSE ($\hat{\theta}$) was always less for smaller w, eventually reaching the level of MI as w approaches 0. As before, RMSE ($\hat{\tau}$) was extremely low and essentially equivalent for all methods.

(2) *Mean and standard deviation of testing times*: Larger w led to lower $\overline{\text{tt}}$ from $v = 0$ to about 1, at which point $\overline{\text{tt}}$ equalized for all w, then the trend reversed for v beyond 1. On the other hand, larger w always resulted in lower s_{tt} at any v. For any w, $\overline{\text{tt}}$ and s_{tt} were minimized at about $v = 0.3$ and $v = 1.1$, respectively. At these minimum points, GMIT far outperformed all other methods.

(3) *Mean and standard deviation of test overlap rates*: Larger w led to higher $\overline{\text{tor}}$ from $v = 0$ to about 0.5, at which point $\overline{\text{tor}}$ equalized for all w, then the trend reversed for v beyond 0.5. On the other hand, larger w led to lower s_{tor} from $v = 0$ to about 0.3, at which point $\overline{\text{tor}}$ equalized for all w, then the trend reversed for v beyond 0.3. For any w, $\overline{\text{tor}}$ and s_{tor} were both minimized at about $v = 1.1$. At this minimum point, GMIT performed comparably to ASBT.

Several of these observed patterns deserve some elucidation. First, perhaps counterintuitively, $\overline{\text{tt}}$ was minimized and $\overline{\text{tor}}$ was maximized not at $v = 0$ but at about $v = 0.3$, which was the approximate minimum of the expected RT at the median of τ: $\min\left(E[T_j|\text{med}(\tau)]\right) \approx 0.3$. Since no items can have an expected RT of 0, $E(T_{ij}|\tau_i)$ centered at the representative minimum will generally be less than $E(T_{ij}|\tau_i)$ itself, thereby having greater weight in IT_j^G. Second, $\overline{\text{tor}}$ and s_{tor} were minimized at about $v = 1.1$, which was the approximate median of the expected RT at the median of τ: $\text{med}\left(E[T_j|\text{med}(\tau)]\right) \approx 1.1$. A heuristic explanation is that centering the expected RT at its centermost value allows the greatest flexibility in selecting items for examinees at both ends of the τ spectrum, thereby optimizing item pool usage. Third, s_{tt} also happened to be minimized at about $v = 1.1$ for this particular data, but a clear pattern could not be discerned in general. Fourth, w instigated a distinct trade-off between RMSE($\hat{\theta}$) and performance on other criteria, specifically $\overline{\text{tor}}$ for $v > 0.5$ and s_{tt}. Nevertheless, the effects of w were relatively minor compared to the influence of v on general performance. Therefore, the best performer for these data seemed to be GMIT with $v = 1.1$, with the less important choice of w mostly depending on the minimum accuracy or maximum average rate of test overlap deemed acceptable.

Study 2: Real Item Pool and Examinees

Method

To further validate the effectiveness of GMIT, the procedure was next implemented on a set of real data from a high-stakes, large-scale standardized CAT (bestowed by a generous source). The data consisted of raw responses and RTs

from about 2,000 examinees, and the item pool contained about 500 multiple-choice items that were precalibrated according to 3PLM. The lognormal model item parameters (α, β) were estimated using a modified version of van der Linden's (2007) Markov chain Monte Carlo (MCMC) routine that fixed the 3PLM item parameters (a, b, c) to the precalibrated values, and the distribution of τ was set to have a mean of 0. All parameters appeared to converge using 10,000 MCMC draws with a burn-in size of 5,000, and the model seemed to fit well enough for the current application.

For CAT simulation, each examinee's responses and RTs were generated for all items. The test length was fixed at $L = 30$, with the first item chosen randomly in order to calculate initial estimates of θ_i and τ_i. As before, estimation was performed using a combination of MLE and EAP. For ASBT, the item pool was divided into five strata of about 100 items each, then 6 items were selected in each successive stage.

Results

Figure 6 shows the results of GMIT with the real data, which exhibit much of the same patterns as the earlier results with simulated data in Figure 5. First, $\overline{\text{tt}}$ was minimized and $\overline{\text{tor}}$ was maximized at about $v = 0.6$, which was the approximate minimum of the expected RT at the median of τ: $\min(E[T_j|\text{med}(\tau)]) \approx 0.6$. Second, $\overline{\text{tor}}$ and s_{tor} were at their minimum at about $v = 1.8$, which was the approximate median of the expected RT at the median of τ: $\text{med}(E[T_j|\text{med}(\tau)]) \approx 1.8$. Third, s_{tt} was minimized at about $v = 1$. Fourth, the trade-off between $\text{RMSE}(\hat{\theta})$ and performance on other criteria were even less salient than with the simulated data. All things considered, an optimal combination for these real data could be $v = 1.3$ and $w = 0.5$, which afforded better accuracy than ASBT, kept average testing time close to MIT, drastically reduced the variability of testing times to near minimum, and provided a level of item exposure control comparable to ASBT.

Discussion

Continual efforts to refine the item selection algorithm in CAT are not only of scholarly interest but also of paramount importance to operational testing. It goes without saying that accurately measuring ability, saving valuable time and resources, minimizing differential speededness among examinees, and strengthening test security are all critical considerations for most high-stakes administrations. In this spirit, the present investigation sought to improve upon the innovative RT-based item selection methods introduced by Fan et al. (2012). The results of extensive simulations, with both real and simulated data, provide strong evidence for the overall superiority of the proposed GMIT over the other evaluated methods. Ultimately, GMIT with carefully chosen centering and weighting values can appreciably increase the validity of test scores, with negligible detriment to

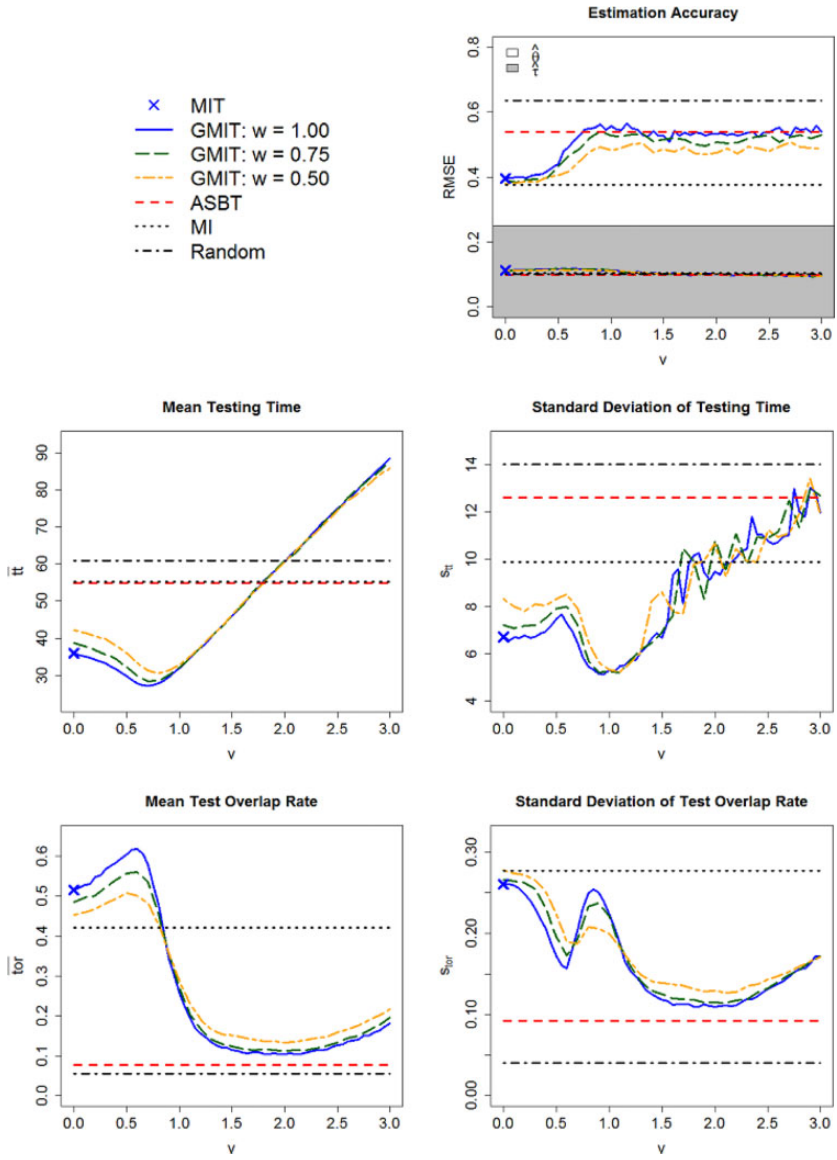

FIGURE 6. *Performances of GMIT for real data. The centering values* v *only apply to GMIT.*

measurement precision, in two distinct aspects: curtailing the likelihood of time pressure–induced rapid guessing by markedly reducing the mean and variance of testing times and decreasing the chances of item preknowledge by dramatically

reducing the mean and variance of test overlap rates. The truly remarkable feature of GMIT is that all of these benefits can be realized without imposing explicit item exposure controls or RT constraints (cf. van der Linden, 1999).

The initialization of GMIT for use in practice requires the following steps: (1) calibrating the item pool with appropriate measurement models for ability and speed given responses and RTs, respectively; (2) generating examinees based on a reasonable or empirically motivated assumption about the joint distribution of ability and speed of the target population; (3) establishing a set of evaluation criteria; (4) conducting a series of CAT simulations with a range of v and w values; and (5) selecting the optimal $\{v, w\}$ according to performance on the evaluation criteria. If performance is evaluated on two or more criteria that involve trade-offs, the "optimal" choice ultimately depends on the minimally acceptable levels on the criteria (e.g., $\overline{\text{tor}} \leq 0.20$) or the user's rational judgment, which can be done via visual inspection of the results as demonstrated.

Alternatively, if a more objective measure is desired to aid in the decision, it is possible to construct an optimality index such as the following:

$$\Omega_{\{v,w\}} = \boldsymbol{\gamma}^T \mathbf{Z}_{\{v,w\}}, \quad \{v, w\} \in V \times W, \tag{27}$$

where $\boldsymbol{\gamma}$ is a vector of weights and $\mathbf{Z}_{\{v,w\}}$ is a vector of standardized values for each evaluation criterion given $\{v, w\}$. Placing all of the criteria on the same scale through standardization is necessary to ensure that the weighted composite is not influenced by the magnitude and spread of the original scales. Provided that lower values indicate better performance for every criterion, the optimal choice would be $\{v, w\}$ that minimizes $\Omega_{\{v,w\}}$, which could be interpreted as a weighted average of the standardized criteria if the values of $\boldsymbol{\gamma}$ are nonnegative and sum to 1. $\boldsymbol{\gamma}$ would be specified according to the importance attributed to each criterion in the overall performance evaluation. As a simple example with the real data results, Table 2 shows an excerpt of rank-ordered $\Omega_{\{v,w\}}$ values computed using weights of $1/6$ for each of the six evaluation criteria. According to this evenly weighted index, $\{v, w\} = \{1.4, 0.75\}$ was the most optimal, whereas the previous choice of $\{v, w\} = \{1.3, 0.50\}$ ranked 10th out of 93. The latter choice placed more emphasis on ability estimation accuracy over the other criteria, but the practical differences between the two choices were relatively slight nonetheless.

Painstaking efforts were taken to assure that the proposed procedure and outcome can be generalized to a broad range of item bank structures and test-taking populations. Although the current investigation was limited to fixed-length CAT with commonly utilized unidimensional 3PLM and lognormal models under the hierarchical framework, the flexibility of GMIT allows for easy implementation and evaluation under a wide variety of schemes. For instance, a recent paper reported success in utilizing the original MIT method in computerized classification testing (CTT) with the sequential probability ratio test (SPRT) stopping rule (Sie, Finkelman, Riley, & Smits, 2015). As a next step, GMIT could be easily tried

TABLE 2.
Average of Standardized Evaluation Criteria, $\Omega_{\{v,w\}}$, for GMIT With Real Data

Rank	$\{v, w\}$	$\Omega_{\{v,w\}}$
1	{1.4, 0.75}	−.4746
2	{1.5, 1.00}	−.4537
3	{1.5, 0.75}	−.4436
4	{1.4, 0.50}	−.4182
5	{1.3, 1.00}	−.4070
6	{1.6, 0.50}	−.4027
7	{1.6, 0.75}	−.3935
8	{1.3, 0.75}	−.3865
9	{1.9, 0.75}	−.3758
10	{1.3, 0.50}	−.3708
⋮	⋮	⋮
93	{3.0, 0.75}	.5055

in the same context with a straightforward modification. Moreover, further scrutiny is certainly warranted to confirm the usefulness of the technique in operational CAT, which is frequently constrained by practical requirements such as content balancing and ordering. This could not be studied at present because the real data at hand did not contain nonstatistical specifications, but there are few compelling reasons to suspect a drastic degradation in GMIT's efficacy under realistic circumstances. Finally, it would be informative to conduct a separate study comparing GMIT to other RT-based methods not considered in this article, including various mathematical optimization approaches (Veldkamp, 2016) and a simplified version of MIT that uses sample-based average log-RTs (in lieu of model-based expected RTs) with randomesque exposure control (Y. Cheng, Diao, & Behrens, 2017).

As a supplemental consideration, although BMIT did not prove to be effective in regard to its originally intended purpose, β-partitioning may have potential in substantive applications. One such possibility could be abating test anxiety caused by perceived speededness. Conceivably, time intensive items at the start of a timed test may elicit subpar performance by those who have not properly "warmed up" and harbor legitimate fears of running out of time. The serious underestimation of ability due to such uncharacteristic errors on initial items is well-documented (Chang & Ying, 2008). By β-partitioning the item pool and selecting items in stages of increasing β, examinees would start off with short items and gradually progress to longer items, which may help allay time-induced anxiety and thus improve test validity. Clearly, empirical studies would need to be conducted to investigate this conjecture.

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) received no financial support for the research, authorship, and/or publication of this article.

References

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444.

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., & van der Linden, W. J. (2003). Optimal stratification of item pools in *a*-stratified computerized adaptive testing. *Applied Psychological Measurement*, *27*, 262–274.

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., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, *40*, 129–145.

Cheng, Y., Chang, H.-H., & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement*, *31*, 467–482.

Cheng, Y., Diao, Q., & Behrens, J. T. (2017). A simplified version of the maximum information per time unit method in computerized adaptive testing. *Behavior Research Methods*, *49*, 502–512.

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.

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37*, 655–670.

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*. Retrieved from http://www.jtla.org

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

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

Klein Entink, R. H., van der Linden, W. J., & Fox, J.-P. (2009). A Box–Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, *62*, 621–640.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Sie, H., Finkelman, M. D., Riley, B., & Smits, N. (2015). Utilizing response times in computerized classification testing. *Applied Psychological Measurement*, *39*, 389–405.

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.

Sympson, J. B., & Hetter, R. D. (1985, October). 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. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*, 195–210.

van der Linden, W. J. (2003). Some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *28*, 249–265.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*, 117–130.

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. (2016). On the issue of item selection in computerized adaptive testing with response times. *Journal of Educational Measurement*, *53*, 212–228.

Wang, C., Chang, H.-H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, *66*, 144–168.

Wang, C., Fan, Z., Chang, H.-H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*, 381–417.

Wang, C., Zheng, Y., & Chang, H.-H. (2014). Does standard deviation matter? Using "standard deviation" to quantify security of multistage testing. *Psychometrika*, *79*, 154–174.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*, 109–135.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 473–492.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, *8*, 347–364.

Authors

EDISON M. CHOE is an associate psychometrician at Graduate Management Admission Council (GMAC), 11921 Freedom Drive, Suite 300, Reston, VA 20190; email: echoe@gmac.com. He was a graduate student at the University of Illinois at Urbana-Champaign when this article was completed. His primary research interests include computerized adaptive testing and psychometric issues concerning test security.

JUSTIN L. KERN is a visiting assistant professor at the University of California, Merced, 5200 North Lake Road, Merced, CA 95343; email: jkern4@ucmerced.edu. His research interests include psychological and educational measurement, statistical modeling in the social sciences, multivariate analysis techniques, and computerized adaptive testing.

HUA-HUA CHANG is a professor of psychology at the University of Illinois at Urbana-Champaign, 603 East Daniel Street, Champaign, IL 61820; email: hhchang@illinois.edu. His primary research interests focus on improving large-scale educational assessments, specifically on issues of test reliability and validity. One of his major contributions is solving both theoretical and practical issues of computerized adaptive testing.

Manuscript received July 24, 2015
First revision received October 26, 2016
Second revision received April 14, 2017
Accepted June 13, 2017