PSYCHOMETRIKA—VOL. 83, NO. 3, 650–673
SEPTEMBER 2018
https://doi.org/10.1007/s11336-017-9596-3

# SEQUENTIAL DETECTION OF COMPROMISED ITEMS USING RESPONSE TIMES IN COMPUTERIZED ADAPTIVE TESTING

EDISON M. CHOE

GRADUATE MANAGEMENT ADMISSION COUNCIL® (GMAC®)

JINMING ZHANG AND HUA-HUA CHANG

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Item compromise persists in undermining the integrity of testing, even secure administrations of computerized adaptive testing (CAT) with sophisticated item exposure controls. In ongoing efforts to tackle this perennial security issue in CAT, a couple of recent studies investigated sequential procedures for detecting compromised items, in which a significant increase in the proportion of correct responses for each item in the pool is monitored in real time using moving averages. In addition to actual responses, response times are valuable information with tremendous potential to reveal items that may have been leaked. Specifically, examinees that have preknowledge of an item would likely respond more quickly to it than those who do not. Therefore, the current study proposes several augmented methods for the detection of compromised items, all involving simultaneous monitoring of changes in both the proportion correct and average response time for every item using various moving average strategies. Simulation results with an operational item pool indicate that, compared to the analysis of responses alone, utilizing response times can afford marked improvements in detection power with fewer false positives.

Key words: test security, response time, computerized adaptive testing, sequential analysis, change-point detection, repeated significance tests.

## 1. Introduction

In a typical administration of computerized adaptive testing (CAT), items are sequentially selected in real time from a large item pool according to the examinee's current performance. Ideally, this provides each examinee with a unique set of items with minimal overlap, thereby discouraging cheating by copying or sharing answers. In practice, however, item selection algorithms based on maximizing information (or minimizing standard error of measurement) are generally prone to highly unbalanced item exposure. Among other concerns, frequently administered items are at great risk for becoming compromised, thereby undermining the integrity of the test.

To counter such a glaring security issue in CAT, much psychometric research in test security has been focused on preventive measures involving some form of item exposure control while still maintaining the efficiency or accuracy of ability estimation as much as possible. A few common strategies in this regard include the Sympson–Hetter (SH) exposure control (Hetter & Sympson, 1997; Sympson & Hetter, 1985), variations or modifications of the SH method (Stocking, 1993; van der Linden, 2003), $a$-stratification techniques (Chang, Qian, & Ying, 2001; Chang & Ying, 1999), the randomesque method (Kingsbury & Zara, 1989), and more fine-grained controls conditional on ability (Stocking & Lewis, 1998; Chang, Ansley, & Lin, 2000). Georgiadou, Triantafillou, and Economides (2007) provide a fairly comprehensive review of item exposure control strategies. However, even the most successful exposure controls cannot entirely prevent the problem of

Correspondence should be made to Edison M. Choe, Graduate Management Admission Council® (GMAC®), 11921 Freedom Drive, Suite 300, Reston, VA 20190, USA. Email: echoe@gmac.com

compromised items in practice, simply because a realistic item pool size is usually much smaller than the number of examinees. Since most items will necessarily be administered multiple times, they are inevitably vulnerable to compromise by unscrupulous test-takers. Therefore, there is a great need for diagnostic measures to spot anomalous behavior of both examinees and items alike.

From the examinee perspective, the general strategy is to detect an aberrant pattern of responses or response times (RTs) across all items that have been administered to the test-taker. There is extensive literature on the use of person misfit statistics and related methods for this general purpose, including but certainly not limited to the following: the $l_z$ statistic and variations thereof (Drasgow, Levine, & Williams, 1985; Armstrong, Stoumbos, Kung, & Shi, 2007), caution indices (Tatsuoka, 1984; McLeod & Lewis, 1999), score ratio (Karabatsos, 2003), likelihood ratio (Levine & Drasgow, 1988), KL divergence and K-Index (Belov, Pashley, Lewis, & Armstrong, 2007; Belov & Armstrong, 2010), posterior shift (Belov, 2015), data forensics (Impara & Kingsbury, 2005), effective response time (Meijer & Sotaridona, 2006), Bayesian checks (van der Linden & van Krimpen-Stoop, 2003; van der Linden & Guo, 2008; van der Linden & Lewis, 2015; Marianti, Fox, Marianna, Veldkamp, & Tijmstra, 2014), CUSUM techniques (van Krimpen-Stoop & Meijer, 2001; Meijer, 2002; Armstrong & Shi, 2009; Egberink, Meijer, Veldkamp, Schakel, & Smid, 2010; Tendeiro & Meijer, 2012), and outlier detection (Mavridis & Moustaki, 2008, 2009; Moustaki & Knott, 2014; Öztürk & Karabatsos, 2017).

From the item perspective, a common strategy is to detect item parameter drift (IPD). In brief, IPD methods broadly focus on identifying items whose parameters may have drifted over time, for a host of reasons ranging from poor initial calibration to changes in curriculum (see, for example, Risk (2015) for a recent review of the literature). Within the vast literature, two papers stand out in their novel use of CUSUM to sequentially monitor IPD for the specific purpose of detecting compromised items. A study by Veerkamp and Glas (2000) employed a standardized CUSUM statistic for detecting drift in the restricted 3PLM (i.e., fixed $c$ parameter), and a recent study by Kang and Chang (2016) extended the technique by using a log-likelihood CUSUM statistic for detecting overall drift in both the unrestricted 3PLM and the lognormal model of RTs within the hierarchical framework (van der Linden, 2007). Although these methods demonstrated great promise, their major drawback is the need for repeated item calibration at each sequential step, which may be infeasible or impractical due to inadequate sample size and tremendous computational burden. Consequently, for practical implementation, CUSUM in this context can only be performed at intervals throughout the usage lifetime of an item (e.g., every 100 times the item is exposed).

Another strategy from the item perspective is to detect an aberrant pattern of responses or RTs across all examinees that have been administered the item. However, literature on this front is relatively scarce, a handful of examples including a merged information theory and combinatorial optimization algorithm (Belov, 2014), a dual differential person functioning (DPF) and differential item functioning (DIF) approach (O'Leary & Smith, 2017), and a log-odds ratio index of item fit (McLeod & Schnipke, 1999). These particular methods can be effective in detecting compromised items, but only after one or more groups of aberrant (or a larger set of potentially aberrant) examinees have first been identified at the end of a testing cycle. Alternatively, Lu and Hambleton (2003) demonstrated the use of an item misfit index $K_1$ (originally reported by Zhu, Yu, and Liu (2002) as $Z_c$) to detect disclosed items in CAT at a single occasion, while Han and Hambleton (2004) utilized the same index to pioneer a procedure for real-time detection of compromised items in linear computer-based testing (CBT). The essential idea in the latter paper is that, for a testing period with a set item bank, each item can be continuously monitored after every exposure for any significant increase in the proportion of correct responses via moving averages. In this fashion, Zhang (2014) and Zhang and Li (2016) investigated the real-time detection procedure in the specific context of CAT. The various implementations of this technique are illustrated shortly after explaining the requisite theoretical framework in the next section. In brief, the procedures

were shown to be capable of detecting compromised items quickly with relatively high accuracy under certain conditions, albeit with room for improvement.

Therefore, in efforts to build upon this promising work, the current study proposes the use of RTs in addition to responses. More specifically, examinees' RTs are incorporated into the process by simultaneously monitoring any significant decrease in the average RT of each item over repeated exposures. By evaluating abnormal changes in both the number of correct responses and the average RTs for items, the procedure can provide even greater statistical power for detecting compromise as well as stronger substantive evidence that an item is indeed compromised. The efficacy of this enhanced method is investigated in detail.

## 2. CAT Framework

The primary purpose of CAT is to measure an examinee's latent trait(s) of interest as efficiently as possible, in terms of either maximal accuracy with a fixed number of items or a certain level of accuracy with a minimal number of items. As such, the core of any CAT system is an adaptive algorithm that strives to select the most appropriate sequence of items for the test-taker. Any such algorithm requires a way to relate the latent trait(s) to the psychometric properties of items, which is principally fulfilled by a class of models within the item response theory (IRT) framework. In particular, the three parameter logistic model (3PLM; Lord & Novick, 1968) is routinely used for applications measuring univariate ability with dichotomous items. It is typically parameterized as

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

in which $X_{ij}$ is a binary random variable mapping the $i$th examinee's response to the $j$th item as either 1 for correct or 0 for incorrect, and $\theta$ is the latent ability parameter. Hence, function $P_j(\theta_i)$ outputs the conditional probability of correctly answering item $j$ given the examinee's ability $\theta_i$, where $a_j$, $b_j$, and $c_j$ represent the item discrimination, difficulty, and pseudo-guessing parameters, respectively. $\theta_i$ and $b_j$ are always scaled on the same continuous metric, which grants a direct and meaningful link between the test-taker's ability and the item's difficulty.

Item selection algorithms are commonly based on the Fisher information, which can be derived for a 3PLM item as

$$I_j(\theta_i) = -E\left(\frac{\partial^2}{\partial\theta_i^2}\log L(\theta_i|x_{ij})\right) = a_j^2\left(\frac{1 - P_j(\theta_i)}{P_j(\theta_i)}\right)\left(\frac{P_j(\theta_i) - c_j}{1 - c_j}\right)^2. \tag{2}$$

Note that $L(\theta_i|x_{ij})$ is the likelihood function of $\theta_i$ given an observed response $x_{ij}$:

$$L(\theta_i|x_{ij}) = P_j(\theta_i)^{x_{ij}}[1 - P_j(\theta_i)]^{1-x_{ij}}. \tag{3}$$

The classic maximum Fisher information (MFI) method chooses the next item with the largest $I_j(\hat{\theta}_i)$, where $\hat{\theta}_i$ is the interim maximum likelihood estimate (MLE) of $\theta_i$ based on the examinee's answers to the previous items (Lord, 1980). Specifically, given observed responses to a set of $k$ items, $\mathbf{x} = \{x_1, x_2, \ldots, x_k\}$, the MLE of $\theta_i$ is computed as

$$\hat{\theta}_i^{ML} = \arg\max_{\theta_i} L(\theta_i|\mathbf{x}_i) = \arg\max_{\theta_i}\prod_{j=1}^{k} P_j(\theta_i)^{x_{ij}}[1 - P_j(\theta_i)]^{1-x_{ij}}, \tag{4}$$

and its estimated standard error is the inverse square root of the cumulative Fisher information across the $k$ items:

$$\mathrm{SE}\left(\hat{\theta}_i^{ML}\right) \approx \frac{1}{\sqrt{I^{(k)}\left(\hat{\theta}_i^{ML}\right)}} = \frac{1}{\sqrt{\sum_{j=1}^{k} I_j\left(\hat{\theta}_i^{ML}\right)}}. \tag{5}$$

Technically, $\hat{\theta}_i^{ML}$ relies on 3PLM's fundamental assumption of local independence (i.e., conditional independence of $X_{i1}, \ldots, X_{ik}$ given $\theta_i$), which is necessarily violated in CAT (Mislevy & Chang, 2000). Nevertheless, under mild regularity conditions with an infinite item pool, $\hat{\theta}_i^{ML}$ for MFI still converges in distribution as follows:

$$\hat{\theta}_i^{ML} \xrightarrow{d} \mathcal{N}\left(\theta_i, \frac{1}{I^{(k)}(\theta_i)}\right) \quad \text{as} \quad k \to \infty \tag{6}$$

(Chang & Ying, 2009; Chang, 2015). As a result, the MFI method theoretically produces an unbiased estimate and maximizes the measurement precision of $\theta_i$ when using MLE.

A notable drawback of MLE, however, is that the estimation can be volatile or even infeasible when there is little to no variation in responses, especially early on when only a few items have been answered. Therefore, expected a posteriori (EAP) estimation is often employed as an alternative, which calculates the expected value of the posterior distribution of $\theta$ given $\mathbf{x}$ (over the parameter space $\Theta$) as follows:

$$\hat{\theta}_i^{EAP} = E_{\theta_i} f(\theta_i|\mathbf{x}_i) = \int_{\Theta} \theta_i \frac{L(\theta_i|\mathbf{x}_i)g(\theta_i)}{\int_{\Theta} L(\theta_i|\mathbf{x}_i)g(\theta_i)\mathrm{d}\theta_i}\mathrm{d}\theta_i = \frac{\int_{\Theta} \theta_i L(\theta_i|\mathbf{x}_i)g(\theta_i)\mathrm{d}\theta_i}{\int_{\Theta} L(\theta_i|\mathbf{x}_i)g(\theta_i)\mathrm{d}\theta_i}. \tag{7}$$

(Bock & Mislevy, 1982). Note that $g(\theta_i)$ is a prior density function of $\theta_i$, which is usually set as uniform or standard normal in the absence of a more informative prior. Since the above expression is often analytically intractable, it can be numerically approximated as

$$\hat{\theta}_i^{EAP} \approx \frac{\sum_Q \theta_q L(\theta_q|\mathbf{x}_i)g(\theta_q)}{\sum_Q L(\theta_q|\mathbf{x}_i)g(\theta_q)}, \tag{8}$$

where $Q$ is a finite set of quadrature nodes ($\theta_q \in Q$) that is representative of $\Theta$. The relationship between $\hat{\theta}_i^{EAP}$ and $\hat{\theta}_i^{ML}$ is theoretically established by the asymptotic posterior normality of $\theta_i$ under weak regularity conditions (Chang & Stout, 1993), which can be interpreted as

$$f(\theta_i|\mathbf{x}_i) \approx \mathcal{N}\left(\hat{\theta}_i^{ML}, \frac{1}{I^{(k)}\left(\hat{\theta}_i^{ML}\right)}\right). \tag{9}$$

In other words, $\hat{\theta}_i^{EAP} \approx \hat{\theta}_i^{ML}$ and $\mathrm{SE}(\hat{\theta}_i^{EAP}) \approx I^{(k)}(\hat{\theta}_i^{ML})^{-1/2}$ for large $k$, thereby justifying the use of the EAP estimator with MFI in the long run.

Regardless of the choice between estimators, the unrestricted form of MFI is highly efficient in terms of ability estimation. Its optimal measurement efficiency, however, comes at the heavy cost of extremely unbalanced item pool usage, as items with large $a$ parameters are disproportionately favored due to their high information (Chang & Ying, 1999; Chang et al., 2001; Hau & Chang,

2001). In fact, low discrimination items are seldom if ever used, which is clearly an inefficient management of resources. Furthermore, high exposure items are at greater risk of compromise to the detriment of test security. Therefore, MFI is almost always restricted with some form of item exposure control in practice.

Among numerous schemes that have been proposed, the classic Sympson–Hetter (SH) method of exposure control (Hetter & Sympson, 1997; Sympson & Hetter, 1985) is perhaps the most well known. It probabilistically enforces a maximum exposure rate as follows: (1) representing the event of selecting item $j$ as $S_j$ and the event of administering item $j$ as $A_j$, the probability of administering an item given that it has been selected is $P(A_j|S_j) = P(A_j \cap S_j)/P(S_j)$; (2) recognizing that an item can only be administered if it has been selected, or $A_j \subseteq S_j$, $P(A_j \cap S_j) = P(A_j)$, which is the actual exposure rate of the item; (3) setting the maximum exposure rate at $r$, or $P(A_j) = r$, the probability of administering the selected item $j$ is set to be $P(A_j|S_j) = r/P(S_j)$. Although effective in theory, a practical limitation of the SH method is that the probability of selecting item $j$, $P(S_j)$ can only be estimated through iterated CAT simulations until a stable value is obtained, which may take as many as 100–150 repetitions (van der Linden, 2003). Furthermore, since unselected items cannot be administered, SH is unable to increase exposure for underexposed items (Chang & Ying, 1999).

A notable alternative to MFI (with or without an externally imposed constraint such as SH) is the $a$-stratified with $b$-blocking design (ASB; Chang et al., 2001), which achieves balance in item pool usage through an innovative item selection procedure. The ASB method first partitions the item bank into several blocks according to the magnitude of $b$ values, sorts each block according to the magnitude of $a$ values, then forms new strata by grouping items with the same rank order of $a$ across the blocks. Ultimately, the CAT administration is divided into successive stages, proceeding from the stratum with the lowest to highest $a$ values for best results (Hau & Chang, 2001). At any given stage, the next item chosen is the one that maximizes the $b$-matching criterion:

$$B_j(\theta_i) = |\theta_i - b_j|^{-1}. \tag{10}$$

In other words, the item whose difficulty is closest to the interim ability estimate is selected next from the current stratum. Note that $b$-matching is equivalent to MFI for Rasch or 1PLM items (i.e., $a = 1$ and $c = 0$), which is suboptimal for 3PLM items in terms of maximizing information. Nevertheless, by coercing items to be drawn more evenly across the item pool in this way, ASB has been shown to dramatically improve the balance of item exposure with a marginal decrease in the accuracy of $\theta$ estimation (Chang & Ying, 2008).

Therefore, in the interest of stronger test security and better item pool usage, the ASB method was employed in the present investigation of utilizing RTs in the sequential detection of compromised items. Additionally, ability estimation was performed with a combination approach, in which EAP was used as a provisional fail-safe whenever an infeasibility occurred with MLE. In contrast, the original sequential detection study by Zhang (2014) implemented MFI with SH exposure control and exclusive EAP estimation, and the follow-up study by Zhang and Li (2016) used a shadow test engine with all interim estimates in EAP and the final estimates in MLE. The shadow test methodology is not discussed for brevity.

## 3. Sequential Monitoring Procedures

### *3.1. Using Responses*

The goal is to detect a significant increase in the number of correct responses to an item over time, which can be accomplished by periodically comparing the sum of recent responses

to a benchmark value that is expected when the item is not compromised. To this end, define a moving sample to be the most recent $m$ examinees to item $j$, which gradually isolates potentially compromised responses after a leak. The sum of responses in the moving sample is then calculated as,

$$Y_j^{(m)} = \sum_{i=n-m+1}^{n} X_{ij}, \tag{11}$$

where the superscript $(m)$ denotes moving sample, $m$ is the moving sample size, and $n\ (> m)$ is the updated total sample size for item $j$. Under the null hypothesis that the item is not compromised, $X_{ij}$ is a Bernoulli random variable with the following expectation and variance:

$$E(X_{ij}) = P_j(\theta_i), \quad Var(X_{ij}) = P_j(\theta_i)(1 - P_j(\theta_i)). \tag{12}$$

Since $X_{ij}$'s are independently but not identically distributed, $Y_j^{(m)}$ is a Poisson-binomial random variable with the following expectation and variance:

$$E\left(Y_j^{(m)}\right) = \sum_{i=n-m+1}^{n} P_j(\theta_i), \quad Var\left(Y_j^{(m)}\right) = \sum_{i=n-m+1}^{n} P_j(\theta_i)(1 - P_j(\theta_i)). \tag{13}$$

Hence, under the null assumption, the following test statistic has an asymptotic standard normal distribution:

$$\frac{Y_j^{(m)} - \sum_{i=n-m+1}^{n} P_j(\theta_i)}{\sqrt{\sum_{i=n-m+1}^{n} P_j(\theta_i)(1 - P_j(\theta_i))}}, \tag{14}$$

which is the moving average index used by both Han and Hambleton (2004) and Zhang and Li (2016). Noting that $\hat{p}_j^{(m)} = Y_j^{(m)}/m$ is a sample proportion, the test statistic can be equivalently expressed as

$$\frac{\hat{p}_j^{(m)} - \sum_{i=n-m+1}^{n} P_j(\theta_i)/m}{\sqrt{\sum_{i=n-m+1}^{n} P_j(\theta_i)(1 - P_j(\theta_i))/m^2}} \xrightarrow{d} \mathcal{N}(0, 1) \text{ under } H_0, \tag{15}$$

where the null hypothesis, $H_0 : p_j^{(m)} = \sum_{i=n-m+1}^{n} P_j(\theta_i)/m$, is tested against the one-sided alternative hypothesis, $H_1 : p_j^{(m)} > \sum_{i=n-m+1}^{n} P_j(\theta_i)/m$. However, true $\theta_i$ is never known in reality, so Zhang and Li (2016) approximated the test statistic by substituting with $\hat{\theta}_i$. This method was shown to be very powerful, but only when ability estimation was relatively uncorrupted by item preknowledge. As an item pool becomes progressively compromised, an examinee would likely have preknowledge of a greater number of administered items. In effect, $\hat{\theta}_i$'s would become increasingly positively biased, thereby inflating $E(Y_j^{(m)})$ and diminishing the power to detect compromise.

As an alternative approach, Zhang (2014) proposed framing the problem as a comparison of two sample proportions. Specifically, the moving sample is compared to a reference sample,

which is defined as the first $n - m$ examinees to item $j$. The proportion of correct responses in this complementary sample is then computed as

$$\hat{p}_j^{(r)} = \frac{\sum_{i=1}^{n-m} X_{ij}}{n-m}, \tag{16}$$

where the superscript $(r)$ denotes reference sample. $\hat{p}_j^{(r)}$ serves as an appropriate empirical benchmark as long as the item has not been compromised before $n - m$. Thus, the test statistic for two sample proportions can be constructed as

$$\frac{\hat{p}_j^{(m)} - \hat{p}_j^{(r)}}{\sqrt{p_j(1-p_j)\left(\frac{1}{m} + \frac{1}{n-m}\right)}} \xrightarrow{d} \mathcal{N}(0,1) \text{ under } H_0, \tag{17}$$

where $H_0 : p_j^{(m)} = p_j^{(r)}$ is tested against $H_1 : p_j^{(m)} > p_j^{(r)}$. Since the true $p_j$ is unknown, the original study substituted with $\hat{p}_j^{(r)}$ in the denominator to approximate the test statistic. Nevertheless, the present study opts to use the more conventional method of estimating $p_j$ by pooling $\hat{p}_j^{(r)}$ and $\hat{p}_j^{(m)}$ as

$$\hat{p}_j = \frac{(n-m)\hat{p}_j^{(r)} + m\hat{p}_j^{(m)}}{(n-m)+m} = \frac{\sum_{i=1}^{n} X_{ij}}{n}, \tag{18}$$

which is simply the proportion correct out of all $n$ responses for item $j$. Ultimately, the approximated test statistic is given as

$$Z_j = \frac{\hat{p}_j^{(m)} - \hat{p}_j^{(r)}}{\sqrt{\hat{p}_j(1-\hat{p}_j)\left(\frac{1}{m} + \frac{1}{n-m}\right)}} = \frac{\hat{p}_j^{(m)} - \hat{p}_j^{(r)}}{\sqrt{\hat{p}_j(1-\hat{p}_j)/m}}\sqrt{\frac{n-m}{n}}, \tag{19}$$

which is used to conduct the test each time the item is administered to a new examinee by comparing it to a chosen critical value, $z_c$. In other words, if $Z_j > z_c$, then $H_0$ is rejected and the item is flagged as compromised since there is evidence that the number of correct responses has increased significantly. Figure 1 illustrates the sequential process of monitoring an item starting at a predesignated exposure point followed by three possible decision scenarios: (1) type I error of flagging an uncompromised item; (2) correct decision of flagging a compromised item, where the number of exposures from the point of compromise (also known as the change point) to point of flag is called the lag; (3) type II error of failing to flag a compromised item by the end of the CAT cycle.

The choice of $z_c$ depends on the desired rate of type I error, $\alpha$, which is complicated by the fact that many items are each being tested over repeated occasions. In other words, multiplicity occurs both between and within items, resulting in different interpretations of $\alpha$ depending on how we define the "family" of tests for which type I error should be controlled. In the simplest case, a "family" consists of a single monitored item on a single occurrence, so $\alpha$ is the probability of incorrectly flagging a given item on any given exposure. In other words, there is a $100(\alpha)\%$ chance of flagging an uncompromised item every time it is tested. This level of error is easily controlled by setting $z_c = \Phi^{-1}(1-\alpha)$, where $\Phi$ is the standard normal CDF. On the other

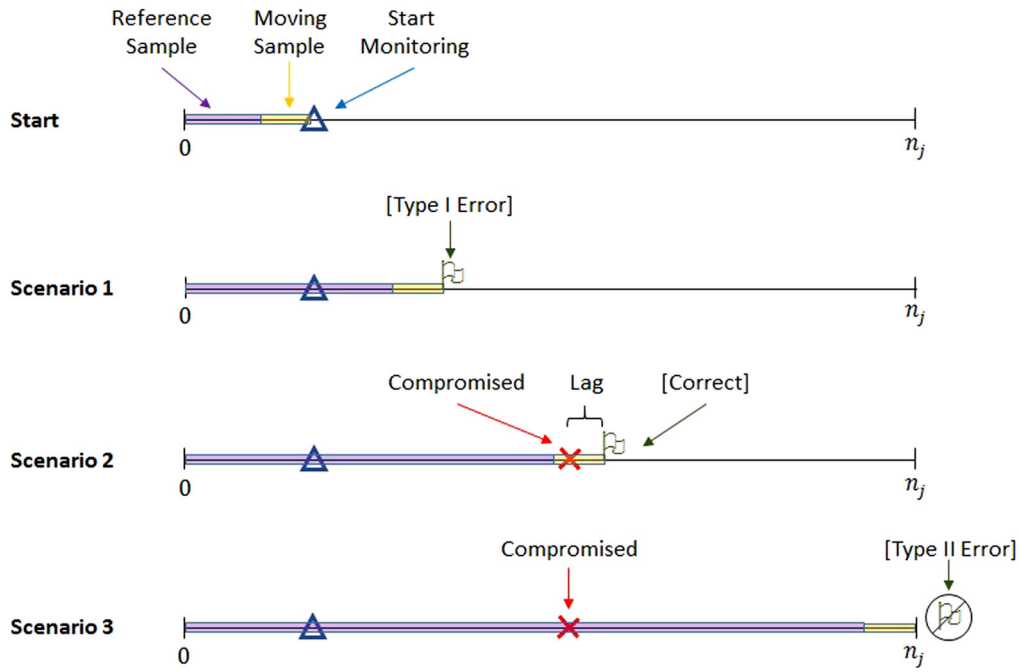


FIGURE 1.
An illustration of the sequential monitoring process with three possible decision scenarios. In all scenarios, the number line represents the sequential exposure count of item $j$, the blue triangle represents the exposure at which the monitoring process starts, and the purple and yellow bars represent the reference and moving samples, respectively. In scenario 1 where the item has not been compromised, the white flag indicates the exposure at which the item was incorrectly flagged (type I error). In scenario 2 where the item has been compromised at the exposure indicated by the red $\times$, the white flag indicates the exposure at which the item was correctly flagged, and the exposure difference between the flag and $\times$ is the lag. In scenario 3 where the item has been compromised at the exposure indicated by the red $\times$, the white flag within the no symbol indicates that the item has been incorrectly missed (type II error) (Color figure online).

extreme, a "family" could be defined as all monitored items on all occurrences, in which case $\alpha$ is the probability of incorrectly flagging at least once across all items and their exposures for the duration of a given CAT cycle. In other words, we can be $100(1-\alpha)\%$ confident that none of items in the bank will be incorrectly flagged. Determining a precise $z_c$ to control for this level of error is much more difficult due to an unknown degree of dependence between items as well as heavy dependence within items without prior knowledge of exposure counts. Note that the strongest dependence within an item occurs on two consecutive tests, since the latter shares all of the same data with the former except for a single new observation added to the moving sample and the oldest observation in the moving sample transferred to the reference sample. In this study, a "family" is defined more moderately as a single monitored item across all occurrences, so $\alpha$ is the probability of incorrectly flagging an item across all of its exposures. In other words, for a given CAT cycle, we are willing to tolerate flagging $100(\alpha)\%$ of uncompromised items in the bank. Lacking more convenient analytic methods, Monte Carlo simulations can be conducted to determine $z_c$ for desired values of $\alpha$.

### *3.2. Using Response Times*

In general, examinees with preknowledge of an item would be expected to respond quicker than usual. Granted, it is theoretically plausible for some crafty cheaters to game the system

by deliberately stalling for time. However, for most operational CATs, this concern is largely immaterial for three reasons. First, this would be a rare occurrence because cheaters would need to be reasonably sure that they have preknowledge of all or most items they will encounter on the test. Of course, this is only possible if somehow the entire item pool was compromised and they managed to memorize everything. Otherwise, for a timed exam, cheaters would be far more inclined to quickly answer the few familiar items to allocate more time for the unfamiliar ones. Second, foolhardy cheaters who attempt this strategy are unlikely to accurately judge how long to delay for a given item before moving on. Most would probably play it safe and not wait too long before moving on. Third, a tiny fraction of highly sophisticated cheaters may slightly hinder the performance of the detection procedures, but the overall impact would probably be trivial.

Thus, the goal is to detect a significant decrease in RTs to an item over repeated administrations, which can be accomplished by periodically comparing the average of recent RTs to a benchmark value that is expected when the item is not compromised. This requires a model of RTs that, at the very least, parameterizes the speededness of individual items across examinees. Among a variety of options, the lognormal model (van der Linden, 2006) remains a popular choice for its relative simplicity and practicability for typical RT data.

Given the latent speed of the $i$th examinee ($\tau_i$), the density function of RT for the $j$th item ($T_{ij}$) is defined as

$$f(t_{ij}|\tau_i) = \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} \mathrm{e}^{-[\alpha_j(\log t_{ij} - \beta_j + \tau_i)]^2/2}, \tag{20}$$

where $\alpha_j$ (not to be confused with type I error rate) and $\beta_j$ are, respectively, the time discrimination and time intensity parameters, and $\tau_i$ and $\beta_j$ are scaled on the same metric. Rewriting the density function in standard form for a lognormal random variable,

$$f(t_{ij}|\tau_i) = \frac{1}{t_{ij}\sqrt{2\pi(1/\alpha_j)^2}} \mathrm{e}^{-[\log t_{ij} - (\beta_j - \tau_i)]^2/[2(1/\alpha_j)^2]}, \tag{21}$$

it becomes clear that $\mu_{ij} = \beta_j - \tau_i$ and $\sigma_j^2 = (1/\alpha_j)^2$. In other words, conditional on examinee speed, the log of RT is normally distributed as follows:

$$\log T_{ij}|\tau_i \sim \mathcal{N}\left[\beta_j - \tau_i, 1/\alpha_j^2\right]. \tag{22}$$

Hence, a moving sample technique is proposed in which the average log RT of the last $m$ examinees for item $j$ is first computed as

$$\hat{\mu}_j^{(m)} = \frac{1}{m} \sum_{i=n-m+1}^{n} \log T_{ij}. \tag{23}$$

The expectation and variance of $\hat{\mu}_j^{(m)}$ under the null are

$$E\left(\hat{\mu}_j^{(m)}\right) = \frac{1}{m} \sum_{i=n-m+1}^{n} (\beta_j - \tau_i), \quad Var\left(\hat{\mu}_j^{(m)}\right) = \frac{1}{m\alpha_j^2}, \tag{24}$$

so the following test statistic can be constructed:

$$\frac{\hat{\mu}_j^{(m)} - \sum_{i=n-m+1}^{n}(\beta_j - \tau_i)/m}{(1/\alpha_j)/\sqrt{m}} \xrightarrow{d} \mathcal{N}(0,1) \text{ under } H_0, \tag{25}$$

where $H_0 : \mu_j^{(m)} = \sum_{i=n-m+1}^{n}(\beta_j - \tau_i)/m$ is tested against $H_1 : \mu_j^{(m)} < \sum_{i=n-m+1}^{n}(\beta_j - \tau_i)/m$. Although $\log T_{ij}$'s are independently but not identically distributed, the asymptotic normality of the test statistic is assured by Lyapunov's CLT (see Appendix for proof). But since the true $\tau_i$'s are unknown, the test statistic could be approximated by substituting with the MLE's of $\tau_i$, which are conveniently calculated as

$$\hat{\tau}_i = \frac{\sum_{j=1}^{K} \alpha_j^2(\beta_j - \log T_{ij})}{\sum_{j=1}^{K} \alpha_j^2} \tag{26}$$

given an examinee's RTs on all $K$ items administered (van der Linden, 2006). Just as with ability, however, speed would be routinely overestimated for those with preknowledge of administered items, thereby reducing the power of the test.

To avoid having to determine specific $\tau_i$'s for each item, a general assumption could be made that $\tau_i$ follows a standard normal distribution for every item $j$. Defining $g(\tau_i)$ to be the standard normal density function, it can be shown that the marginal density of RT for item $j$ is

$$f(t_j) = \int_{-\infty}^{\infty} f(t_j|\tau_i)g(\tau_i)\mathrm{d}\tau_i = \frac{1}{t_j\sqrt{2\pi\left(1 + 1/\alpha_j^2\right)}} \mathrm{e}^{-[\log t_j - \beta_j]^2 \Big/ \left[2\left(1+1/\alpha_j^2\right)\right]}, \tag{27}$$

which is lognormal with $\mu_j = \beta_j$ and $\sigma_j^2 = 1 + 1/\alpha_j^2$. In other words, the marginal distribution of log RT is as follows:

$$\log T_j \sim \mathcal{N}[\beta_j, 1 + 1/\alpha_j^2], \tag{28}$$

which simplifies the null expectation and variance of $\hat{\mu}_j^{(m)}$ to,

$$E\left(\hat{\mu}_j^{(m)}\right) = \beta_j, \quad Var\left(\hat{\mu}_j^{(m)}\right) = \left(1 + 1/\alpha_j^2\right)/m. \tag{29}$$

As a result, the following test statistic can be constructed:

$$\frac{\hat{\mu}_j^{(m)} - \beta_j}{\sqrt{\left(1 + 1/\alpha_j^2\right)/m}} \sim \mathcal{N}(0,1) \text{ under } H_0, \tag{30}$$

where $H_0 : \mu_j^{(m)} = \beta_j$ is tested against $H_1 : \mu_j^{(m)} < \beta_j$. Nevertheless, even if it is true that $\tau_i$ is standard normal in the general population, this convenient formulation only holds when $\theta_i$ and $\tau_i$ are independent. Otherwise, ASB would indirectly influence the distribution of $\tau_i$'s for an

item. For instance, if $\theta_i$ and $\tau_i$ are positively correlated, an item with high $b_j$ would generally be selected for examinees with high $\theta_i$'s and in turn higher $\tau_i$'s. Consequently, $\tau_i$'s for this item would no longer be distributed as standard normal, rendering the above test statistic inaccurate.

Alternatively, an empirical route can be taken in which the moving sample is compared to the reference sample via a two-sample means $t$-test. The mean of log RTs for the reference sample is

$$\hat{\mu}_j^{(r)} = \frac{1}{n-m}\sum_{i=1}^{n-m} \log T_{ij}, \tag{31}$$

and the variances of log RTs for the moving and reference samples are

$$\hat{\sigma}_j^{2(m)} = \frac{\sum_{i=n-m+1}^{n}\left(\log T_{ij} - \hat{\mu}_j^{(m)}\right)^2}{m-1} \quad \text{and} \quad \hat{\sigma}_j^{2(r)} = \frac{\sum_{i=1}^{n-m}\left(\log T_{ij} - \hat{\mu}_j^{(r)}\right)^2}{n-m-1}, \tag{32}$$

respectively. Assuming that $\sigma_j^{2(m)} = \sigma_j^{2(r)}$, the pooled sample variance is

$$\hat{\sigma}_j^2 = \frac{(m-1)\hat{\sigma}_j^{2(m)} + (n-m-1)\hat{\sigma}_j^{2(r)}}{n-2}. \tag{33}$$

Therefore, the test statistic is given as

$$W_j = \frac{\hat{\mu}_j^{(m)} - \hat{\mu}_j^{(r)}}{\sqrt{\hat{\sigma}_j^2\left(\frac{1}{m} + \frac{1}{n-m}\right)}} = \frac{\hat{\mu}_j^{(m)} - \hat{\mu}_j^{(r)}}{\hat{\sigma}_j/\sqrt{m}}\sqrt{\frac{n-m}{m}} \quad \sim \mathcal{T}(n-2) \text{ under } H_0, \tag{34}$$

where $H_0 : \mu_j^{(m)} = \mu_j^{(r)}$ is tested against $H_1 : \mu_j^{(m)} < \mu_j^{(r)}$ each time the item is administered to a new examinee by comparing $W_j$ to a specified critical value, $t_c$. In other words, if $W_j < t_c$, then $H_0$ is rejected and the item is flagged as compromised since there is evidence that the average log RT has dropped significantly. As with $z_c$ when testing proportions, $t_c$ for desired levels of $\alpha$ can be found via Monte Carlo.

### *3.3. Using Responses and Response Times Jointly*

The sequential monitoring of responses and RTs, as described above, can be run concurrently but independently as dual univariate (DU) procedures. Within this scheme, define two ways to deem an item compromised:

DU - 1: Flag item $j$ if $[(Z_j > z_c) \cap (W_j < 0)] \cup [(Z_j > 0) \cap (W_j < t_c)]$;
DU - 2: Flag item $j$ if $(Z_j > z_c) \cap (W_j < t_c)$.

DU-1 presumes that a significant result for either responses or RTs is sufficient evidence for compromise, as long as the insignificant result is in the direction of $H_1$. On the other hand, DU-2 presumes that significant results for both responses and RTs are necessary to make an informed decision. To avoid the complication of having to determine separate critical values for the response and RT processes, the latter can just be set as $t_c = -z_c$.

Alternatively, responses and RTs can be monitored simultaneously within a single multivariate (SM) framework, which accounts for the possible dependence between responses and RTs. Dropping the subscript $j$ to reduce notational clutter, define the following moving sample statistics for item $j$: $\hat{\mu}_1^{(m)} = \hat{p}^{(m)}$ is the mean of responses (i.e., proportion of correct responses), $\hat{\mu}_2^{(m)}$ is the mean of log RTs, $\hat{\sigma}_1^{2(m)}$ is the variance of responses, $\hat{\sigma}_2^{2(m)}$ is the variance of log RTs, and $\hat{\sigma}_{12}^{(m)}$ is the covariance between responses and log RTs. Unbiased estimators are used in all cases, including the sample variance of responses: $\hat{\sigma}_1^{2(m)} = \hat{p}^{(m)}(1 - \hat{p}^{(m)})(m/(m-1))$. Thus, the estimated mean vector and covariance matrix for a moving sample can be specified as

$$\hat{\boldsymbol{\mu}}^{(m)} = \begin{bmatrix} \hat{\mu}_1^{(m)} \\ \hat{\mu}_2^{(m)} \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}^{(m)} = \begin{bmatrix} \hat{\sigma}_1^{2(m)} & \hat{\sigma}_{12}^{(m)} \\ \hat{\sigma}_{12}^{(m)} & \hat{\sigma}_2^{2(m)} \end{bmatrix}, \tag{35}$$

respectively. Likewise, for the reference sample,

$$\hat{\boldsymbol{\mu}}^{(r)} = \begin{bmatrix} \hat{\mu}_1^{(r)} \\ \hat{\mu}_2^{(r)} \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}^{(r)} = \begin{bmatrix} \hat{\sigma}_1^{2(r)} & \hat{\sigma}_{12}^{(r)} \\ \hat{\sigma}_{12}^{(r)} & \hat{\sigma}_2^{2(r)} \end{bmatrix}. \tag{36}$$

Although the joint distribution of responses and RTs is clearly not normal, the asymptotic bivariate normality of the mean vectors can be inferred by the multivariate CLT. Therefore, computing the unbiased pooled covariance matrix as

$$\hat{\boldsymbol{\Sigma}} = \frac{m-1}{n-2}\hat{\boldsymbol{\Sigma}}^{(m)} + \frac{n-m-1}{n-2}\hat{\boldsymbol{\Sigma}}^{(r)}, \tag{37}$$

the two-sample Hotelling's $T^2$ statistic can be constructed as

$$T^2 = \left[\hat{\boldsymbol{\mu}}^{(m)} - \hat{\boldsymbol{\mu}}^{(r)}\right]' \left[\hat{\boldsymbol{\Sigma}}\left(\frac{1}{m} + \frac{1}{n-m}\right)\right]^{-1} \left[\hat{\boldsymbol{\mu}}^{(m)} - \hat{\boldsymbol{\mu}}^{(r)}\right], \tag{38}$$

which is approximately related to the $F$-distribution as follows:

$$F = \frac{n-3}{2(n-2)}T^2 \sim \mathcal{F}(2, n-3) \text{ under } H_0. \tag{39}$$

The null hypothesis, $H_0 : \boldsymbol{\mu}^{(m)} = \boldsymbol{\mu}^{(r)}$, is tested against the directional alternative hypothesis, $H_1 : \mu_1^{(m)} > \mu_1^{(r)} \;\; \& \;\; \mu_2^{(m)} < \mu_2^{(r)}$, after each item exposure until significance is reached. In other words, an item is flagged as compromised if $F > F_c$, provided that $\hat{p}^{(m)} > \hat{p}^{(r)}$ and $\hat{\mu}_2^{(m)} < \hat{\mu}_2^{(r)}$. The imposed constraints ensure that the specific directionality of the test is achieved, and the critical value $F_c$ can be determined for any level of $\alpha$ through Monte Carlo. Note that the conventional Hotelling's $T^2$ test with the non-directional alternative, $H_1 : \boldsymbol{\mu}^{(m)} \neq \boldsymbol{\mu}^{(r)}$, would be inefficient in this context.

## 4. Methods

### *4.1. Data*

The sequential monitoring procedures were evaluated through simulations based on real data from a high-stakes, large-scale standardized CAT. The data consisted of raw responses and RTs (in minutes) from about 2000 examinees with an item pool of about 500 items whose 3PLM parameters were already estimated. The lognormal model parameters were calibrated under the two-level hierarchical framework (van der Linden, 2007), which accounts for the relationship between accuracy and speed. The first level consisted of the 3PL and lognormal models, and the second level specified the covariance structure between the person parameters $(\theta_i, \tau_i)$ and among the item parameters $(a_j, b_j, c_j, \alpha_j, \beta_j)$. Note that this modeling framework disregards the classic within-person speed-accuracy tradeoff, or in other words, the compromise between $\theta_i$ and $\tau_i$ within an individual examinee during the course of the test. Instead, a reasonable assumption is made that an individual's latent parameters remain constant as long as the test is not unduly speeded. Ultimately, $\alpha_j$, $\beta_j$, $\theta_i$, and $\tau_i$ were estimated using a modified version of van der Linden's (2008) MCMC routine that fixed $a_j$, $b_j$, and $c_j$ to the pre-calibrated values and centered the distribution of $\tau_i$ at 0. Using 10,000 MCMC draws with a burn-in size of 5000, trace plots with multiple chains displayed rapid mixing for all estimated parameters. All converged estimates from this calibration step were regarded as the true parameter values when simulating CAT. Note that as a consequence of disregarding errors in the item parameter estimates in the simulation, the absolute performances of the detection methods may be overly optimistic. Nevertheless, the primary purpose of this study is to compare their relative performances, which should not be affected.

### *4.2. Simulation Design*

The CAT system was built upon the ASB item selection algorithm, with the item pool divided into 5 strata of about 100 items each. Fixing the test length at 30 items, the first 5 were chosen randomly from each stratum in order to calculate initial estimates of $\theta_i$ and $\tau_i$, then subsequent items were selected using the $b$-matching criterion at each of the five stages. Additionally, the maximum exposure rate was set at 0.2 to ensure a relatively balanced usage of items even under extreme simulation conditions. Based on the true parameters, the $i$th examinee's response to the $j$th administered item was randomly generated in real time from the Bernoulli distribution with success probability $P_j(\theta_i)$; likewise, response time was randomly generated from $\log \mathcal{N}(\beta_j - \tau_i, 1/\alpha_j^2)$.

There are two broad manifestations of item compromise in CAT: (1) a simple situation in which a set of items leak to the general public, thereby giving any test-taker an opportunity to gain preknowledge of any leaked item (e.g., overexposed items spreading through word of mouth or discussions in online forums); and (2) a more complex situation in which one or more subsets of examinees gain preknowledge of different subsets of the item pool (e.g., groups of colluders sharing stolen items). Realistically, it is possible for either or both forms of compromise to transpire during a CAT cycle. For the purposes of this study, however, simulating just the simple scheme was deemed adequate to evaluate the performances of the detection methods without loss of generality. This is because the methods are only trying to detect a suspicious change in the response and RT processes for an item, irrespective of the mechanism of compromise and pattern of preknowledge. In other words, it is largely irrelevant how the item got disclosed or who knows the answers to which items. The only thing of consequence for the monitoring procedures is that, for any given item, there is a sudden increase in the number of correct responses and/or decrease in average RTs after a certain point, however that may have occurred.

Formally, every examinee was assumed to be a potential beneficiary of a compromised item with stationary probability $\psi$. In other words, $\psi$ is the probability of any examinee having pre-

knowledge of any given compromised item:

$$\psi = P(\text{preknowledge} \mid \text{compromised}). \tag{40}$$

Thus, the preknowledge distribution of responses to any compromised item was modeled as

$$P^*(X = x) = 0.999^x \cdot 0.001^{(1-x)} \quad \Leftrightarrow \quad X \sim \text{Bernoulli}(0.999), \tag{41}$$

which specifies a correct response with near but not absolute certainty to allow for inadvertent mistakes by even those with preknowledge. Also, the preknowledge distribution of RTs (in minutes) on any compromised item was modeled as

$$f^*(t_{ij}) = \frac{3.5}{t_{ij}\sqrt{2\pi}} \mathrm{e}^{-3.5^2(\log t_{ij}+2)/2} \quad \Leftrightarrow \quad \log T \sim \mathcal{N}(-2, 1/3.5^2), \tag{42}$$

which specifies a reasonable range from about 2 to 30 s with a mean of about 8.5 s. Therefore, responses and RTs to an item, from the point of compromise onward, follow the preknowledge distributions with probability $\psi$ and the regular distributions with probability $1 - \psi$, which can be expressed in terms of mixture distributions as follows:

$$\widetilde{P}_j(\theta_i) = \psi P^*(X_{ij} = 1) + (1 - \psi) P_j(\theta_i), \tag{43}$$

$$\widetilde{f}_j(t_{ij}|\tau_i) = \psi f^*(t_{ij}) + (1 - \psi) f_j(t_{ij}|\tau_i). \tag{44}$$

The monitoring process was set to start for every item at the 40th exposure using a moving sample size of $m$. For instance, using $m = 10$, the moving and reference samples of the initial test would consist of the last 10 and first 30 examinees to have been administered the item, respectively. A random quarter of the item pool (about 125 items) were queued to be compromised, each starting at a randomized exposure count between 40 and 100. Any examinee administered a compromised item had preknowledge with a designated probability of $\psi$. Defining $C$ as the set of all compromised items and $F$ as the set of all flagged items, type I error rate and power were estimated as

$$P(\text{Type I Error}) \approx P(F|C') = \frac{P(F \cap C')}{P(C')} = \frac{|F \cap C'|}{|C'|}, \tag{45}$$

$$\text{Power} \approx P(F|C) = \frac{P(F \cap C)}{P(C)} = \frac{|F \cap C|}{|C|}. \tag{46}$$

If an item in $C$ was prematurely flagged before the designated change point, it was moved to the uncompromised set $C'$ and counted as a type I error. Any flagged item, whether or not in error, was recorded but otherwise kept operational in the item pool. Additionally, the average lag $\bar{L}$ from the change point $l_j$ to flag point $n_j$ for the set of correctly flagged items $(F \cap C)$ was calculated as

$$\bar{L} = \frac{\sum_{j \in F \cap C}(n_j - l_j)}{|F \cap C|} \tag{47}$$

to evaluate how quickly compromised items could be detected on average. The performances of the sequential monitoring procedures were comparatively evaluated on these three criteria instead

of the average run length (ARL) that is commonly utilized in conventional change-point detection. The reason is simply that ARL assumes that the sequential process continues ad infinitum until a significant change is detected, which is clearly not the case in CAT due to a finite number of examinees. Higher $\psi$ is expected to yield greater power at a given type I error rate, since a higher prevalence of preknowledge makes it easier to detect. Likewise, smaller $m$ is expected to yield shorter lag at a given type I error rate, since a smaller moving sample retains less older data that may act as dead weight.

## 5. Results

The first set of simulations compared the performances of the five monitoring schemes: responses alone (R), RTs alone (T), dual univariate 1 (DU-1), dual univariate 2 (DU-2), and single multivariate (SM). Every technique was evaluated on each of 2 sample sizes ($m = 5, 20$) at each of 3 preknowledge probabilities ($\psi = 0.15, 0.25, 0.35$) for a total of 6 conditions. The results, which were averaged across 100 replications, are presented as receiver operating characteristic (ROC) curves in Fig. 2 and lag plots in Fig. 3. Note that the results are only shown up to a type I error rate of 0.1, since anything beyond that is generally unacceptable in practice. The most salient observation is that the performances of T, DU-1, and SM were all nearly identical with the highest power and lowest lag at any given type I error rate. On the contrary, R was worst by far and DU-2 was somewhere in the middle in terms of general performance. In other words, R and T were effectively the lower and upper performance baselines, respectively, indicating that preknowledge RT's were much easier to detect than preknowledge responses. Consequently, DU-1 and SM were overwhelmingly dominated by RTs, while DU-2 was evenly influenced by both responses and RTs. Moreover, for every procedure, lag was shorter for higher $\psi$ and smaller $m$, and power was greater for higher $\psi$ regardless of $m$ as expected. However, a closer look at the ROC curves reveals an interesting pattern: Power was greater for larger $m$ at $\psi = 0.35$, very similar for both $m = 5$ and 20 at $\psi = 0.25$, and actually greater for smaller $m$ at $\psi = 0.15$. This suggested an interaction between $\psi$ and $m$, which warranted a follow-up study.

The second set of simulations compared the performances of 5 moving sample sizes ($m = 2, 5, 10, 20, 30$) at each of 6 preknowledge probabilities ($\psi = 0.05, 0.10, 0.15, 0.25, 0.35, 0.45$) exclusively for SM. As before, the results were averaged across 100 replications and presented as ROC curves in Fig. 4 and lag plots in Fig. 5. The particular interaction effect becomes quite noticeable here: $\psi$ strongly moderated the effect of $m$ on power at any given type I error rate. For $\psi < 0.25$, smaller $m$ resulted in greater power, with larger differences in effect for lower $\psi$; at $\psi = 0.25$, $m$ had no appreciable effect on power; for $\psi > 0.25$, larger $m$ resulted in greater power, with larger differences in effect for higher $\psi$. This phenomenon occurs because when $\psi$ is very low, there is a dearth of preknowledge responses and RTs. As a result, a smaller moving sample can more easily isolate them, thereby increasing power even at the cost of larger sampling error. In the current context, $\psi = 0.25$ happened to be the point of equilibrium at which the opposing forces of preknowledge isolation and sampling error balanced out to the same power for every $m$. Also, for $\psi > 0.25$, there were negligible improvements in power for $m$ greater 5, most likely due to the ceiling effect. On the other hand, moderator effects were not observed for lag. Just as in the earlier results, lag was always shorter for smaller $m$ and higher $\psi$.

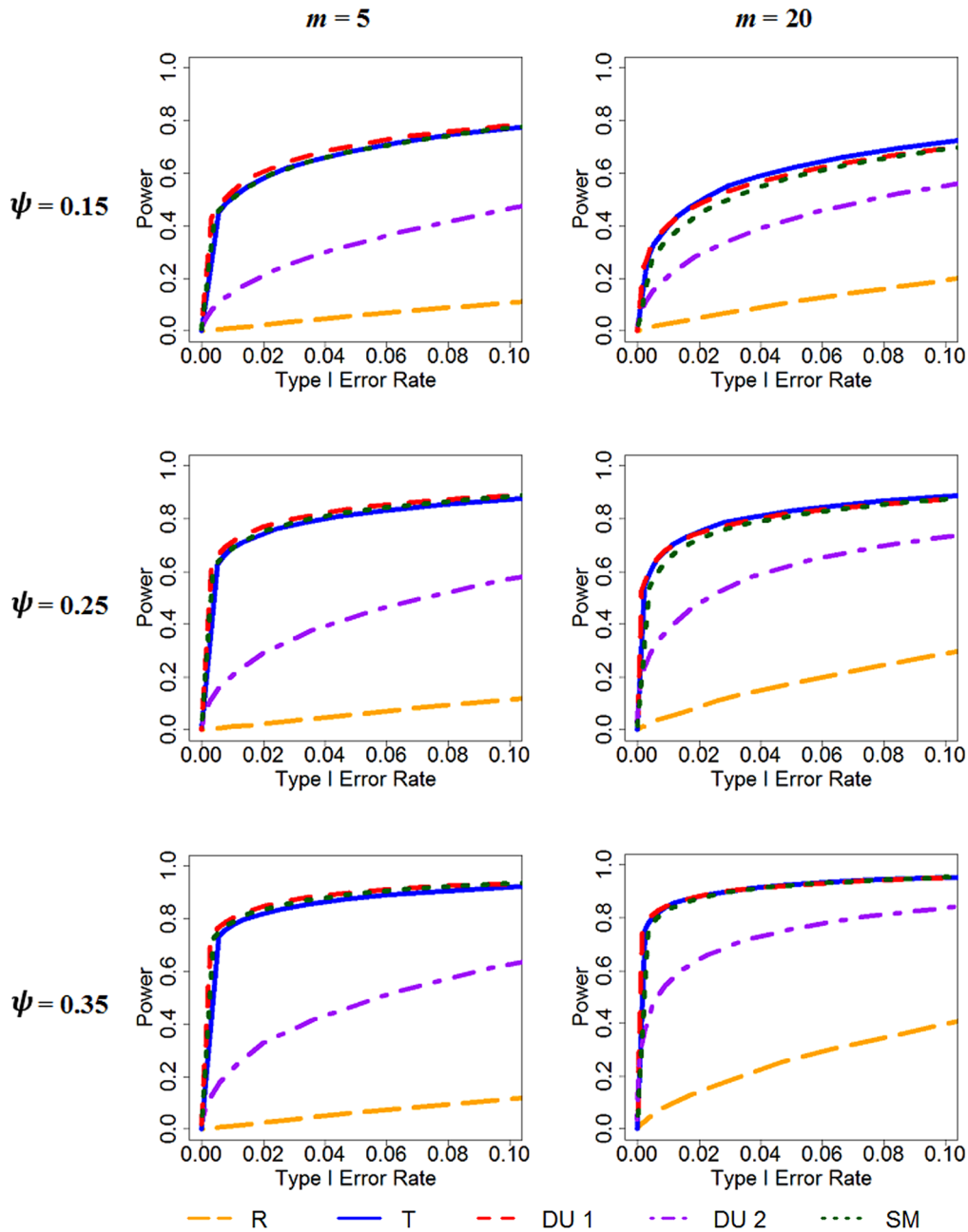


FIGURE 2.
ROC curves for each of the five sequential procedures (R, T, DU-1, DU-2, SM) across six conditions ($\mathbf{m} = \{5, 20\} \times \boldsymbol{\psi} = \{0.15, 0.25, 0.35\}$). Results are averaged across 100 replications with about 500 items and 2000 examinees.

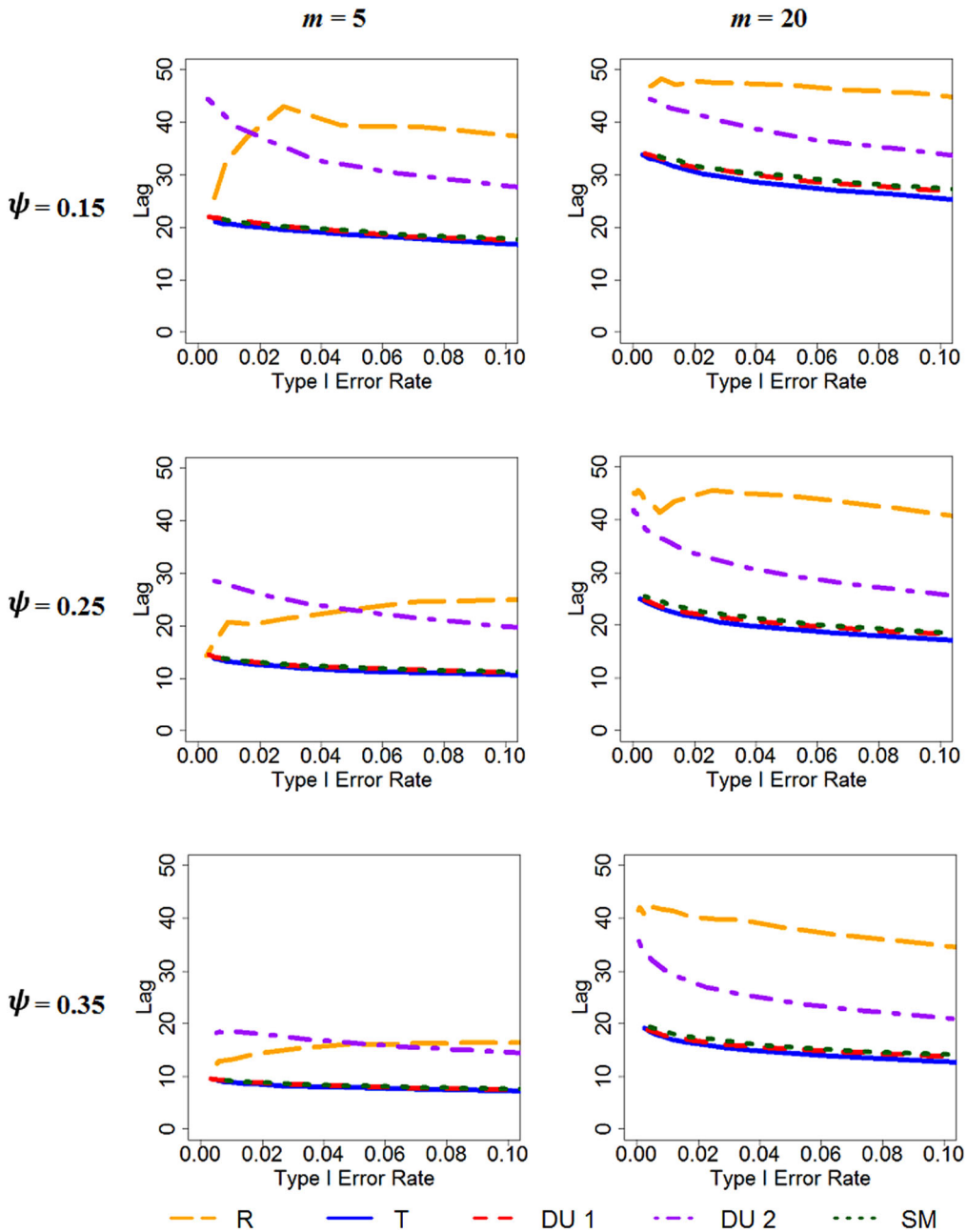


FIGURE 3.
Lag plots for each of the five sequential procedures (R, T, DU-1, DU-2, SM) across six conditions ($\mathbf{m} = \{5, 20\} \times \boldsymbol{\psi} = \{0.15, 0.25, 0.35\}$). Results are averaged across 100 replications with about 500 items and 2000 examinees.

## 6. Discussion

Optimistic simulation conditions notwithstanding, the results demonstrate that response times can be effectively utilized in conjunction with responses to improve the sequential detection of

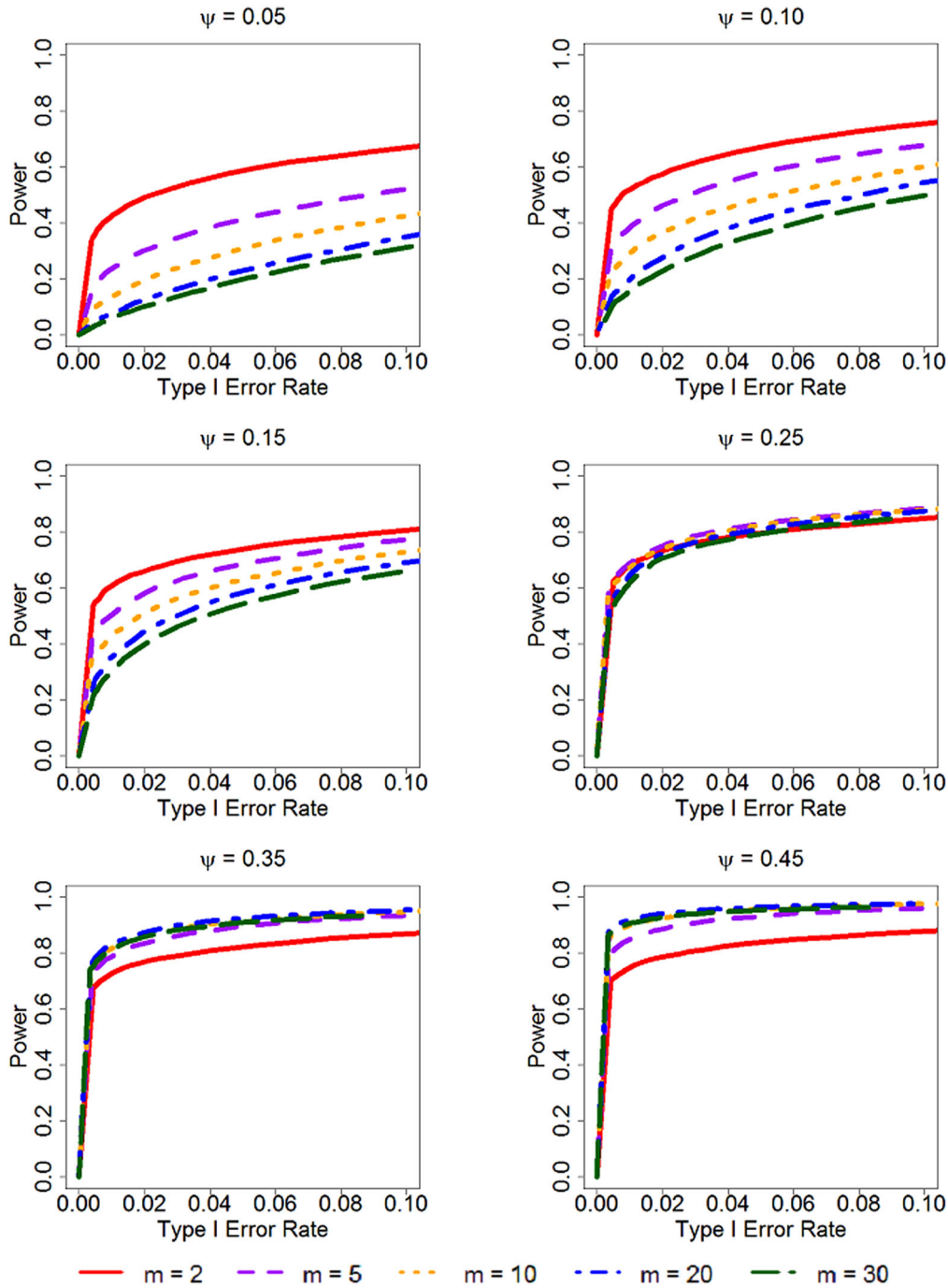


FIGURE 4.
ROC curves for the SM procedure with five moving sample sizes ($m = 2, 5, 10, 20, 30$) at each of six levels of item preknowledge ($\psi = 0.05, 0.10, 0.15, 0.25, 0.35, 0.45$). Results are averaged across 100 replications with about 500 items and 2000 examinees.

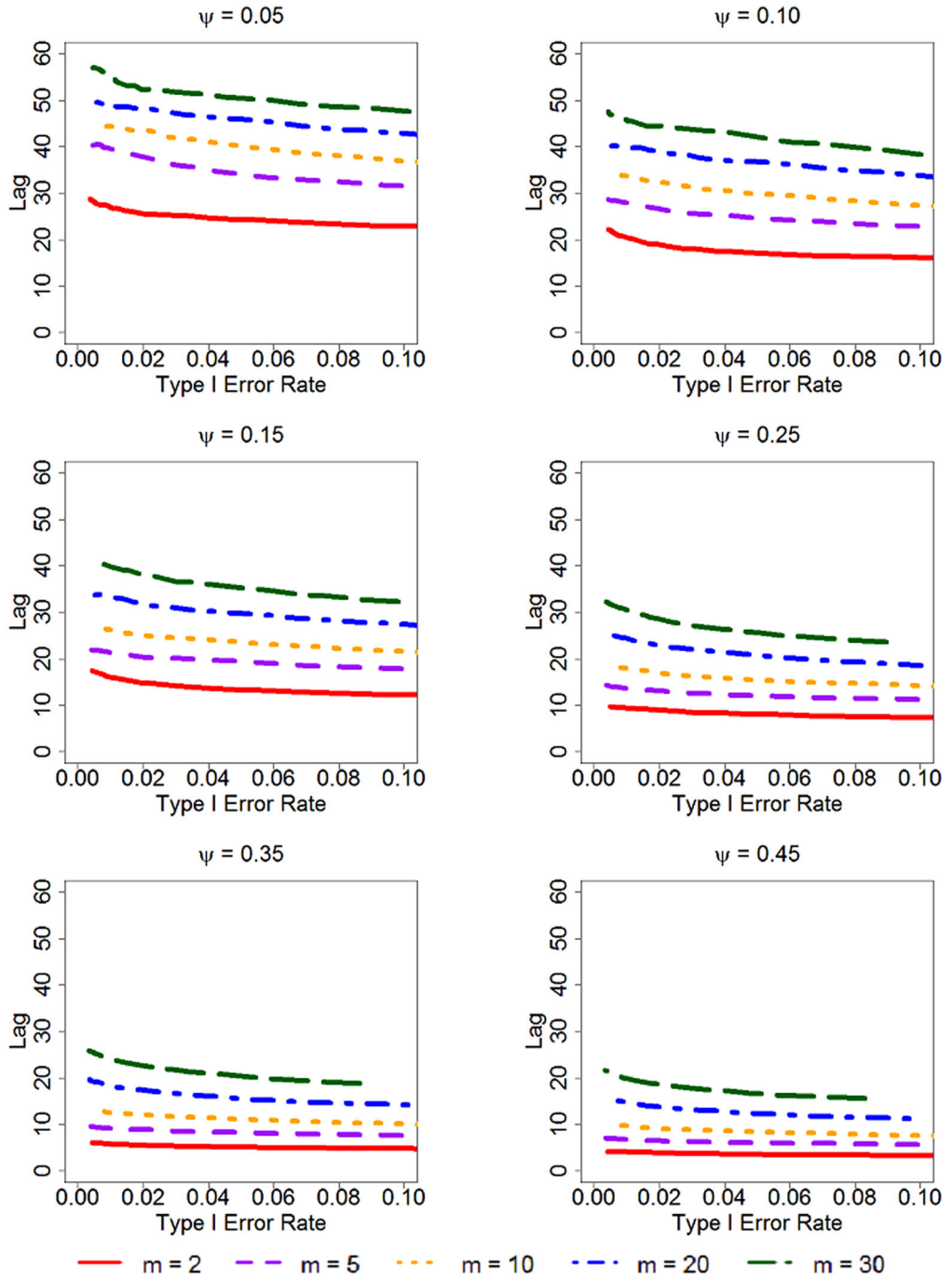


FIGURE 5.
Lag plots for the SM procedure with five moving sample sizes ($m = 2, 5, 10, 20, 30$) at each of six levels of item preknowledge ($\psi = 0.05, 0.10, 0.15, 0.25, 0.35, 0.45$). Results are averaged across 100 replications with about 500 items and 2000 examinees.

compromised items. Both DU-1 and SM were shown to be equally superior over DU-2 in detection accuracy and speed. Nevertheless, SM has two distinct advantages over DU-1: First, SM is easier to implement since only a single process needs to be tracked as opposed to two separate streams. Second, SM can be seen as a more holistic approach that combines all information into a single evidentiary criterion instead of cherry-picking the favorable outcome. Choosing an appropriate moving sample size is a trickier matter, since it depends on the unknown probability that a random examinee has preknowledge of any given compromised item. Because the optimal $m$ is most likely unique for every CAT, it must be determined by the user through a series of simulations. This can be accomplished by first finding the equilibrium point, $\psi_e$. If true $\psi$ is believed to be less than $\psi_e$, use $m = 2$ for best results; otherwise, choose the largest $m$ beyond which there seem to be insubstantial improvements in power. Once $m$ is determined, an item can be monitored as soon as $n = m + 2$. For future study, it may be interesting to allow the moving sample size to vary over the course of the monitoring process, for instance, in inverse proportion to the item exposure count.

At this point, a word of caution regarding the interpretation of power would be prudent. It may be tempting to interpret power as the probability that a flagged item is compromised, or $P(C|F)$, which would be of primary interest in practice. However, doing so would be committing an inverse fallacy, recalling that power is actually the probability that a compromised item is flagged, or $P(F|C)$. Succinctly, power $= P(F|C) \neq P(C|F)$; instead, we properly apply Bayes' theorem to obtain

$$P(C|F) = \frac{P(F|C)P(C)}{P(F|C)P(C) + P(F|C')P(C')} = \frac{\text{Power} \times P(C)}{[\text{Power} \times P(C)] + [\alpha \times (1 - P(C))]}. \quad (48)$$

Note that $P(C)$ is the base rate of item compromise, which is typically unknown. Nevertheless, to illustrate the substantial impact of the base rate, say we have 90% power at 5% type I error, but the base rate is relatively low at 5.5%. Then, there is only about a 50% chance that a flagged item is actually compromised even with such high power. Although somewhat discouraging, this is a typical phenomenon in diagnostic testing in general, such as in medical screening for a rare disease. As with any such tool, the sequential detection procedures should be utilized responsibly, preferably with corroborating evidence of compromise. On a related note, it would be unwise to settle on a high type I error rate in efforts to afford greater power. Over-flagging and subsequently deactivating perfectly usable items would be a supreme waste, considering that items are enormously time-consuming and costly to develop.

There are several issues that have not been explicitly accounted for in this study. First, the particular lognormal distribution used to model preknowledge RTs is certainly plausible and suitable for the purposes of this study, but it is admittedly an uninformed choice. Currently, no empirically supported alternatives have been proposed in literature, most likely due to the paucity of real RT data from examinees verified to have item preknowledge. Second, sequential monitoring assumes that the general characteristics of the examinee population are consistent over the course of item usage. Hence, the simulations did not consider scenarios of drastic changes in response patterns due to reasons unrelated to item compromise, such as a sudden shift in the demographics of test-takers. Third, for a given CAT window, the probability of item preknowledge ($\psi$) was assumed to be constant across all compromised items and test-takers; moreover, all those with preknowledge were simplistically simulated to respond correctly with near certainty (99.9%). In actuality, these probabilities are likely to vary for each examinee-item pair. Fourth, the dependence between ability and speed may play a role in the performance of the joint detection procedures. The correlation between $\theta$ and $\tau$ was 0.77 for the empirical data at hand, but the impact of varying degrees of association on power and type I error warrants further investigation. Fifth, information regarding non-statistical considerations, such as content balancing and avoiding enemy items,

was unavailable for the item pool used in this study. Test constraints add a complex dimension to item selection and exposure control strategies, which may affect the performance of the detection methods. Acknowledging all of these various limitations, it would be worthwhile to extend the models and simulations to reflect the more complex reality. Moreover, empirical studies need to be conducted to assess the applicability and efficacy of the proposed procedures in practice.

Lastly, development of more powerful tests and efficient techniques are currently underway. For one thing, the classic Hotelling's $T^2$ statistic used in the SM procedure may not be the most appropriate choice, given that the mean vector is not bivariate normal. Especially at the beginning of a CAT cycle when there are relatively few examinees, there may exist more suitable nonparametric tests that can afford greater power. Also, the presented methods only monitor the sample means of responses and response times. For RTs in particular, the variance is most likely smaller as well for compromised items. Thus, testing the difference in the empirical distribution functions (EDF) between the moving and reference samples, using the Kolmogorov–Smirnov or Kuiper's tests for instance, might prove to be more powerful since both the location and spread of the sample distributions are taken into account. Additionally, the detection problem can be approached from a more traditional sequential analysis framework. More specifically, changes in response patterns could be monitored via control charts such as CUSUM, or the sequential hypothesis testing procedure could be framed as a series of generalized likelihood ratio tests (GLRT). The feasibility of these methods remains to be seen.

## Appendix

*Application of Lyapunov's Central Limit Theorem*

Assume that log RT is normally distributed as follows: $\log T_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_j^2)$, where $\mu_{ij} = \beta_j - \tau_i$ and $\sigma_j^2 = 1/\alpha_j^2$. The mean log RT of the moving sample for item $j$ is then given as $\hat{\mu}_j^{(m)} = \frac{1}{m}\sum_{i=n-m+1}^{n} \log T_{ij}$. Also, define the following: $s_m^2 = \sum_{i=n-m+1}^{n} \sigma_j^2 = m\sigma_j^2$. In this context, Lyapunov's CLT states that

$$\frac{1}{s_m}\sum_{i=n-m+1}^{n}(\log T_{ij} - \mu_{ij}) = \frac{\hat{\mu}_j^{(m)} - \sum_{i=n-m+1}^{n}\mu_{ij}/m}{\sigma_j/\sqrt{m}} \xrightarrow{d} \mathcal{N}(0,1) \tag{A.1}$$

if, for any $\delta > 0$, the following condition is met:

$$\lim_{m\to\infty}\frac{1}{s_m^{2+\delta}}\sum_{i=n-m+1}^{n} E\left(|\log T_{ij} - \mu_{ij}|^{2+\delta}\right) = 0. \tag{A.2}$$

Recognizing that the expectation term is a central absolute moment of $\log T_{ij}$,

$$E\left(|\log T_{ij} - \mu_{ij}|^{2+\delta}\right) = \sigma_j^{2+\delta}(1+\delta)!! \cdot \begin{cases} \sqrt{2/\pi} & \text{if } 2+\delta \text{ is odd} \\ 1 & \text{if } 2+\delta \text{ is even} \end{cases}. \tag{A.3}$$

Therefore, using $\delta = 2$ for simplicity,

$$\begin{aligned}\lim_{m\to\infty} \frac{1}{s_m^4} \sum_{i=n-m+1}^{n} E\left(|\log T_{ij} - \mu_{ij}|^4\right) &= \lim_{m\to\infty} \frac{1}{m^2\sigma_j^4} \sum_{i=n-m+1}^{n} 3\sigma_j^4 \\ &= \lim_{m\to\infty} \frac{m\left(3\sigma_j^4\right)}{m^2\sigma_j^4} \\ &= \lim_{m\to\infty} \frac{3}{m} \\ &= 0,\end{aligned}$$

thereby meeting Lyapunov's condition for the asymptotic normality of the test statistic.

## References


Armstrong, R. D., & Shi, M. (2009). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement*, *33*, 391–410.

Armstrong, R. D., Stoumbos, Z. G., Kung, M. T., & Shi, M. (2007). On the performance of the lz person-fit statistic. *Practical Assessment Research and Evaluation*, *12*(16).

Belov, D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing*, *2*, 37–58.

Belov, D. I. (2015). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, *40*, 83–97.

Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback–Leibler divergence and K-Index. *Applied Psychological Measurement*, *34*, 379–392.

Belov, D. I., Pashley, P. J., Lewis, C., & Armstrong, R. D. (2007). Detecting aberrant responses with Kullback–Leibler distance. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 7–14). Tokyo: Universal Academy Press.

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. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*, 1–20.

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., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 37–52.

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.

Chang, H.-H., & Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, *37*, 1466–1488.

Chang, S. W., Ansley, T. N., & Lin, S. H. (2000). Performance of item exposure control methods in computerized adaptive testing: Further explorations. In *Paper presented at the annual meeting of the American Educational Research Association*, New Orleans, LA.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67–86.

Egberink, I., Meijer, R. R., Veldkamp, B. P., Schakel, L., & Smid, N. G. (2010). Detection of aberrant item score patterns in computerized adaptive testing: An empirical example using the CUSUM. *Personality and Individual Differences*, *48*, 921–925.

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*.

Han, N., & Hambleton, R. (2004). Detecting exposed test items in computer-based testing. In *Paper presented at the annual meeting of the National Council on Measurement in Education*, San Diego, CA.

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.

Impara, J. C., & Kingsbury, G. (2005). Detecting cheating in computer adaptive tests using data forensics. In *Paper presented at the annual meeting of the National Council on Measurement in Education*, Montreal, Cananda.

Kang, H.-A., & Chang, H.-H. (2016). Online detection of item compromise in CAT using responses and response times. In *Paper presented at the annual meeting of the National Council on Measurement in Education*, Washington, D.C.
Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*, 277–298.
Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*, 359–375.
Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, *53*, 161–176.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of mental test scores*. Reading, MA: Addison-Wesley.
Lu, Y., & Hambleton, R. (2003). Statistics for detecting disclosed items in a CAT environment (Research Report No. 498). Amherst, MA: University of Massachusetts, School of Education, Center for Educational Assessment.
Marianti, S., Fox, J.-P., Marianna, A., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*, 426–451.
Mavridis, D., & Moustaki, I. (2008). Detecting outliers in factor analysis using the forward search algorithm. *Multivariate Behavioral Research*, *43*, 435–475.
Mavridis, D., & Moustaki, I. (2009). The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data. *Journal of Computational and Graphical Statistics*, *18*, 1016–1034.
McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement*, *23*, 147–160.
McLeod, L. D., & Schnipke, D. L. (1999). Detecting items that have been memorized in the computerized adaptive testing environment. In *Paper presented at the annual meeting of the National Council on Measurement in Education*, Montreal, Canada.
Meijer, R. R. (2002). Outlier detection in high-stakes certification testing. *Journal of Educational Measurement*, *39*, 219–233.
Meijer, R. R., & Sotaridona, L. S. (2006). Detection of advance item knowledge using response times in computer adaptive testing. Technical Report 03-03, Law School Admission Council.
Mislevy, R. J., & Chang, H.-H. (2000). Does adaptive testing violate local independence? *Psychometrika*, *65*, 149–156.
Moustaki, I., & Knott, M. (2014). Latent variable models that account for atypical responses. *Journal of the Royal Statistical Society, Series C*, *63*, 343–360.
O'Leary, L. S., & Smith, R. W. (2017). Detecting candidate preknowledge and compromised content using differential person and item functioning. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 151–163). New York, NY: Routledge.
Öztürk, N. K., & Karabatsos, G. (2017). A Bayesian robust IRT outlier-detection model. *Applied Psychological Measurement*, *41*, 195–208.
Risk, N. M. (2015). *The impact of item parameter drift in computer adaptive testing (CAT)* (Unpublished doctoral dissertation). University of Illinois at Chicago.
Stocking, M. L. (1993). Controlling item exposure rates in a realistic adaptive testing paradigm. *ETS Research Report Series* (pp. 1–31).
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). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association*, San Diego, CA: Navy Personnel Research and Development Center.
Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, *49*, 95–110.
Tendeiro, J. N., & Meijer, R. R. (2012). A CUSUM to detect person misfit: A discussion and some alternative for existing procedures. *Applied Psychological Measurement*, *36*, 420–442.
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., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384.
van der Linden, W. J., & Lewis, C. (2015). Bayesian checks on cheating on tests. *Psychometrika*, *80*, 689–706.
van der Linden, W. J., & van Krimpen-Stoop, E. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*, 251–265.
van Krimpen-Stoop, E., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, *26*, 199–218.
Veerkamp, W. J. J., & Glas, C. A. W. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, *25*, 373–389.
Zhang, J. (2014). A sequential procedure for detecting compromised items in the item pool of a CAT system. *Applied Psychological Measurement*, *38*, 87–104.
Zhang, J., & Li, J. (2016). Monitoring items in real time to enhance CAT security. *Journal of Educational Measurement*, *53*, 131–151.

Zhu, R., Yu, F., & Liu, S. (2002). Statistical indexes for monitoring item behavior under computer adaptive testing environment. In: *Paper presented at the annual meeting of the American Educational Research Association*, New Orleans, LA.

*Manuscript Received: 11 JAN 2017*
*Published Online Date: 22 NOV 2017*