Behav Res (2017) 49:502–512
DOI 10.3758/s13428-016-0712-6

A simplified version of the maximum information per time unit method in computerized adaptive testing

Ying Cheng[1] · **Qi Diao**[2] · **John T. Behrens**[1,3]

Published online: 23 February 2016

Abstract In this article, we propose a simplified version of the maximum information per time unit method (MIT; Fan, Wang, Chang, & Douglas, *Journal of Educational and Behavioral Statistics 37*: 655–670, 2012), or MIT-S, for computerized adaptive testing. Unlike the original MIT method, the proposed MIT-S method does not require fitting a response time model to the individual-level response time data. It is also computationally efficient. The performance of the MIT-S method was compared against that of the maximum information (MI) method in terms of measurement precision, testing time saving, and item pool usage under various item response theory (IRT) models. The results indicated that when the underlying IRT model is the two- or three-parameter logistic model, the MIT-S method maintains measurement precision and saves testing time. It performs similarly to the MI method in exposure control; both result in highly skewed item exposure distributions, due to heavy reliance on the highly discriminating items. If the underlying model is the one-parameter logistic (1PL) model, the MIT-S method maintains the measurement precision and saves a considerable amount of testing time. However, its heavy reliance on time-saving items leads to a highly skewed item exposure distribution. This weakness can be ameliorated by using randomesque exposure control, which successfully balances the item pool usage. Overall, the MIT-S method with randomesque exposure control is recommended for achieving better testing efficiency while maintaining measurement precision and balanced item pool usage when the underlying IRT model is 1PL.

Keywords Maximum information per time unit · Response time · Computerized adaptive testing · Item exposure control · Test efficiency

Tailored testing or adaptive testing is known for its "efficiency" over traditional linear testing. The idea is to find the most suitable items from a large bank for each examinee. In the simplest case of educational testing, highly capable test takers should not be asked many easy questions and struggling test takers should not be presented with too many difficult questions. Delivering questions that are too easy may cause boredom while delivering items that are too difficult may cause anxiety or other forms of construct irrelevant meta-cognitive activity. In addition, responses to items that are too easy or too hard provide little information from the perspective of testing efficiency and are not helpful in quickly zeroing in on an examinee's ability.

Adaptive testing seeks to avoid these difficulties by delivering assessment tasks that are tailored to each examinee's ability. When the maximum information (MI) method is used for item selection (Weiss, 1982) adaptive testing can achieve the same level of measurement precision with as few as half the number of items required of linear tests. Following this approach, the ability estimate of an examinee is updated every time he or she responds to a question (Lord, 1980). The MI method then identifies the most informative item in the bank for a particular examinee with certain ability, best approximated by the most recent ability estimate—and administers that item to the examinee. By applying this method at every step, the MI method asymptotically leads to the largest test

✉ Ying Cheng
ycheng4@nd.edu

[1] Department of Psychology, University of Notre Dame, 118 Haggar Hall, Notre Dame, IN 46556, USA

[2] Pacific Metrics Corporation, Monterey, CA, USA

[3] Advanced Computing and Data Science Lab, Pearson, South Bend, IN, USA

information possible given the number of items administered to each examinee

Although it is commonplace to conceptualize test length in terms of the number of items delivered, in practice the length of time involved with a test is often of equal or greater importance. Wainer et al. (2000) noted that adaptive testing in this fashion allows "individuals to work at their own pace . . . aside from the practical necessity of having rough limits on the time of testing" (p. 11). Hypothetically, a test can run long in time if the informative items chosen under the MI method also tend to be time-consuming. van der Linden, Scrams, and Schnipke (1999) point out that this is not unlikely to happen for capable examinees, because for them, adaptive testing "results in more difficult items, and more difficult items generally require more time" (van der Linden & van Krimpen-Stoop, 2003, p. 251).

Accordingly, an "efficient" test from the perspective of accumulating more information given a fixed number of items may turn out to be inefficient in terms of the length of administration time. Naturally another approach to evaluating test efficiency is to achieve high test information in a fixed amount of time. We will call the first concern "form length efficiency" (FLE) and the second, "delivery time efficiency" (DTE). When concerned with delivery time efficiency, a maximizing information per time unit (MIT) strategy makes better sense than the MI method. If an adaptive test is of fixed length in terms of the number of items, the MIT strategy would lead to reduced testing time.

Fan, Wang, Chang, and Douglas (2012) introduced the first implementation of the MIT method. The Fan et al. (2012) approach requires individual-level response time information, and fitting the log-normal model to the response time data. It also requires real-time update of the working speed parameter estimate for each examinee. Issues of potential model misfit, as well as high demand in data and computational resources of the original MIT method have prompted us to propose a simplified version, MIT-S, in this article. The performance of the MIT-S method is evaluated under various item response theory (IRT) models against the performance of the MI method in terms of measurement precision, savings in test completion time, and item pool usage.

Background

The MI method is one of the most widely used item selection strategies in IRT-based computerized adaptive testing (CAT) (van der Linden, 2003). Mathematically, the MI method selects the $(t + 1)$th item as

$$\max_{l} \left\{ I_l\left(\hat{\theta}^{(t)}\right) : l \in R_t \right\}, \tag{1}$$

where θ is ability, $\hat{\theta}^{(t)}$ is the ability estimate after t items, R_t is the eligible set of items in the pool after t items have been administered, and $I_l\left(\hat{\theta}^{(t)}\right)$ is the information of item l evaluated at $\hat{\theta}^{(t)}$. Asymptotically the resulting test maximizes information at the true θ, or equivalently, reaches the highest measurement precision, for a given number of items (form length) L.

The MI method has its disadvantages. For example, it is recognized that the MI method favors more discriminating items. Under the IRT models, except for the model that assumes an equal discrimination parameter across items (i.e., the one-parameter logistic model or 1PL model), item information relies heavily on the discrimination parameter. For example, the two-parameter logistic (2PL) model specifies the probability of an examinee of latent trait θ giving a correct response to item l, denoted by u_l, as follows:

$$P(u_l = 1|\theta, \gamma_l) = \frac{\exp[a_l(\theta - b_l)]}{1 + \exp[a_l(\theta - b_l)]}, \tag{2}$$

where $u_l = 1$ if the answer is correct and $u_l = 0$ otherwise. The item parameters $\gamma_l = (a_l, b_l)'$, where a_l and b_l are the item discrimination and difficulty parameters, respectively. Hereafter $P(u_l = 1|\theta, \gamma_l)$ will be denoted by $P_l(\theta)$ for short. Compared to the 2PL model, the three-parameter logistic (3PL) model adds one extra item parameter, the pseudoguessing parameter, c_l. The 1PL model, on the other hand, simplifies the 2PL model by assuming that all items have the same discrimination parameter, and a_l therefore drops out of Eq. (2). The item information function for the 2PL model is given by

$$I_l(\theta) = a_l^2 P_l(\theta) Q_l(\theta). \tag{3}$$

From Eq. 3, it is clear that other things being equal, a higher discrimination parameter leads to higher information. This holds true for item information computed on the basis of the 3PL model, as well. Given a limited item pool, the MI method therefore keeps selecting the most discriminating items. Depending on the item pool size and the presence of content constraints, such highly exposed items may pose a security threat to the CAT program (Chang & Zhang, 2002). This is why the evaluation of a CAT program will typically not only involve evaluation of measurement precision, but also item pool usage in terms of how frequently each item is used or exposed.

Despite its disadvantage in exposure control, the MI method is the most widely used item selection method for adaptive testing, due to its superior efficiency in FLE. This method, however, does not take into account response time, and thereby does not address DTE. An item may provide more information but also require much longer to answer. Since all tests

are likely to have time limits based on construct-relevant constraints or practical limitations of delivery contexts, and examinee and proctor time always has some cost, considering statistical approaches to DTE as well as FLE is an important issue in the psychometrics of applied assessment.

To address DTE, Fan et al. (2012) therefore proposed selecting the next item that maximizes the information per time unit as

$$\max_l \left\{ \frac{I_l\left(\hat{\theta}^{(t)}\right)}{E\left[T_l\middle|\hat{\tau}^{(t)}\right]} : l \in R_t \right\}, \tag{4}$$

where τ is the latent working speed (analogous to the latent trait θ in the 2PL model). The denominator, $E\left[T_l|\hat{\tau}^{(t)}\right]$, is the expected time of an examinee takes to answer item l, given his or her most recent estimate of working speed, $\hat{\tau}^{(t)}$. The person parameters, θ and τ, are assumed to follow multivariate normal distribution, often with a positive correlation assumed between θ and τ. In a CAT, θ and τ can be updated simultaneously.

Fan et al. (2012) used the log-normal model for response time that was proposed by van der Linden (2006):

$$T_{il} \sim f(t_{il}) = \frac{\alpha_l}{\sqrt{2\pi} t_{il}} exp\left\{ -\frac{1}{2}\left[\alpha_l(lnt_{il} - (\beta_l - \tau_i))\right]^2 \right\}, \tag{5}$$

where T_{il} is the response time for the ith examinee to answer item l. The item parameters are α_l and β_l, which are analogous to the item discrimination (a_l) and difficulty parameters (b_l) in a usual IRT model. Here β_l is the time intensity parameter. Larger β_l suggests the item is more time-consuming. Similar to the 2PL model, which can be estimated if item response data are available, the α_l and β_l in the log-normal response time model can be calibrated given individual and item level response time data and used in CAT. In CAT, similar to$\hat{\theta}$, $\hat{\tau}$ can be updated using the maximum likelihood method every time an item is answered. Then $E\left[T_l|\hat{\tau}^{(t)}\right]$, the expected time required to finish item l given the examinee's speed of $\hat{\tau}^{(t)}$, can be computed as

$$E\left[T_l\middle|\hat{\tau}^{(t)}\right] = \exp\left(\beta_l - \hat{\tau}^{(t)} + \frac{1}{2\alpha_l^2}\right). \tag{6}$$

Fan et al. (2012) found that on average, the MIT method, as compared to the MI method, saves substantial testing time, with only a small loss of measurement precision. Unfortunately, it also results in substantially worse item pool usage. Fan et al. (2012) proposed addressing these concerns by combining the MIT method with the a-stratified-with-b-blocking method (Chang, Qian, & Ying, 2001) to control item exposure. They found that the hybrid method is effective in balancing item pool usage. These findings hold when a) the item responses follow the 3PL model; and b) when the item response time follows the log-normal model of van der Linden (2006).

The implementation of the MIT method as introduced in Fan et al. (2012) requires individual- and item-level response time information, as well as fitting the log-normal model to the response time data. Good model fit is therefore a prerequisite to using this method. There are a plethora of models to choose from to model response time in the psychological and educational-testing literatures, including parametric models based on the log-normal (van der Linden, 2006) distribution, which is motivated by distribution fitting, as well as the Weibull distribution (Rouder, Sun, Speckman, Lu, & Zhou, 2003), gamma distribution (Maris, 1993), and Poisson counter process (Ratcliff & Smith, 2004), which are motivated by modeling the psychological process producing the response time. Some of these models try to capture the cognitive process underlying reaction time of simple tasks in psychological experiments. For example, the sequential sampling models characterize response time of respondents making decisions between two choices (Ratcliff & Smith, 2004), where these decisions are simple, "rapid, one-process decisions (e.g., less than 1,000–1,500-ms mean RT at a maximum)" (Ratcliff & Smith, 2004, p. 335). Such models are therefore not suitable for modeling response time in educational testing. In addition, it has been found that the shapes of empirical response time distributions for items within a test and of similar types of tests can vary (Klein Entink, Kuhn, Hornke, & Fox, 2009; Ranger & Kuhn, 2012). Thus, even though many models for response time can be chosen from, "given the diversity of response time distributions that the items in a large, operational item pool might exhibit" in educational testing, no single model may universally fit well all items in an item bank (Patton, 2014, p. 58). As a consequence, the original MIT method could face serious challenges.

The MIT method also requires real-time updating of the person working speed parameter estimate $\hat{\tau}$. The time intensity measure in the denominator of Eq. 4 is thus individualized. Interestingly, such an individualized measure does not make any difference in rank-ordering the items for item selection. In other words, the rank-ordering of the items in terms of $\frac{I_l\left(\hat{\theta}^{(t)}\right)}{E\left[T_l|\hat{\tau}^{(t)}\right]}$ does not change for students with high versus low working-speed estimate. This is because, $E\left[T_l|\hat{\tau}_i^{(t)}\right]$ is equal to exp $\left(\beta_l + \frac{1}{2\alpha_l^2} - \hat{\tau}_i^{(t)}\right)$ when the log-normal model holds, where $\hat{\tau}_i^{(t)}$ is the working speed estimate of examinee i after t items are answered. The working speed estimate $\hat{\tau}_i^{(t)}$ does not change with l, and does not affect the rank-ordering of the items. It therefore calls into question if it is necessary to have an individualized estimate in the denominator of the objective function in Eq. 4.

To broaden the applicability of methods focused on improving CAT delivery in terms of DTE, we propose a simplified version of the MIT method, denoted by *MIT-S*, which does not require model fitting or real-time estimation of the working-speed parameter. Aside from the simplification in item selection, there are two additional important differences between this study and Fan et al. (2012). First, we evaluated the item selection algorithms under various IRT models, in particular the 1PL. This model is a popular candidate for CAT (e.g., Chuesathuchon & Waugh, 2010), especially with the increasing use of IRT in medical research (Elhan et al., 2008; Öztuna et al., 2010; Velozo, Wang, Lehman, & Wang, 2008), and the rise of adaptive and computer-based delivery of assessment in large scale learning systems (e.g., Behrens, Mislevy, DiCerbo & Levy, 2012). The investigation of the 1PL case is particularly important, because it is unclear whether the time-adjusted item selection methods will have exposure control issues under the 1PL model. Recall that under the 2PL and 3PL models, both the MI and the MIT methods favor highly discriminating items. Under the 1PL model, all items are equally discriminating. Does that mean there will not be any issue regarding item exposure under the 1PL model? Will highly time-saving items be favored under the MIT strategy? If so, to what degree will time-saving items be favored when item discrimination is not a factor? These questions need to be answered.

The second difference between the present study and Fan et al. (2012) is that we retain relationships among the item parameter estimates by using an empirical instead of a simulated item bank. This may have important implications on the performance of the item selection method, especially in terms of exposure control and testing time saving. For example, it is well-known that in an operational item pool, the item discrimination and difficulty parameters are often positively correlated. It also makes intuitive sense that difficult items tend to be more time-consuming. If item discrimination, item difficulty, and time intensity are a positively correlated trio, then the appeal of highly discriminating items may be offset by their time intensity. In that case, would the highly discriminating items still be favored? With these questions in mind, we next examine the performance of the MIT-S method under various IRT models using an empirical item bank.

The MIT-S Method

The simplified MIT method, or MIT-S, replaces the denominator in Eq. 4 with the average time it takes to answer an item. In other words, the item is selected following

$$\max_{l}\left\{\frac{I_l\left(\hat{\theta}^{(t)}\right)}{\overline{lnT_{il}}}, l\in R_t\right\}, \tag{7}$$

where $\overline{lnT_{il}}$ is the average of log-transformed response time to item l. If the log-normal model holds,

$$lnT_{il}\Big|(\alpha_l,\beta_l,\tau_i)\sim N\left((\beta_l-\tau_i),\ \frac{1}{\alpha_l^2}\right).$$

So, $E(lnT_{il}|(\alpha_l,\beta_l,\tau_i))=\beta_l-\tau_i$, and

$$\overline{lnT_{il}}\Big|(\alpha_l,\beta_l)\sim N\left((\beta_l-\mu_\tau),\ \frac{1}{\alpha_l^2}\right).$$

In this context, $\overline{lnT_{il}}$ as the average log-transformed response time is the maximum likelihood estimate of the mean of the log-normal distribution over the examinee population, which is $E(\beta_l-\tau_i)=\beta_l-\mu_\tau$, where $\mu_\tau$is the mean working speed of the examinee population. In other words, $\overline{lnT_{il}}$ serves as an estimate of the difference between the time intensity parameter of an item and the group-level speed. A more time-consuming item and slower test takers will lead to larger $\overline{lnT_{il}}$. Apparently, $\overline{lnT_{il}}$ ignores the individual difference in working speed, thereby serving as a non-individualized time intensity measure. But as discussed earlier, in item selection only the rank order of items is of concern. Consequently, the individualized time intensity measure in the denominator of Eq. 4 does not matter, and $\overline{lnT_{il}}$ may suffice as a measure of time intensity for the purpose of item selection.

This simplification means that we do not need response time data from each individual on every item, but only require the average time spent on each item. The simplification also means that it is not necessary to fit a particular response time model to the data and this type of model misfit is no longer a concern. Furthermore, we do not need to obtain precise estimates of the parameters of the response time model for them to be used in CAT. Additionally, real-time updating of the working speed estimate, $\hat{\tau}$, is not needed. In summary, the MIT-S method is less demanding in terms of data, preprocessing (i.e., model fit and item calibration), and computational resources.

Whereas this mathematical simplification is demonstrable, it is still important to investigate whether the MIT-S method maintains the time-saving advantage of the MIT method. In question is also how the time-adjusted item selection indices will perform when the underlying IRT model is 1PL. In addition, it is important to understand the implications of these insights on applications using a real item bank in which one keeps the relationship among the item parameters intact. Such a case may, for example, maintain a correlation between item discrimination and difficulty, as well as between item discrimination and the time intensity of an item (now measured by $\overline{lnT_{il}}$).

The Fan et al. (2012)'s simulation study investigated the case in which (a) an examinee's working speed and ability have a moderate positive correlation (.5) or no correlation (0), and (b) an item's time intensity (measured by β_l) and difficulty have a small positive correlation (.25) or no

correlation (0). The correlation between speed and ability is supported by theory and empirical research; that is, across examinees, speed and ability have a positive correlation (van der Linden, 2009), even though within an examinee there can be accuracy–speed trade-off. For the correlation between time intensity and difficulty, empirical studies show mixed results: Some studies have indicated that the time intensity and item difficulty of an item are positively correlated, and other studies have yielded negative correlations (van der Linden, 2009). The results are mixed, but most of the empirical correlations are positive and significant. Intuitively, this makes sense—difficult items tend to be more time-consuming. Some of the nonsignificant, or negative, correlations may be artifacts of missing data, because the response time data often come from CAT, meaning that the difficult items are only seen by capable examinees. The reader is referred to van der Linden (2009) for a discussion of conceptual issues and a summary of empirical findings with respect to response time modeling.

In their work, Fan et al. (2012) assumed that item difficulty (b_I) and discrimination (a_I) are independent. These authors also assumed that item discrimination (a_I) and time intensity (measured by β_I) are independent. It is well-known that in operational item banks the item difficulty (b_I) and discrimination (a_I) have substantial correlation (Chang, Qian, & Ying, 2001). It is yet unknown whether item discrimination and time intensity are independent. We argue that the assumption of independence between these parameters has practical implications for exposure control using the MIT method. If highly discriminating items tend to be highly time-consuming, the MIT method may not favor the highly discriminating items as the MI method would, even under the 2PL or 3PL model. For these reasons, we used item parameters from a real rather than a simulated item bank in this study, and we chose to use non-CAT data to calibrate the items.

Simulation study 1

We used data from a large item bank of mathematics placement test items. Instructors or administrators are given access to a large catalog of mathematics topics and corresponding items that can be used to create postsecondary mathematics placement examinations. All of the students of one instructor receive the same questions. Across instructors, many of the data are missing. However, no question is chosen tailored to an individual examinee. The missingness is therefore accounted for by class membership and is not related to the underlying latent constructs, such as ability or work speed. We consider this an advantage for fitting IRT models.

Using a subset of the data from recent years of the large-scale math placement test, we calibrated the 3PL model parameters for the 595 items of pre-algebra. The average item response time ranged from 23.17 to 539.01 s, with a skew of 1.48 and a kurtosis of 3.98. Clearly the average item response time was positively skewed. As expected, a positive and significant correlation (.564) was found between item difficulty and $\overline{lnT_{il}}$ (the time intensity measure of the MIT-S method). The correlation between item discrimination and difficulty was .350, which is statistically significant. This positive and significant correlation is consistent with previous findings reported (Chang et al., 2001). The correlation between item discrimination and time intensity was .205, which is statistically significant and reflects the fact that discriminating items tend to be more time-consuming. Descriptive statistics and correlations of the item parameter estimates and the time intensity measure $\overline{lnT_{il}}$ are provided in Table 1. The first three rows of the table show correlations, and the bottom rows show descriptive statistics. The item bank, with an average difficulty of .44, is a little on the difficult side. It has over twice as many items that are hard ($b_I > 0$) as items that are easy ($b_I < 0$). The items are also highly discriminating, with an average discrimination parameter of 1.61.

On the basis of the item parameter estimates of usual IRT models and the time intensity measure $\overline{lnT_{il}}$ from the large-scale math placement test data, we conducted a simulation study to examine the performance of the MIT-S method. Parallel to Fan et al. (2012), the performance of the MIT-S and MI methods were compared in terms of measurement precision, time saving, and exposure control. In the simulation study, tests of length 20 and 40 were simulated, because test length has implications for test completion time. As expected, the MIT strategy would likely select the best items that were both discriminating and time saving. There might not be many items of this type in the bank, however. If the test is long, the MIT strategy could exhaust its favorite type of items and start selecting less favorable items. As a result, the advantage of time saving might not be as pronounced if the test is long. This is the rationale for including two test lengths.

A sample of 5,000 examinees was generated, and their abilities followed N(0, 1). The first item of the CAT was selected randomly. The ability estimate $\hat{\theta}^{(t)}$ was updated by an expected a posteriori for which the prior of θ was N(0, 1). Measurement precision was evaluated by means of the bias and the mean squared error (MSE) of $\hat{\theta}$, as well as the

Table 1 Correlation and descriptive statistics of IRT item parameter estimates and time intensity

	3PL_*a*	3PL_*b*	3PL_*c*	Time Intensity
3PL_*a*	1	.350**	–.363**	.205**
3PL_*b*	.350**	1	–.226**	.564**
3PL_*c*	–.363**	–.226**	1	–.043
Minimum	0.27	–3.65	0.00	4.77
Maximum	2.88	4.12	0.40	6.29
Mean	1.61	0.51	0.10	4.77

correlation between the true θ and the final $\hat{\theta}$ —that is, $\rho\left(\theta, \hat{\theta}\right)$. For each examinee, $\sum\overline{lnT_{il}}$ was computed. Then the average of $\sum\overline{lnT_{il}}$ over all 5,000 examinees was computed as a measure of the average time to finish the fixed-length (in terms of the number of items administered) CAT. Note that in Tables 2, 3, 4, 5 and 6, the completion time is reported on the raw response time scale—that is, $\exp\left(\sum\overline{lnT_{il}}\right)$ is reported.

Item pool usage was measured as the percentages of unused (i.e., exposure rate = 0), underused (i.e., exposure rate < 2%), and overexposed (i.e., exposure rate > .20) items. Ideally, we would like to see the item exposure spread among items, instead of concentrating on a few items. In other words, low percentages of unused, underused, and overused items are desirable. A summary statistic of the item pool usage (Chang & Ying, 1999) was also used, given by

$$\chi^2 = \sum_{j=1}^{M} \frac{\left(er_j - \bar{er}\right)^2}{\bar{er}}, \tag{8}$$

where M is the total number of items in the bank, er_j is the observed exposure rate of item j, and $\bar{er}$ is the expected item exposure rate if no item is favored over others, or $\bar{er} = \frac{L}{M}$. Large χ^2 values indicate unbalanced item pool usage.

Results of study 1

Table 2 shows the performances of the MIT-S method and the MI method in terms of measurement precision, exposure control, and testing time under the 3PL model. Again, measurement precision is captured by bias, MSE, and the correlation between the true θ and the final $\hat{\theta}$ —that is, $\rho\left(\theta, \hat{\theta}\right)$. According to these indices, the MIT-S method is comparable to the MI method in measurement precision, regardless of test length. In terms of exposure control, the χ^2 index indicates that neither the MI nor the MIT-S method does very well in balancing item pool usage. If the item pool usage is completely balanced (i.e., every item is used equally frequently), the χ^2 index should be 0. A larger χ^2 index suggests worse item pool usage. These results suggest that the MIT-S method results in only slightly worse item pool usage than does the MIT method. This is true regardless of test length. The percentages of never used, underused, and overused items are very comparable between the MI and MIT-S methods. This is in stark contrast to the finding in Fan et al. (2012), which indicated that trying to select time-saving items with the MIT method led to worse item pool usage than did the MI method. As the test lengthens, both the MI and MIT-S methods do better balancing item pool usage, because some items that were never used are now exposed. At the same time, the proportion of overexposed items also increases. With respect to time saving, the MIT-S method does better, with on average a 12% reduction in testing time, regardless of the test length. This is not as impressive as the value reported in Fan et al. (2012), which showed a reduction in response time by about 1/3. So the simplification of the MIT-S as compared to the MIT method results in some loss in time saving.

In summary, the MIT-S method maintains the measurement precision and reduces testing time by about 12%. The performance of MIT-S in exposure control is very similar to that of the MI method. For MIT-S, the correlation between the item exposure rate and item discrimination is .39 when the test length is 20. The item exposure rates under the MIT-S method and the MI method correlate at .95. Correlations of similar magnitudes are observed when the test length is 40. This indicates that item selection under the MIT-S method, as under the MI method, still favors highly discriminating items. This is the reason that the *a*-stratified method was effective in controlling item exposure with the MIT method in Fan et al. (2012). For the same reason, we expect the *a*-stratification method would be effective with the MIT-S method, as well, when the underlying IRT model was 2PL or 3PL.

Table 2 Performance of the MIT-S and MI methods under the 3PL model

		20 Items		40 Items	
		MI	MIT-S	MI	MIT-S
Average test completion time (mins)		37.76	33.32	79.19	70.03
Measurement precision	Bias	.001	.000	.001	.003
	MSE	.033	.032	.019	.020
	$\rho\left(\theta, \hat{\theta}\right)$	.984	.985	.991	.990
Exposure control	χ^2	115.36	118.52	102.00	104.09
	No exposure	73.9%	74.1%	54.3%	53.9%
	Underexposed (<.02)	79.5%	79.8%	61.3%	60.5%

Next, we fixed the pseudo-guessing parameter to 0 for all of the items, and the resulting item parameters were used for the 2PL simulation. These results are summarized in Table 3. The patterns are very similar to those under the 3PL model (see Table 2). The MIT-S method achieves the same level of measurement precision as the MI method. The item bank usages are very similar when we look at the proportions of items that are not used at all, underused, and overused. Meanwhile, the MIT-S method reduces the testing time by 14% (for test length = 20) and 12% (for test length = 40). Although the time saving is indeed more pronounced with a shorter test, the difference is small under the 2PL model.

A key question we are trying to answer is how MIT-S performs when the underlying IRT model is the 1PL model. In this case, all items have the same item discrimination parameter. Therefore, we do not expect item selection under the MI or MIT-S to favor highly discriminating items. However, item selection under the MIT-S method might favor highly time-saving items, so we further constrained the discrimination parameters from the 2PL model to 1.0. Note that the correlation between item difficulty and time intensity is still maintained. The results are summarized in Table 4 for test lengths of 20 and 40.

This table indicates that the MIT-S method results in a very skewed item exposure distribution. When the test length is 20, 78.2% of the items are never used, and 82.7% of the items in the pool are used less than 2% of the time. This indicates that the majority of the items are dormant. When the test length is 40, about 66.6% of the items are underused. In terms of measurement precision, the MIT-S method does as well as the MI method, as indicated by the extremely similar bias, MSE, and $\rho\left(\theta,\hat{\theta}\right)$. In addition, the MIT-S method leads to substantial time saving when the underlying model is 1PL. The total testing time is reduced by 54.4% when the test length is 20, and by 46.4% when the test length is 40. Here again, we see the time saving is more pronounced when the test is short.

In summary, the MIT-S method does well in maintaining the measurement precision, while reducing the testing time dramatically in comparison to the MIT approach, when the underlying true IRT model is 1PL. However, it performs much worse than the MI method in exposure control, insofar as a large portion of the item bank is wasted, and the exposure concentrates on a very small subset of items. In fact, the highly exposed items tend to be the highly time-saving ones. The correlations between item exposure rate and time intensity are –.40 (when the test length is 20) and –.54 (when the test length is 40), indicating that time-consuming items are used less frequently.

Exposure control under the 1PL model and simulation study 2

We investigated how we can address the exposure control problem with the MIT-S method when the model is 1PL. Recall that Fan et al. (2012) found that the item bank usage was very unbalanced under the 2PL and 3PL models; therefore, they combined the *a*-stratification design with the MIT strategy to control and balance exposure rates. The *a*-stratification design, first proposed by Chang and Ying (1999) and later modified by Chang, Qian, and Ying (2001), was a novel design for exposure control. The design used in Fan et al. (2012) was *a*-stratified CAT with *b*-blocking, the modified version introduced in 2001. It involves first rank-ordering the items in the pool according to their difficulty parameters. Items that are similar in difficulty are then grouped to form a "*b*-block," and the items within each block are rank-ordered again by 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

61.2%

Table 3 Performance of the MIT-S and MI methods under the 2PL model

		20 Items		40 Items	
		MI	MIT-S	MI	MIT-S
Average test completion time (mins)		34.52	29.83	72.98	64.06
Measurement precision	Bias	.002	.002	.001	.001
	MSE	.025	.025	.015	.015
	$\rho\left(\theta,\hat{\theta}\right)$	.988	.988	.993	.993
Exposure control	χ^2	108.5	114.78	95.68	100.18
	No exposure	73.4%	73.9%	52.9%	54.8%
	Underexposed (<.02)	78.8%	78.8%	60.3%	
	Overexposed (>.20)	5.7%	6.1%	13.4%	13.8%

Table 4 Performance of the MIT-S and MI methods under the 1PL model

		20-Item		40-Item	
		MI	MIT-S	MI	MIT-S
Average test completion time (mins)		37.58	17.13	75.70	40.56
Measurement precision	Bias	.001	.001	–.002	–.002
	MSE	.076	.075	.038	.039
	$\rho\left(\theta, \hat{\theta}\right)$	.963	.963	.982	.981
Exposure control	χ^2	14.81	153.26	23.25	133.51
	No exposure	6.4%	78.2%	0.5%	59.8%
	Underexposed (<.02)	50.6%	82.7%	15.1%	66.6%
	Overexposed (>.20)	0	5.9%	2.9%	13.1%

formed by pooling the high-*a* items. In this way, the strata have increasing levels of average item discrimination, but the distributions of difficulty remain similar across strata. Examinees are only allowed to see items from the low-*a* stratum in the beginning of the test when our knowledge of the true θ is minimal; higher-*a* items are reserved for the later stage. As a result, not only the highly discriminating items, but also the less discriminating items, are used, and the item bank usage is more balanced. By forming *b*-blocks in addition to *a*-stratification, the stratification design takes into account the positive correlation between the discrimination and difficulty parameters in operational item banks, and consequently maintains measurement precision while balancing item pool usage (Chang et al., 2001). When used in conjunction with the MIT strategy, the method was successful in controlling for item exposure, as well (Fan et al., 2012).

This method is, however, no longer applicable when the underlying model is 1PL. The discrimination parameters are all equal, so there is no basis for *a*-stratification. Some other well-known item exposure control methods are still appropriate, though, such as the randomesque method (Kingsbury & Zara, 1989) and the progressive restrictive method (Revuelta & Ponsoda, 1998).

In contrast to the MI or MIT methods, which choose the best single item in the bank according to an item selection criterion, the randomesque method randomly picks an item out of the *n* best items, where *n* can be 5 or 10, for example. In our study, we used $n = 5$, and the resulting item selection method is denoted as MIT-S-R5. In this approach, at every step the five items that lead to the largest ratio of $\frac{I_l\left(\hat{\theta}^{(t)}\right)}{\sum lnT_{il}}$ will be identified, and one will be randomly chosen from those five to be given to the examinee.

The progressive restrictive method also tries to impose some randomness onto the item selection process. When combined with the MIT-S method, instead of choosing the item that yields the largest ratio of item information to average response time, the progressive restrictive method tries to maximize the weighted sum of the ratio and a random number generated within the range of 0 and the maximum ratio of $\frac{I_l\left(\hat{\theta}^{(t)}\right)}{\sum lnT_{il}}$ achieved on previous items, evaluated at the current $\hat{\theta}^{(t)}$. Early in the test, the weight on the random component is heavy, and the weight decreases as testing progresses. In Revuelta and Ponsoda (1998), the weight on the random component was set to be $1 - t/L$. If the test length is 20, and ten items have been administered, the weight is .5. This is the "progressive" part of the progressive restrictive method. On the other hand, the exposure rates of items are monitored over time. Once an item's exposure rate is over a set limit, the item is set aside, or "hibernates," and becomes ineligible for subsequent item selection. As testing progresses and the item's exposure rate drops below the preset limit, the item is "awakened" and becomes eligible once again for item selection. This is the "restrictive" part of the progressive restrictive method. It effectively prevents an item from being overexposed. These types of techniques are important in a high-stakes assessment situation that requires high security, and thereby strict exposure control. We used the progressive restrictive method with MIT-S and denoted the resulting procedure as *MIT-S-PR*. Under the MIT-S-PR method, items are selected sequentially following

$$\max_l \left\{ \left(1-\frac{t}{L}\right) r^{(t)} + \frac{t}{L} \frac{I_l\left(\hat{\theta}^{(t)}\right)}{\sum \overline{lnT_{il}}}, l \in R_t \right\}, \tag{9}$$

where $r^{(t)}$ is the random number generated at stage *t*, and R_t is the set of eligible items at this stage. Note that because of the "restrictive" nature of the method, items with real-time exposure rates higher than the set limit (in this study, .2) are ineligible for selection.

Other aspects of the simulation study remained the same as those in Simulation Study 1. We fixed the discrimination parameter from the 3PL model parameter estimates to 1, and the pseudoguessing parameter to 0 for all the items. The resulting item parameters were used for the 1PL simulation. Two test lengths were again chosen, 20 and 40, and the item selection methods were compared in terms of measurement precision, time saving, and item bank usage. Table 5 summarizes the performance of the MIT-S method in conjunction with the randomesque method when $n = 5$ (i.e., MIT-S-R5) and the progressive restrictive method (i.e., MIT-S-PR), and contrasts them with the MI and basic MIT-S methods. To facilitate the comparison, the corresponding columns for the MI and MIT-S methods from Table 4 are included in Table 5. The test length is 20, and the underlying true IRT model is 1PL. Note that in this case the MIT-S-R5 method improves the item pool usage substantially. As compared to the MIT-S method, for MIT-S-R5 the χ^2 decreases from 153.26 to 25.61, the percentage of unused items in the pool is reduced by more than half (from 78.2% to 33.9%), and no item is overexposed. It maintains the time-saving advantage of the MIT-S method, so that the testing time is reduced from 37.58 (for the MI method) to 23.19 min, a reduction of almost 40%. The MIT-S-PR method also helps improve item pool usage, but it is not as effective as the randomesque method. Its advantage in time saving is slightly more pronounced, though (a reduction of the average total testing time by 43%). Note that it also keeps any item from being overexposed, indicating that the progressive restrictive method is effective in protecting the security of the test items. Comparison of the MIT-S, MIT-S-PR, and MIT-S-R5 methods suggests that exposure control slightly reduces the time-saving advantage, because the item selection algorithm is forced to use some of the more time-consuming items. However, MIT-S-R5 still leads to substantial reductions in testing time and balances the item pool usage, with nearly no loss of measurement precision, since the differences in bias, MSE, and $\rho\left(\theta, \hat{\theta}\right)$ occur only in the third decimal place.

Table 6 provides the same comparison of methods when the test length is 40. The corresponding columns of Table 4 are included in Table 6 to facilitate comparison. The general patterns we observed in Table 5 are again observed in Table 6. Between the two exposure control methods, the randomesque approach performs better in exposure control, because it more effectively promotes exposure of the items that are selected less often. Both approaches are successful in keeping any item from being overexposed. The MIT-S-R5 method leads to a 10% reduction in testing time, whereas the MIT-S-PR method leads to a 30% reduction.

Overall, the MIT-S method does very well reducing testing time, but it leads to very unbalanced item pool usage under the 1PL model. When used in conjunction with the randomesque method, however, it successfully maintains measurement precision, saves testing time (though not as much as the original MIT-S method), and leads to good exposure control. The time-saving advantage of the MIT-S-R5 method is more evident when the test is short.

Summary and Discussion

In this study, we investigated the performance of the MIT-S method under various IRT models. This is a simplified version of the original MIT method proposed by Fan et al. (2012). Our results indicate that the MIT-S method effectively reduces testing time without compromising measurement precision. In terms of item pool usage, it is comparable to the MI method when the underlying IRT model is 2PL or 3PL. This result is in contrast to the prior findings of Fan et al. (2012), which suggest that trying to save time exacerbates the unbalanced item pool usage. The difference in results may be due to the fact that item discrimination and time intensity are correlated in the empirical item bank. From this perspective, Fan et al.

Table 5 Performance of the MIT-S and MI methods with exposure control under the 1PL model, test length = 20

		MI	MIT-S	MIT-S-R5	MIT-S-PR
Average test completion time (mins)		37.58	17.13	23.19	21.50
Measurement precision	Bias	.001	.001	.005	.001
	MSE	.076	.075	.082	.078
	$\rho\left(\theta, \hat{\theta}\right)$	.963	.963	.960	.961
Exposure control	χ^2	14.81	153.26	25.61	54.22
	No exposure	6.4%	78.2%	33.9%	37.5%
	Underexposed (<.02)	50.6%	82.7%	57.0%	72.3%
	Overexposed (>.20)	0	5.9%	0	0

Table 6 Performance of the MIT-S and MI methods with exposure control under the 1PL model, test length = 40

		MI	MIT-S	MIT-S-R5	MIT-S-PR
Average test completion time (mins)		75.70	40.56	68.66	51.52
Measurement precision	Bias	–.002	–.002	.002	.001
	MSE	.038	.039	.044	.040
	$\rho\left(\theta, \hat{\theta}\right)$	.982	.981	.978	.980
Exposure control	χ^2	23.25	133.51	18.41	49.36
	No exposure	0.5%	59.8%	5.9%	13.3%
	Underexposed (<.02)	15.1%	66.6%	20.3%	50.4%
	Overexposed (>.20)	2.9%	13.1%	0	0

(2012) might have sold the MIT method a little short. Its advantage in testing time savings does not necessarily come at the expense of worse exposure control than under the MI method. This indicates that item bank structure might have substantial influence on CAT outcomes.

The MIT-S method leads to substantial time savings when the underlying IRT model is 1PL, but as we would expect, it results in poor item bank usage, because it favors highly time-saving items. When the MIT-S method is used with randomesque exposure control, the item bank usage is much improved.

These results suggest the need for future research in several directions. First, the MIT method and the MIT-S method should be compared when the log-normal model fits and when the assumption of the log-normal model is violated (i.e., model misfit). We expect that the MIT-S method will be more robust against model misfit. On the other hand, as a reviewer pointed out, both the MIT and MIT-S methods conduct item selection by maximizing an information-gain-to-time-cost ratio. When the time cost function is defined differently, the information gain is expressed in a different metric. Many arbitrary choices exist regarding the gain-to-cost-ratio metric, some more justifiable than others. A well-justified metric that deserves further study is to use the estimate of the average raw response time reconstructed from the maximum likelihood estimates (MLEs) of the log-normal distribution, instead of the average log-transformed response time, as the denominator in Eq. 7. This would be asymptotically equivalent to the ratio of expected gain to expected time cost in Kujala (2010). Further research is warranted to investigate the performance of such a metric, and to compare its performance against the original MIT and MIT-S methods.

Second, these results suggest that additional research should be conducted to take into account realistic constraints on CAT—for example, content balancing and item type balancing. It may be necessary to consider these constraints, because items of various types and from various content areas may differ qualitatively in their time intensities. For example, short essay questions usually take much longer to finish than multiple-choice items. Consequently, the MIT or MIT-S method, by favoring time-saving items, may disproportionally choose multiple-choice items. As a result, the test composition could deviate substantially from the desired blueprint. Because of this, we should explore combining the MIT or MIT-S method with popular constraint management methods, such as the weighted deviation modeling method (Stocking & Swanson, 1993), the maximum priority index method (Cheng & Chang, 2009), and the shadow test approach (van der Linden, 2010).

Finally, we recommend examination of the issue of controlling the total response time under the MIT or the MIT-S method. Van der Linden (2009) and van der Linden and Xiong (2013) are pioneering studies on controlling the total test time using item-level and individual-level response time information. Although the MIT and MIT-S methods save testing time on average, there is still a risk that the total testing time limit for an examinee will be exceeded. It is therefore essential to keep the total testing time under control for each individual examinee.

Although a large literature exists concerning optimizing CAT from the perspective of FLE, the literature concerning optimizing CAT from a DTE perspective is relatively new. We have extended this nascent literature by simplifying the operational assumptions required to follow the MIT approach with the MIT-S formulation. In addition, we show how the interaction of item bank structure, with variations in IRT models and item selection algorithms, can lead to higher DTE. The increase in the use of additional (time) data regarding examinee performance, combined with additional variations in algorithmic details, suggests directions for continued improvement in CAT methods across a broad range of assessment and learning contexts.

Author note This work was supported by a 2014 CTB/McGraw-Hill R&D Grant and a grant to the first author from the National Science Foundation: DRL-1350787. The authors thank Adele Brandstrom for her generous help with editing the second draft of the manuscript.

References

Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2012). Evidence centered design for learning and assessment in the digital world. In M. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 13–54). Charlotte, NC: Information Age.

Chang, H.-H., & Ying, Z. (1999). Alpha-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23,* 211–222.

Chang, H.-H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67,* 387–398.

Chang, H.-H., Qian, J., & Ying, Z. (2001). Alpha-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement, 25,* 333–341.

Cheng, Y., & Chang, H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62,* 369–383.

Chuesathuchon, C., & Waugh, R. F. (2010). Item banking and computerized adaptive testing with Rasch measurement: An example for primary mathematics in Thailand. In R. F. Waugh (Ed.), *Applications of Rasch Measurement in Education* (pp. 1–36). Hauppauge, NY: Nova Science.

Elhan, A. H., Oztuna, D., Kutlay, S., Küçükdeveci, A. A., & Tennant, A. (2008). An initial application of computerized adaptive testing (CAT) for measuring disability in patients with low back pain. *BMC Musculoskeletal Disorders, 9,* 166. doi:10.1186/1471-2474-9-166

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

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., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods, 14,* 54–75. doi:10.1037/a0014877

Kujala, J. V. (2010). Obtaining the best value for money in adaptive sequential estimation. *Journal of Mathematical Psychology, 54,* 475–480.

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

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika, 58,* 445–469.

Öztuna, D., Elhan, A. H., Küçükdeveci, A. A., Kutlay, S., & Tennant, A. (2010). An application of computerised adaptive testing for measuring health status in patients with knee osteoarthritis. *Disability and Rehabilitation, 32,* 1928–1938.

Patton, J. (2014). *Some consequences of response time model misspecification in educational measurement (Unpublished doctoral dissertation)*. IN: Notre Dame.

Ranger, J., & Kuhn, J.-T. (2012). A flexible latent trait model for response times in tests. *Psychometrika, 77,* 31–47.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111,* 333–367. doi:10.1037/0033-295X.111.2.333

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

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika, 68,* 589–606. doi:10.1007/BF02295614

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

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. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46,* 247–272.

van der Linden, W. J. (2010). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 31–55). New York, NY: Springer.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (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., & van Krimpen-Stoop, E. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. *Psychometrika, 68,* 251–265.

van der Linden, W. J., & Xiong, X. (2013). Speededness and adaptive testing. *Journal of Educational and Behavioral Statistics, 38,* 418–438.

Velozo, C. A., Wang, Y., Lehman, L., & Wang, J. H. (2008). Utilizing Rasch measurement models to develop a computer adaptive self-report of walking, climbing and running. *Disability and Rehabilitation, 30,* 458–467.

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., . . . Thissen, D. (Eds.) (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Erlbaum.

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