Article

Journal of Educational and Behavioral Statistics
Vol. XX, No. X, pp. 1–31
DOI: 10.3102/10769986211059085

Item Pool Quality Control in Educational Testing: Change Point Model, Compound Risk, and Sequential Detection

Yunxiao Chen
London School of Economics and Political Science, United Kingdom

Yi-Hsuan Lee
Educational Testing Service, Princeton, NJ, USA

Xiaoou Li
University of Minnesota Twin Cities, Minneapolis, MN, USA

In standardized educational testing, test items are reused in multiple test administrations. To ensure the validity of test scores, the psychometric properties of items should remain unchanged over time. In this article, we consider the sequential monitoring of test items, in particular, the detection of abrupt changes to their psychometric properties, where a change can be caused by, for example, leakage of the item or change of the corresponding curriculum. We propose a statistical framework for the detection of abrupt changes in individual items. This framework consists of (1) a multistream Bayesian change point model describing sequential changes in items, (2) a compound risk function quantifying the risk in sequential decisions, and (3) sequential decision rules that control the compound risk. Throughout the sequential decision process, the proposed decision rule balances the trade-off between two sources of errors, the false detection of prechange items, and the nondetection of postchange items. An item-specific monitoring statistic is proposed based on an item response theory model that eliminates the confounding from the examinee population which changes over time. Sequential decision rules and their theoretical properties are developed under two settings: the oracle setting where the Bayesian change point model is completely known and a more realistic setting where some parameters of the model are unknown. Simulation studies are conducted under settings that mimic real operational tests.

Keywords: *standardized testing; test security; item preknowledge; sequential change point detection; multistream data; compound decision*

1. Introduction

The administration of a standardized educational test typically relies on an item pool, where items are repeatedly chosen from the pool to assemble test forms. To maintain the validity and reliability of a standardized test over time, it is important to ensure that the psychometric properties of items in the pool remain unchanged. An item may need to be revised or removed from the pool once its psychometric properties encounter a significant change, which may be caused by various reasons such as its leakage to the public or change of the corresponding curriculum. An important and challenging problem that test administrators face is to periodically review their testing data and detect the changed items as early as possible (see Chapter 4, American Educational Research Association et al., 2014).

Following the discussion in Lee and Haberman (2021), we divide educational tests into two categories—tests with infrequent and frequent testing schedules. Infrequent tests include college admission tests like American College Test (ACT) and Scholastic Assessment Test (SAT) and survey assessments like the National Assessment of Educational Porgress and the Programme for International Student Assessment, where ACT and SAT deliver seven administrations per year and survey assessments are typically delivered once every a few years. Frequent tests include both continuous tests that are delivered daily and frequent but noncontinuous tests that are available once or more times per week. Examples of continuous tests include the Graduate Record Examinations general test and the Praxis Core Academic Skills for Educators tests that follow the multistage testing (MST; Yan et al., 2016) and fixed-form testing designs, respectively. Frequent but noncontinuous tests are also very common. For example, the Test of English for International Communication speaking test had 281 administrations in 3 years (Qu et al., 2017), and another assessment of English proficiency had 498 administrations in 6 years (Qian & Li, 2021), both of which are fix-form tests. This article considers item pool monitoring for frequent tests, in which items are reused more frequently and thus are more likely to be leaked. In particular, we focus on frequent but noncontinuous tests with a fixed-form design. Generalization to other frequent test settings is also discussed.

To tackle this problem, we adopt a multistream sequential change point formulation. Each item corresponds to a data stream, for which data are collected sequentially from test administrations over time. Each data stream is associated with its own change point. The change point corresponds to a distributional change in some monitoring statistics which reflect certain psychometric properties of the item. That is, the monitoring statistics follow one distribution at any time before the change point and follow a different distribution after the change point. Roughly speaking, our goal is to detect as many postchange items as possible at each time point, without making too many false detections of prechange items. The detected items will be reviewed by the test developers to check their validity. Further actions, such as removing items, may follow.

To provide a sensible solution to this change detection problem for item quality control, we propose a statistical decision framework. This framework consists of (1) a multistream Bayesian change point model describing the data streams with change points, (2) a compound risk function quantifying the risk in sequential decisions, and (3) sequential decision rules that aim at detecting as many postchange items as possible while controlling the compound risk to be below a prespecified level. Specifically, our risk function can be viewed as a measure of the proportion of postchange items among the undetected items given all the up-to-date information from the monitoring statistics, where the information will be formalized by an information filtration (see Section 2.1 for the definition). We emphasize that this risk function measures the overall item-pool-level risk, instead of item-level risk. It is thus suitable for the purpose of controlling item pool quality as a whole. The quality of undetected items can be guaranteed by controlling their compound risk. Consequently, only the detected items need a validity check. Our development considers two different settings, including an oracle setting for which the Bayesian change point model is completely known and a more realistic setting where only partial information is available about the model.

The current development is a significant extension of Y. Chen and Li (2021), where the compound sequential change detection framework is first proposed. First, a more general model is considered that is more suitable for item quality control. Specifically, it takes into account item exposure and addition of new items, two important features of test administration and maintenance. Second, we extend the development in Y. Chen and Li (2021) to a more realistic setting when only partial information is available about the Bayesian change point model. A change detection procedure is proposed that is shown to control the compound risk. Third, a monitoring statistic is proposed based on an item response theory (IRT) model. This statistic adjusts for confounding from examinee population changes, so that changes in item properties can be better detected. Finally, simulation studies are conducted under settings that mimic the administrations of operational tests.

The proposed framework is closely related to, but substantially different from, the classical sequential change detection problem for a single data stream (Page, 1954; Roberts, 1966; Shewhart, 1931; Shiryaev, 1963), as well as recent developments on change detection for multiple streams (e.g., Mei, 2010; Chan, 2017; H. Chen, 2019; H. Chen & Zhang, 2015; J. Chen et al., 2020; Xie & Siegmund, 2013). The major difference is that the existing works, except for Y. Chen and Li (2021) and J. Chen et al. (2020), consider the detection of a single change point, even with multistream data. Consequently, they do not handle a compound risk that aggregates information on the change points of different data streams.

This framework also closely connects to compound decision theory (see, e.g., C.-H. Zhang, 2003) which dates back to the seminal works of Robbins (1951, 1956). Specifically, the compound risk that we control at each time point can be viewed as a local false nondiscovery rate studied in Efron et al. (2001) and Efron (2004, 2008, 2012) for testing multiple hypotheses. The same risk measure has been applied in Y. Chen et al. (2019) for the detection of leaked items and cheating examinees in a single test administration. The proposed method shares the same scalability as the local false discovery and nondiscovery rates for multiple testing. That is, no matter how large the item pool is, it is always sensible to use the proposed procedure without changing the threshold for compound risk, while, on the other hand, error metrics like the familywise error rate are far less scalable. In the sequential analysis literature, the idea of compound decision is rarely explored, except in Song and Fellouris (2019) and Bartroff (2018) where compound decision theory for sequential multiple testing is developed and in Y. Chen and Li (2021) where the compound decision framework for multistream change detection is first proposed.

The sequential monitoring of test quality has also been an important topic in the field of educational testing. For example, to monitor item quality, Veerkamp and Glas (2000) applied the cumulative sum control chart (CUSUM) method (Page, 1954) to sequentially detect changes in item difficulty. J. Zhang (2014) and J. Zhang and Li (2016) proposed a series of sequential statistical hypothesis tests for monitoring the item pool of a computerised adaptive testing system. Choe et al. (2018) proposed sequential change detection procedures for the detection of compromised items based on both item responses and response times. Lee and Lewis (2021) used CUSUM statistics to monitor item performance and detect item preknowledge in continuous testing. The existing methods focus on detecting changes in individual items, while, as we will discuss in Section 2.2, the proposed compound decision framework provides better integrative decisions for the entire item pool. To monitor general test quality, Lee and von Davier (2013) proposed sequential procedures to monitor score stability and assess scale drift of an educational assessment over time.

The rest of this article is organized as follows. In Section 2, we propose a general Bayesian change point model and a compound sequential detection procedure, followed by a specific model that can be easily implemented in operational tests. The theoretical properties of the proposed decision rule are established. Section 3 extends the development to a more realistic setting where some parameters of the Bayesian change point model are unknown. A compound decision rule is proposed under this setting and its statistical properties are proved. Section 4 discusses the problem of item quality control and the confounding due to the examinee population change over time. A statistic based on an IRT model is proposed, where this confounding factor is controlled. The performance of the proposed method is further evaluated in Section 5 via simulation studies. We conclude this article with remarks in Section 6.

2. Bayesian Change Point Detection

2.1. General Framework

We start with a general statistical framework for change detection in multiple data streams. For ease of exposition, we consider a frequent but noncontinuous test with fixed forms. Example tests of this type are discussed in Section 1. As will be discussed in Section 2.2, the proposed procedure can be generalized to other frequent tests, such as continuous tests with computerized adaptive testing (CAT) or MST designs. We use $t = 1, 2, ...$ to record test administrations; for example, $t = 1$ denotes the first test administration. Let S_t denote the item pool at time t that contains the items available for the tth test administration. The item pool is allowed to change over time, due to (1) the deletion of problematic items (e.g., items detected to have changed) and (2) the addition of new items.

For each item k, we monitor a certain statistic that reflects the psychometric properties of the item, denoted as X_{kt}, based on the data from all examinees in the tth test administration. This statistic may be univariate or multivariate, calculated based on data from the tth test administration. The monitoring statistic needs to be constructed carefully to adjust for confounding due to the change of the examinee population (e.g., seasonal effect), in order to reflect the real changes in individual items. For example, it is not a good idea to simply monitor the item percent correct. This is because the item percent correct may significantly increase in an test administration, due to that its examinee population overall has higher ability. As will be discussed in Section 4, one possible way to adjust for population change is by using an IRT model (Lord, 1980; Lord & Novick, 1968). In addition, the value of X_{kt} may be missing, as only a subset of items from the item pool S_t will be used in the test administration. We use $S_t^* \subset S_t$ to denote the set of items being used in the tth test administration. That is, X_{kt} is observed, if and only if $k \in S_t^*$. Figure 1 provides a flowchart illustrating this stochastic process. For the tth test administration, we start with an item pool S_t which is determined by information from the previous test administrations. Then, S_t^* is selected from S_t as the set of items for the tth test administration by test developers or certain test assembly algorithms. For these items, response data are collected, leading to monitoring statistics X_{kt}, $k \in S_t^*$. Historical information will be used to determine the item pool for the $(t + 1)$ th administration, by deleting problematic items and/or adding new items.

Each item k is associated with a change point, denoted by τ_k, which takes value in $\{1, 2, ...\} \cup \{\infty\}$. More precisely, τ_k is the time at which the change in the monitoring statistic X_{kt} occurs to item k. Here, we rule out the possibility that $\tau_k = 0$ as it is sensible to assume that items can only change after some exposure.[1] For example, the change of an item may be due to its leakage to the public at that time. The change time $\tau_k = \infty$ means that the item never changes. Throughout this article, we view τ_k as a random variable, whose prior

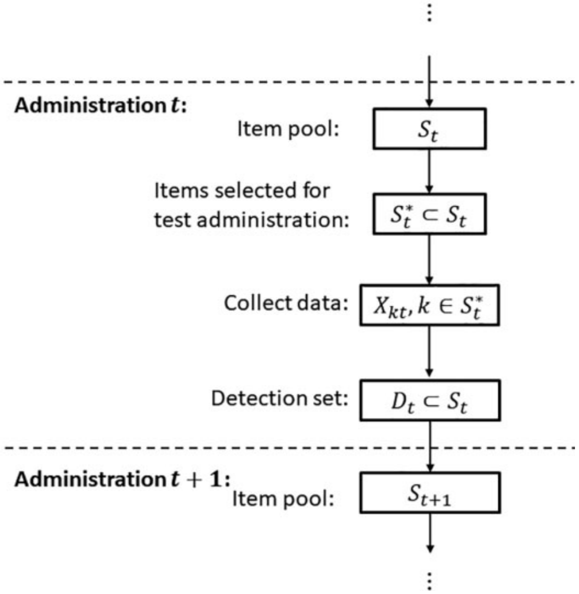

FIGURE 1. *A flowchart for the stochastic process of sequential test administration.*

distribution is allowed to vary across different streams. The distribution of X_{kt} only depends on the change point τ_k and is independent of other variables in the model. That is,

$$X_{kt}|\tau_k \sim \begin{cases} p_{kt} & \text{if}\, t \leq \tau_k \\ q_{kt} & \text{if}\, t > \tau_k, \end{cases} \tag{1}$$

where p_{kt} and q_{kt} are the density functions of the prechange and postchange distributions, respectively. For now, we assume p_{kt} and q_{kt} are both known, for example, two normal distributions whose means and variances are given. We discuss in Section 3 the situation when only partial information is available about these two distributions. Note that both the prechange and postdistributions may depend on t, for example, through the number of examinees in the corresponding test administration. Once data have been collected at time t, we would like to detect the postchange items, given all the information that is currently available. The sequential detection rule can be described by a detection set $D_t \subset S_t$, consisting of items that are likely to have changed. We point out that the proposed framework is very general, allowing the detected items in D_t to be removed or kept in the item pool S_{t+1} at the next time point.

Figure 2 provides a toy example to illustrate this change point model. In this example, at $t = 1$, the item pool only contains Items 1 and 2, and both are used in

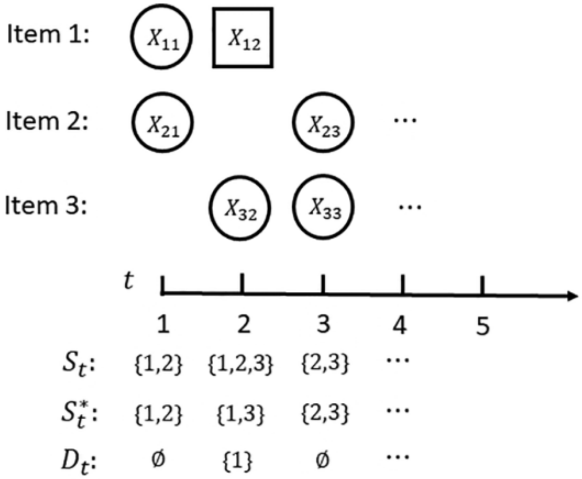

FIGURE 2. *A toy example illustrating the stochastic process of test administration. The prechange and postchange distributions of a monitoring statistic are indicated by circle and square, respectively.*

this test administration. As no change has occurred to these two items yet, the monitoring statistics for both items follow their prechange distributions as indicated by the circles. After the first test administration, a change point occurs to Item 1, recorded by $\tau_1 = 1$. At $t = 2$, Item 3 is added to the item pool, leading to $S_2 = \{1, 2, 3\}$. In this test administration, Items 1 and 3 are used and thus $S_2^* = \{1, 3\}$. As Item 1 has already changed, the monitoring statistic X_{12} now follows a postchange distribution, as indicated by the square. Based on information from the first two test administrations, Item 1 is detected and removed from the pool in this toy example, resulting in $S_3 = \{2, 3\}$. The process can further evolve at $t = 4, 5, ...$, following the flowchart in Figure 1.

We detect changed items by monitoring existing data at each time point t. More precisely, the existing information after the tth test administration is recorded by a sigma-field $\mathcal{F}_t$ defined recursively as $\mathcal{F}_t = \sigma(\mathcal{F}_{t-1}, S_t, S_t^*, X_{kt}, k \in S_t^*)$ with $\mathcal{F}_1 = \sigma(S_1, S_1^*, X_{k1}, k \in S_1^*)$. Note that the sigma-field $\mathcal{F}_t$ defines a information filtration satisfying $\mathcal{F}_s \subset \mathcal{F}_t$ for all $s \leq t$, meaning that as time passes, more and more information becomes available about the stochastic process of X_{kt} s. Note that information filtration is a key concept in stochastic processes. We refer the readers to Florescu (2014) for the mathematical details of an information filtration and its interpretation as accumulated information up to each time point.

We consider a Bayesian setting where the change points τ_k are viewed as random variables, for which the posterior probabilities $P(\tau_k < t|\mathcal{F}_t)$ can be evaluated at each time point t for any $k \in S_t$. The detection of postchange items will be based on these posterior probabilities. More precisely, we consider sequential detection rules D_t that are adaptive to the information filtration $\mathcal{F}_t$; that is, D_t is a random set that is measurable with respect to $\mathcal{F}_t$. It means that the sequential decision D_t is made only based on the information that is currently available as recorded in $\mathcal{F}_t$. We remark that, under this framework, S_{t+1} may be determined not only by the change detection results suggested by statistical algorithms like the one proposed herein but also by the domain knowledge of the test developers, which is common practice in the educational testing industry. That is, the detection results only provide the testing program warnings on potentially problematic items. These items will be reviewed by the test developers and then decisions will be made on whether to delete some existing items from the item pool and whether to add new items.

2.2. Proposed Compound Detection Procedure

We now propose a sequential change detection rule under the above general model. Following the discussions above, at each time t, it seems natural to flag the items whose posterior probability $P(\tau_k < t|\mathcal{F}_t)$ is large, as a larger posterior probability implies a higher chance of having changed. The question is, what cutoff value should we choose when making the decision? There is a trade-off behind this decision. On the one hand, we would like to detect as many post-change items as possible. On the other hand, we want to avoid making many false detections of prechange items, as false detections lead to unnecessary labor cost on item development, review, and maintenance. In what follows, an optimization program will be proposed to balance this trade-off.

Recall that the goal of our monitoring procedure is to maintain the quality of an item pool that may be measured by the proportion of postchange items in the pool at each time t. For example, the quality of S_t can be measured by $(\sum_{k\in S_t} 1_{\{\tau_k<t\}})/|S_t|$. The smaller the proportion, the better the quality of the item pool. As the change points τ_k for items $k \in S_t$ are unknown, this proportion is random. At a given time t, the best estimate of this quantity (under the mean squared error loss) is its conditional mean adaptive to the current information sigma-field $\mathcal{F}_t$.

The same quality measure can be used in the detection procedure. More precisely, let $D \subset S_t$ be the detection set after the tth test administration. The risk associated with the detection set D adaptive to the information filtration can be measured by

$$R(D|\mathcal{F}_t) = E\left(\frac{\sum\limits_{k\in S_t\setminus D} 1_{\{\tau_k<t\}}}{\max\{|S_t \setminus D|, 1\}}|\mathcal{F}_t\right), \tag{2}$$

where the denominator is chosen, so that $R(D|\mathcal{F}_t)$ is well-defined even when $D = S_t$. The smaller value of $R(D|\mathcal{F}_t)$ implies the better quality of the undetected items and thus a lower risk. Therefore, a reasonable criterion is to control the risk $R(D|\mathcal{F}_t)$ to be below a prespecified threshold (e.g., 1%). Under this criterion, the expected proportion of postchange items in the undetected set is below the threshold and thus the item pool is overall of high quality. Consequently, in preparation for future test administrations, only the detected items need investigation.

Following the terminology of compound decision theory (Efron, 2012), we refer to the ratio $(\sum_{k\in S_t\setminus D} 1_{\{\tau_k<t\}})/\max\{|S_t \setminus D|, 1\}$ as the false nondiscovery proportion (FNP) and the risk $R(D|\mathcal{F}_t)$ as the local false nondiscovery rate (FNR). Similarly, we define the false discovery proportion (FDP) as $\sum_{k\in D} 1_{\{\tau_k\geq t\}}/\max\{1, |D|\}$, that is the proportion of prechange items in the detection set. The local false discovery rate (FDR) is defined as the conditional expectation of the FDP.

The proposed decision rule at time t is to minimize the size of the detection set while controlling local FNR to be below a given threshold α, that is,

$$D_t = \arg\min_{D\subset S_t} |D|, s.t.R(D|\mathcal{F}_t) \leq \alpha. \tag{3}$$

Minimizing the detection set avoids making too many false detections of prechange items, and the constraint on local FNR ensures the detection of a sufficient number of postchange items. The proposed decision rule D_t can be obtained by Algorithm 1 shown below. With given posterior probabilities, the computation of Algorithm 1 is dominated by the sorting step whose complexity is $O(|S_t|\log|S_t|)$. The computation of posterior probabilities $W_{kt} = P(\tau_k < t|\mathcal{F}_t)$ under a specific change point model will be discussed in Section 2.3. The detection rule (3) is adaptive, in the sense that it makes use of up-to-date information $\mathcal{F}_t$. It is also compound, as the threshold on the posterior probabilities W_{kt} is determined by the posterior probabilities of all the items in the current item pool S_t. The proposed procedure is optimal in the sense described in Proposition 1.

Proposition 1: The sequential decision rule D_t given by Algorithm 1 satisfies that $R(D_t|\mathcal{F}_t) \leq \alpha$. In addition, for any other sequential decision rule $D'_t \subset S_t$ that is $\mathcal{F}_t$ measurable and satisfies $R(D_{t'}|\mathcal{F}_t) \leq \alpha$, we have $|D_t| \leq |D'_t|$ and

$$E\left(\sum_{k\in D_t} 1_{\{\tau_k\geq t\}}\right) \leq E\left(\sum_{k\in D'_t} 1_{\{\tau_k\geq t\}}\right).$$

Proportion 1 implies that the proposed sequential decision rule minimizes the expected number of false detections of prechange items at the current step, among all sequential decision rules that control the local FNR below the same level. Under some arguably more restrictive assumptions on the change point model, the proposed decision rule is not only optimal at the current step but also

ALGORITHM 1.
Proposed detection rule

```
Input: Threshold $\alpha$, the current item pool $S_t$, and posterior probabilities $(W_{kt})_{k \in S_t}$, where
    $W_{kt} = P(\tau_k < t|\mathcal{F}_t)$.
  1: Sort the posterior probabilities in an ascending order. That is,
    $W_{k_1,t} \leq W_{k_2,t} \leq \cdots \leq W_{k_{|S_t|},t}$,
  where $S_t = \{k_1, ..., k_{|S_t|}\}$. To avoid additional randomness, when there exists a tie
  ($W_{k_i,t} = W_{k_{i+1},t}$), we require $k_i < k_{i+1}$.
  2: For $n = 1, ..., |S_t|$, define
    $V_n = \frac{\sum_{i=1}^{n} W_{k_i,t}}{n}$.
  and define $V_0 = 0$.
  3.: Find the largest $n \in \{0, 1, ..., |S_t|\}$ such that $V_n \leq \alpha$.
Output: $D_t = S_t \setminus \{k_1, ..., k_n\}$ if $n \geq 1$ and $D_t = S_t$ if $n = 0$.
```

uniformly optimal throughout the entire sequential decision process. This optimality result is given in Supplemental Appendix B, available in the online version of this article.

Remark 1 (comparison with existing procedures): We compare the proposed change detection framework with existing works on sequential item pool monitoring. Two different approaches are taken in the existing works. Specifically, Veerkamp and Glas (2000) and Lee and Lewis (2021) apply the CUSUM procedure to sequential data for each individual item and declare a change once the CUSUM statistic exceeds a prespecified threshold. J. Zhang (2014), J. Zhang and Li (2016), and Choe et al. (2018) test the prechange hypothesis for each individual item at each time point. A change is declared if the *p* value from the hypothesis test is smaller than a prespecified threshold. The posterior probabilities in the current method play a similar role as the CUSUM statistics and *p* values in these works to measure the likelihood of each item having changed.

More specifically, we provide the connection between the posterior probabilities monitored by the proposed method and the *p* values monitored by the sequential hypothesis testing method. For a given item *k* and at each time point *t*, the sequential hypothesis testing approach tests the null hypothesis $H_0 : \tau_k \geq t$ versus the alternative hypothesis $H_1 : \tau_k < t$. Following the routine of frequentist hypothesis testing, a *p* value is obtained based on some carefully designed test statistic. If the *p* value is below a prespecified threshold, H_0 is rejected and the item is declared to have changed. The proposed method tests the same hypotheses but takes a Bayesian approach. The prior distribution for τ_k implies the prior probabilities $P(\tau_k \geq t)$ and $P(\tau_k < t)$ for the null and alternative hypotheses, respectively. Given these prior probabilities, together with the prechange and postchange distributions of data, the Bayes formula provides us the posterior probabilities of the null and alternative hypotheses. Bayesian hypothesis testing makes decision based on these posterior probabilities. We refer the readers to Wagenmakers et al. (2010) for a discussion on Bayesian hypothesis testing and a comparison with the frequentist approach. In this sense, the proposed method can be viewed as a Bayesian version of the sequential hypothesis testing method.

Although both the CUSUM and the sequential hypothesis testing methods can effectively detect postchange items, they do not provide an estimate of the number/proportion of postchange items among the undetected ones. Consequently, they cannot directly assess and control the quality of the item pool. These methods may still be able to control a similar risk as in the proposed method by tuning the corresponding threshold for declaring changes, but choosing a suitable threshold for this purpose is a challenging task that may require external information about the number/proportion of postchange items in the pool.

In contrast, the proposed method can directly control the quality of item pool by controlling the local FNR. As a price, it requires to know the Bayesian model for change points, including the prior distribution for the change points and the prechange and postchange distributions for the data streams. As will be explained in Section 3, the proposed method can be extended to control the local FNR given only some partial information about the Bayesian change point model.

Remark 2 (application to continuous testing): Item leakage may be more likely to occur in continuous testing with CAT, MST, and fixed-form testing designs. Continuous testing implies more frequent item usage from an item pool, for which monitoring item pool quality may be even more important. In particular, most of the existing works on the sequential detection of item changes are developed under a continuous CAT setting (Choe et al., 2018; Veerkamp & Glas, 2000; J. Zhang, 2014; J. Zhang & Li, 2016) or under a continuous MST or linear testing setting (Lee & Lewis, 2021).

We point out that the proposed framework is very general that can also be applied to monitoring the item pool of any continuous test. A setting for frequent but noncontinuous fixed-form tests is adopted in Section 2.1 for the ease of exposition, as our main focus is to introduce the new compound sequential decision framework. To apply the proposed framework to continuous testing, the meaning of each time point t needs to be slightly different from that in the existing works concerning a CAT setting. That is, most of the existing works concerning a CAT setting focus on individual items and examinees, where each time point for an item corresponds to its admission to an examinee. For example, for a given item, $t = 20$ means the item having been administered to 20 examinees. Consequently, the same t does not necessarily correspond to the same calendar time for different items, as different items may have different exposure rates. This choice of time t is thus not suitable for defining our compound risk that measures the item pool quality at some point of calendar time. To apply the proposed method to continuous testing, we can let each time point correspond to a fixed period of calendar time, for example, one day or half a day. The duration of the time period may be chosen based on the test volume to allow adequate sample sizes in computing the monitoring statistics. The monitoring statistics at a time point are constructed based on all the item responses collected during the corresponding period. The posterior distributions of change points can be updated based on the monitoring statistics. The compound risk at each time point can thus be defined and compound sequential decisions can be made accordingly. See Online Appendix E for a further discussion on applying the proposed method to continuous testing.

2.3. A Specific Change Point Model

We now provide a specific model for illustration. We assume that the change point τ_k satisfies

$$|\{s : k \in S_s^*, s \leq \tau_k\}| = \gamma_k,$$

where $\gamma_1, \gamma_2, \ldots$ are independent, each of which follows a geometric distribution. That is,

$$P(\gamma_k = m) = (1 - \rho_k)^{m-1}\rho_k, m = 1, 2, ..., \qquad (4)$$

where ρ_k is an item-specific parameter in the open interval $(0, 1)$. This model implies that, on average, the status of an item changes (e.g., being leaked) after being used in $1/\rho_k$ test administrations (i.e., exposures). We point out that the geometric distribution is widely used in the Bayesian formulation of sequential change detection because of its memoryless property. The proposed methods can be extended to other prior distributions for the change points.

We may further assume that both the prechange and postchange distributions are univariate normal. Specifically, as will be justified by the possible choices of the monitoring statistic as in Section 4, we assume that the prechange distribution p_{kt} is standard normal. The postchange distribution q_{kt} is $N(\mu_{kt}, 1)$. For now, we consider the case where ρ_k and μ_{kt} are both known and leave the unknown case to Section 3.

Under this model, the posterior distribution $P(\tau_k < t|\mathcal{F}_t)$ can be computed in an analytic form. To simplify the notation, we denote $W_{kt} = P(\tau_k < t|\mathcal{F}_t)$. This posterior probability can be obtained by a simple updating rule, summarized in the following proposition.

Proposition 2: Assume γ_k follows a geometric prior described in Equation 4 and let $e_{kt} = \sum_{i=1}^{t} 1_{\{k \in S_i^*\}}$ be the number of exposure of item k up to time t. Then, $W_{kt} = \frac{U_{kt}}{U_{kt}+1/\rho_k}$, where U_{kt} is computed through the following updating rule,

$$if\ e_{kt} \leq 1, U_{kt} = 0, else U_{kt} = \begin{cases} U_{k,t-1}\ if\ k \in S_t \setminus S_t^*, \\ (1 + U_{k,t-1})\dfrac{q_{kt}(X_{kt})}{(1 - \rho_k)p_{kt}(X_{kt})}\ if\ k \in S_t^*. \end{cases} \qquad (5)$$

Recall that p_{kt} and q_{kt} are the density functions of the prechange and postchange distributions, respectively, defined in Equation 1. In particular, if $p_{kt} \sim N(0, 1)$ and $q_{kt} \sim N(\mu_{kt}, 1)$, then

$$\mathit{if}\ e_{kt} \leq 1, U_{kt} = 0, \mathit{else} U_{kt} = \begin{cases} U_{k,t-1}\ \mathit{if}\ k \in S_t \setminus S_t^* \\ \frac{1}{1-\rho_k}(1 + U_{k,t-1})\exp\{\mu_{kt}X_{kt} - \frac{\mu_{kt}^2}{2}\}\ \mathit{if}\ k \in S_t^*. \end{cases} \quad (6)$$

In the above proposition, the statistic U_{kt} is a modification of a classic sequential change detection statistic (Shiryaev, 1963) which gives optimal sequential change detection for a single data stream under a Bayesian decision framework. We point out that U_{kt} is not updated until item k is exposed at least twice (i.e., $e_{kt} > 1$), because we do not allow $\tau_k = 0$. According to Equation 5, the update of the posterior probabilities W_{kt} is straightforward when the prechange and postchange distributions are known. These distributions are not necessarily normal, though it may be convenient to make the normality assumption in the current application as discussed in Section 4.

3. When Model Is Not Completely Known

Now, we consider the situation in which only partial information is available about the change point model. More precisely, we focus on the specific change point model given in Section 2.3. Recall that the change point τ_k follows a geometric distribution (Equation 4) with parameter ρ_k. It is further assumed that the prechange distribution p_{kt} is known, for example, a standard normal distribution. In addition, it is assumed that the postchange distribution q_{kt} can be parameterized as

$$q_{kt}(x) = h_{kt}(x|\pi_k),$$

where h_{kt} is a known function and π_k is an item-specific parameter vector. This parametrization of q_{kt} will be justified under an IRT model in Section 4.

In practice, ρ_k and π_k are unknown, but prior information may be available. Specifically, $1/\rho_k$ represents the average number of exposures (i.e., number of times the item is used) for the item to change. Although it is hard to know $1/\rho_k$ precisely, a reasonable lower bound is often available, which leads to an upper bound for ρ_k, denoted by $\overline{\rho} \geq \rho_k$, for all k. In addition, as will be further justified in Section 4, we assume that $\pi_k \in \Theta$, where Θ is a known compact set. In what follows, we propose a method that controls the compound risk $R(D_t|\mathcal{F}_t)$, when only knowing $\bar{\rho}$, p_{kt}, and Θ.

Let $W_{kt}(\rho, \pi)$ denote the posterior probability $P(\tau_k < t|\mathcal{F}_t)$ when the underlying parameters are $\rho_k = \rho$ and $\pi_k = \pi$, and define $\overline{W}_{kt}$ as

$$\overline{W}_{kt} = \sup_{\rho\in(0,\overline{\rho}],\pi\in\Theta} W_{kt}(\rho, \pi). \quad (7)$$

Note that $\overline{W}_{kt}$ is not a posterior probability, but an upper bound of the posterior probability $W_{kt}(\rho_k, \pi_k)$. We now provide Algorithm 2 that replaces W_{kt} in

ALGORITHM 2
Proposed detection rule

```
Input: Threshold $\alpha$, the current item pool $S_t$, and $\overline{W}_{kt}$ defined in (7).
  1: Sort the $\overline{W}_{kt}$ in an ascending order. That is,
       $\overline{W}_{k_1,t} \leq \overline{W}_{k_2,t} \leq \cdots \leq \overline{W}_{k_{|S_t|},t}$,
     where $S_t = \{k_1, ..., k_{|S_t|}\}$. To avoid additional randomness, when there exists a tie
     ($\overline{W}_{k_i,t} = \overline{W}_{k_{i+1},t}$), we require $k_i < k_{i+1}$.
  2: For $n = 1, ..., |S_t|$, define
       $V_n = \frac{\sum_{i=1}^{n} \overline{W}_{k_i,t}}{n}$.
     and define $V_0 = 0$.
  3: Find the largest $n \in \{0, 1, ..., |S_t|\}$ such that $V_n \leq \alpha$.
Output: $D_t = S_t \setminus \{k_1, ..., k_n\}$ if $n \geq 1$ and $D_t = S_t$ if $n = 0$.
```

Algorithm 1 by $\overline{W}_{kt}$. As shown in Theorem 1, the decision rule given by Algorithm 2 controls the compound risk $R(D_t|\mathcal{F}_t)$ at any time t.

> **Theorem 1:** Suppose that $\rho_k \leq \overline{\rho}$ and $\pi_k \in \Theta$ for all k. Then, the proposed decision rule given in Algorithm 2 guarantees that $R(D_t|\mathcal{F}_t) \leq \alpha$ for all $t = 1, 2, \ldots$

It remains to find a way to compute $\overline{W}_{kt}$, as it is defined by an optimization over an iteratively defined object. Proposition 3 provides guidance to this problem.

Proposition 3: Let $e_{kt} = \sum_{i=1}^{t} 1_{\{k \in S_i^*\}}$ be the number of exposures of item k up to time t. Define $U_{kt}(\overline{\rho}, \pi)$ according to the following iterations,

$$\textit{if } e_{kt} \leq 1, U_{kt}(\overline{\rho}, \pi) = 0, \textit{else} U_{kt}(\overline{\rho}, \pi) = \begin{cases} U_{k,t-1}(\overline{\rho}, \pi) \textit{ if } k \in S_t \setminus S_t^*, \\ \dfrac{1}{1 - \overline{\rho}}(1 + U_{k,t-1}(\overline{\rho}, \pi))\dfrac{h_{kt}(X_{kt}|\pi)}{p_{kt}(X_{kt})} \textit{ if } k \in S_t^*. \end{cases} \tag{8}$$

Let $\overline{R}_{kt} = \sup_{\pi \in \Theta} U_{kt}(\overline{\rho}, \pi)$. Then, $\overline{W}_{kt} = \frac{\overline{R}_{kt}}{\overline{R}_{kt} + 1/\overline{\rho}}$.

It can be shown that $W_{kt}(\rho, \pi)$ is monotone increasing with respect to ρ. Therefore, to obtain $\overline{W}_{kt}$, Proposition 3 plugs $\overline{\rho}$ into $W_{kt}(\rho, \pi)$. When the dimension of π is very low (e.g., one or two), we can discretize the set Θ by grid points, update $U_{kt}(\overline{\rho}, \pi)$ on the grid points in parallel, and then approximate $\overline{W}_{kt}$ accordingly. When the number of parameters in π is not very low, by making use of Equation 8, the gradient of $U_{kt}(\overline{\rho}, \pi)$ with respect to π can be computed iteratively. Thus, $\overline{R}_{kt}$ can be computed, for example, by a gradient ascent algorithm.

4. Monitoring Statistic in IRT-Based Testing

In principle, the monitoring statistic X_{kt} can be any statistic whose distribution is different before and after the change point. In practice, we suggest to choose X_{kt} to be a Wald-type statistic, so that we can approximate the prechange and postchange distributions by normal distributions to simplify the computation. In what follows, we give an example of such a monitoring statistic and explain how confounding due to changing examinee population is adjusted in this statistic.

4.1. Standardized Item Residual (SIR) Statistic

Let N_t be the number of people taking the test at time t. Let $Y_{ktn} \in \{0, 1\}$ denote the nth examinee's response to item k at time t, where $Y_{ktn} = 1$ indicates a correct response and $Y_{ktn} = 0$ otherwise.

Let $\bar{Y}_{kt} = (\sum_{n=1}^{N_t} Y_{ktn})/N_t$ be the percent correct for item k at time t. Suppose that a change has not yet occurred. Then, $\bar{Y}_{kt}$ has expected value $\xi_{kt}^0 := E(\bar{Y}_{kt}|\tau_k \geq t)$. If ξ_{kt}^0 is known and let $SE(\bar{Y}_{kt})$ be the standard error of $\bar{Y}_{kt}$, then the standardized residual $(\bar{Y}_{kt} - \xi_{kt}^0)/SE(\bar{Y}_{kt})$ is approximately standard normal when the change has not yet occurred and N_t is sufficiently large. Note that this statistic can adjust for the examinee population change when ξ_{kt}^0 is defined under the IRT framework.

In practice, we typically do not know ξ_{kt}^0. Specially, when the examinee population changes overtime, ξ_{kt}^0 needs to be estimated based on both historical information and data from the tth test administration. Now suppose that we have a consistent estimator of ξ_{kt}^0, denoted as $\hat{\xi}_{kt}^0$. The way obtaining $\hat{\xi}_{kt}^0$ will be discussed in Section 4.2. Then, the SIR statistic is defined as

$$X_{kt} = \frac{\bar{Y}_{kt} - \hat{\xi}_{kt}^0}{SE(\bar{Y}_{kt} - \hat{\xi}_{kt}^0)},$$

where $SE(\bar{Y}_{kt} - \hat{\xi}_{kt}^0)$ is the standard error of the numerator. Under mild conditions and given that $\tau_k \geq t$, X_{kt} is approximately standard normal for sufficiently large N_t. Similarly, given that $\tau_k < t$, X_{kt} is asymptotically $N(\mu_{kt}, 1)$, where μ_{kt} characterizes the mean change of the SIR statistic. The normal approximation can be justified in sense that X_{kt} can be decomposed as $X_{kt} = \mu_{kt} + Z + o_p(1)$, where Z is a standard normal random variable. This holds even when μ_{kt} diverges. When focusing on change points due to item leakage, it is reasonable to further assume that $\mu_{kt} > 0$.

In general, the SIR statistics X_{kt}, $k \in S_t$, are not independent as assumed in our Bayesian change point model. Specifically, the SIR statistics constructed under an IRT model tend to have weak positive correlations, brought by individual-specific latent factors. As shown in Section 5, with such dependence, the proposed method still controls the compound risk.

4.2. SIR Statistic Under IRT Framework

In what follows, we describe an IRT model for item response data Y_{ktn}. Under this model, the SIR statistic X_{kt} can be computed and the mean change μ_{kt} can be expressed as a function of the parameters in the IRT model.

4.2.1. IRT model for prechange items. IRT provides a popular method in educational testing for linking different test administrations with potentially different examinee populations. In the current context, it is sensible to model item responses using a unidimensional IRT model for prechange items. A unidimensional IRT model assumes that each item is characterized by one or multiple parameters that do not change over time, denoted by β_k, and that each examinee n in test administration t is characterized by one parameter, denoted by θ_{tn}. The parameter θ_{tn} is typically interpreted as the ability of the examinee.

Under this model, the probability of an examinee answering item k correctly is completely determined by the item parameters and the person parameter in the form

$$P(Y_{ktn} = 1|\theta_{tn}, \beta_k) = f(\theta_{tn}|\beta_k),$$

where f is a prespecified inverse-link function that is monotonically increasing in θ_{tn}. For example, one of the most commonly used models in educational testing is the so-called two-parameter logistic (2PL) model (Birnbaum, 1968). Under the 2PL model,

$$f(\theta_{tn}|\beta_k) = \frac{\exp(\beta_{k0} + \beta_{k1}\theta_{tn})}{1 + \exp(\beta_{k0} + \beta_{k1}\theta_{tn})},$$

where β_{k0} and $\beta_{k1} > 0$ are known as the easiness and discrimination parameters, respectively, and $\beta_k = (\beta_{k0}, \beta_{k1})$. Given the person and item parameters, an examinee's responses to different items are assumed to be independent, known as the local independence assumption.

In this article, the item parameters β_k are treated as fixed parameters that do not change over time. The person parameters θ_{tn} are treated as random variables. Specifically, $\theta_{t1}, ..., \theta_{t,N_t}$ are assumed to be independent and identically distributed (i.i.d.) samples from a distribution $N(m_t, 1)$, where the mean is time dependent to reflect the population change over time (e.g., seasonal effect) and the variance is fixed to be one to bypass the scale-indeterminacy. Under the IRT model, the expected percent correct can be calculated as

$$\xi_{kt}^0 = \int f(\theta|\beta_k)\frac{1}{\sqrt{2\pi}}\exp(-(\theta - m_t)^2/2)d\theta. \tag{9}$$

4.2.2. IRT model for postchange items. We now describe an IRT model for response data involving item preknowledge. The same model has been adopted in Lee and Lewis (2021) who developed CUSUM statistics based on this model to monitor item performance and detect item preknowledge. For each k satisfying $t > \tau_k$, we use $\eta_{ktn} \in \{0, 1\}$ to denote preknowlege about the item. That is, $\eta_{ktn} = 1$ indicates that the examinee has preknowledge about the item and $\eta_{ktn} = 0$ otherwise. We assume that the indicators η_{ktn} are i.i.d., following a Bernoulli distribution with $P(\eta_{ktn} = 1) = \pi_k$, and η_{ktn} is independent of θ_{tn}. That is, whether an individual has preknowlege about a leaked item is independent of their ability. We further assume that when examinee n has preknowledge about item k, their answer is always correct. That is, $Y_{ktn} = 1$, given that $\eta_{ktn} = 1$. Finally, it is assumed that when examinee n does not cheat on item k, that is, $\eta_{ktn} = 0$, then Y_{ktn} given η_{ktn} and θ_{tn} still follows the same IRT model as if the item has not changed. This postchange model yields $\xi^1_{kt} = (1 - \pi_k)\xi^0_{kt} + \pi_k$ and therefore $\xi^1_{kt} - \xi^0_{kt} = \pi_k(1 - \xi^0_{kt})$.

4.2.3. Estimation of ξ^0_{kt}. In practice, the expected prechange percent correct ξ^0_{kt} is unknown due to the unknown mean m_t which needs to be estimated based on data from the prechange items in the tth test administration. Suppose that there exists a nonempty subset $S^\dagger_t \subset S^*_t$ that is known to only contain prechange items. For example, $S^\dagger_t$ can be the items that have not been exposed before. Then, m_t can be consistently estimated by maximizing the marginal likelihood

$$\hat{m}_t = \underset{m_t}{\operatorname{argmax}} \sum_{n=1}^{N_t} \log\left(\int \prod_{k \in S^\dagger_t} f(\theta|\beta_k)^{Y_{ktn}} (1 - f(\theta|\beta_k))^{1-Y_{ktn}} \frac{1}{\sqrt{2\pi}} \exp(-(\theta - m_t)^2/2) d\theta \right). \tag{10}$$

Accordingly, ξ^0_{kt} can be estimated by plugging in $\hat{m}_t$. The standard error $SE(\bar{Y}_{kt} - \hat{\xi}^0_{kt})$ can also be computed easily; see the details in Online Appendix B.

4.2.4. Prechange and postchange distributions of SIR statistic. The SIR statistic X_{kt} constructed above is approximately standard normal when $\tau_k \geq t$ and is approximately normal $N(\mu_{kt}, 1)$ when $\tau_k < t$, where

$$\mu_{kt} = \frac{\xi^1_{kt} - \xi^0_{kt}}{\sqrt{\operatorname{Var}(\bar{Y}_{kt} - \hat{\xi}^0_{kt})}}.$$

The value of ξ^1_{kt} is determined by the postchange model introduced above and the value of ξ^0_{kt} is determined by the prechange model that can be estimated from data. Given the leakage proportion π_k, μ_{kt} can be approximated by

$$\hat{\mu}_{kt}(\pi_k) = \frac{\pi_k(1 - \hat{\xi}_{kt}^{0})}{SE(\bar{Y}_{kt} - \hat{\xi}_{kt}^{0})}.$$

In practice, π_k is usually unknown and hard to estimate, though it may be known by prior knowledge that π_k s locate in a certain interval Θ. With such information, we can run Algorithm 2 with

$$h_{kt}(x|\pi_k) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \hat{\mu}_{kt}(\pi_k))^2}{2} \right). \tag{11}$$

5. Simulation Study

5.1. Study I

5.1.1. Known change point model. We start with a simple simulation setting to illustrate the proposed method. We consider an item pool originally containing $|S_1| = 500$ items. During the process, once a subset of items are detected, they will be removed, and the same number of new items will be added to ensure $|S_t| = 500$ for all t. We also assume that 50 items are randomly selected from the item pool for each test administration, that is, $|S_t^*| = 50$.

The parameter ρ_k in the change time distribution is generated from a uniform distribution over the interval $[0, 0.1]$ for different k. It is further assumed that the monitoring statistic X_{kt} follows $N(0, 1)$ when $\tau_k \geq t$ and follows $N(\mu_k, 1)$ when $\tau_k < t$. We generate μ_k from a uniform distribution over the interval $[1, 2]$ for different k.

We investigate the situation when ρ_k and μ_k are known. We run 1,000 independent simulations. In each simulation, we apply Algorithm 1, for $t = 1, ..., 50$, where the threshold α for the compound risk is set to be 0.01. To evaluate the method, three metrics are calculated at each time t, including (a) the FNP

$$\frac{\sum_{k \in S_t \setminus D_t} 1_{\{\tau_k < t\}}}{\max\{1, |S_t \setminus D_t|\}},$$

(b) the FDP

$$\frac{\sum_{k \in D_t} 1_{\{\tau_k \geq t\}}}{\max\{1, |D_t|\}},$$

and (c) the number of detections $|D_t|$. Our results are shown in Figure 3, where panels (a)–(c) show the three metrics, respectively. In each panel, the x-axis shows time t and the y-axis shows the 5%, 25%, 50%, 75%, and 95% quantiles of the empirical distribution of the metric based on 1,000 independent simulations.

We take a closer look at the medians of these metrics over time. The median FNP is zero at the beginning, because at time $t = 1$, all the items are new (i.e., never exposed before). It increases as time goes on and stabilizes at the targeted level 0.01 after about 10 time points. The median FDP is also zero at the beginning but then increases dramatically. It stabilizes around the level 0.8 after about 20 time points. Finally, the median detection size also increases with time t and becomes stable around 10 after about 20 time points.

5.1.2. Unknown change point model. We now look at the situation when ρ_k and μ_k are unknown under the same simulation setting as above. We again run 1,000 independent simulations. In each simulation, we apply Algorithm 2, for $t = 1, ..., 50$, where the threshold $\alpha = 0.01$. When applying Algorithm 2, we only know that the prechange distribution is $N(0, 1)$, $\rho_k \in (0, 0.1]$ and $\mu_k \in [1, 2]$. The results are shown in Figure 4 which take a similar form as those given in Figure 3. As we can see, when the change point model is unknown, the decision given by Algorithm 2 still controls the FNP under the targeted level. However, when comparing the current results with those from Algorithm 1 above, we see that the decision rule given by Algorithm 2 is more conservative, which is a price paid for not knowing the parameters in both the geometric distribution for change points and the postchange distribution. As a result of the conservativeness in the decision (i.e., smaller FNP), the FDP and the number of detections are both larger comparing with the results in Figure 3.

5.1.3. Under model misspecification. We further look at the situation when the change point model is misspecified. Specifically, we consider the case when the monitoring statistics X_{kt}, $k \in S_t^*$, are correlated. More precisely, we assume $(X_{kt} : k \in S_t^*)$ given τ_k, $k \in S_t^*$ is multivariate normal distribution, for which the marginal prechange distribution of X_{kt} is still standard normal and the marginal postchange distribution is $N(\mu_k, 1)$, and the covariance between X_{kt} and $X_{k't}$ is 0.1 for $k \neq k'$. The rest of the setting is the same as above.

We apply Algorithm 1, assuming that ρ_k and μ_k are known and pretending that the data streams are independent. The results are shown in Figure 5. Comparing the results in Figures 3 and 5, it seems that the performance metrics are only slightly affected when we simply run Algorithm 1, ignoring the weak positive dependence between the data streams. We further apply Algorithm 2, with knowledge that all the ρ_k s lie in the interval $(0, 0.1]$ and that all the μ_k s lie in the interval $[1, 2]$. Again, when running Algorithm 2, we pretend that the data streams are independent. The results are shown in Figure 6. Similar to the results when given the known model, the effect of ignoring the weak positive dependence in data also seems small when the change point model is not completely known.

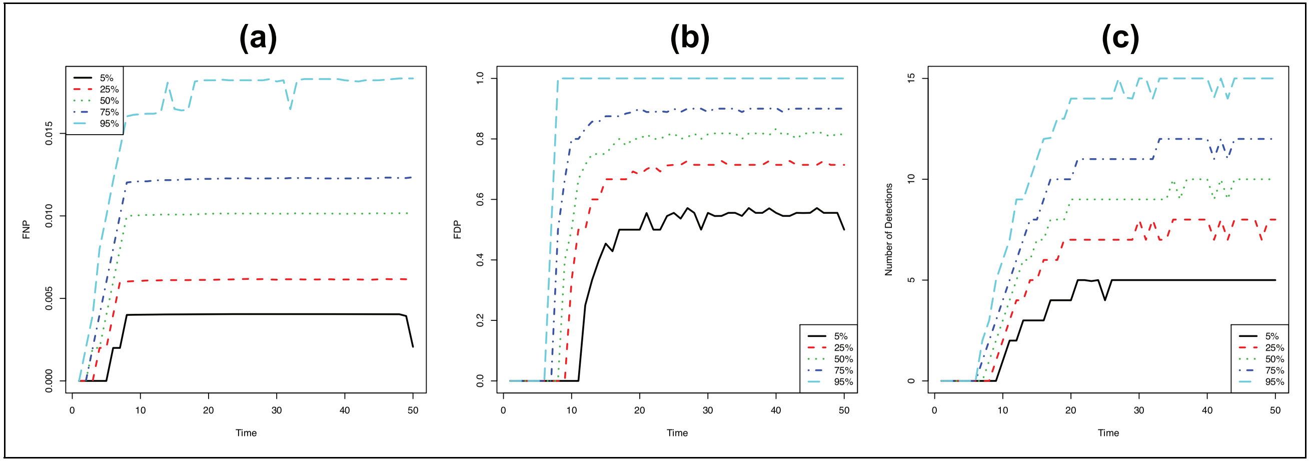

FIGURE 3. *The performance of Algorithm 1 under model correct specification. Panels (a)–(c) show the FNP, FDP, and number of detections, respectively. In each panel, the* x*-axis shows time* t *and the* y*-axis shows the 5%, 25%, 50%, 75%, and 95% quantiles of the empirical distribution of the metric based on 1,000 independent simulations.*

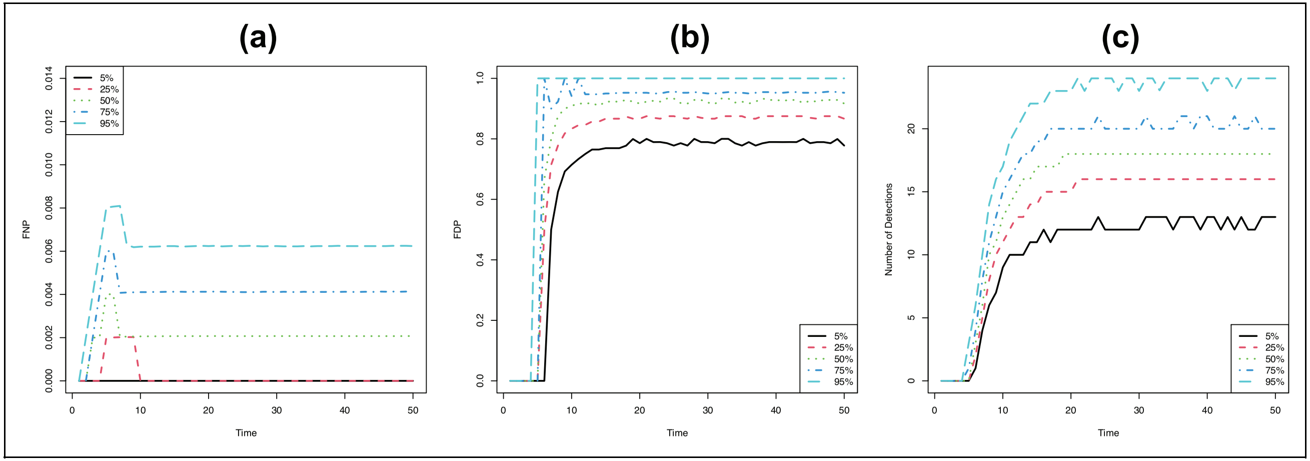

FIGURE 4. *The performance of Algorithm 2 under model correct specification. The plots can be interpreted the same as those in Figure 3.*

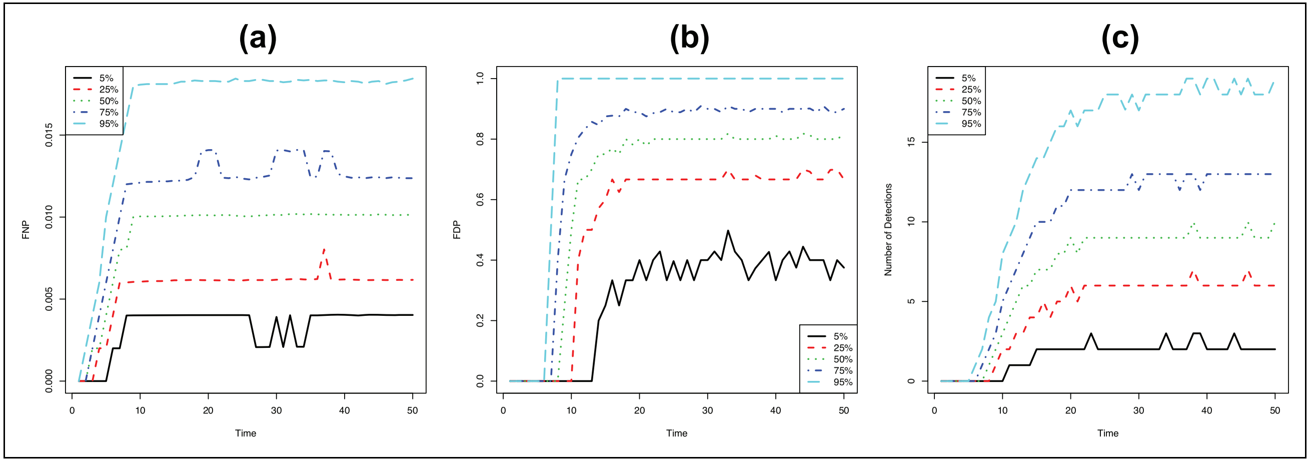

FIGURE 5. *The performance of Algorithm 1 under model mispecification. The plots can be interpreted the same as those in Figure 3.*

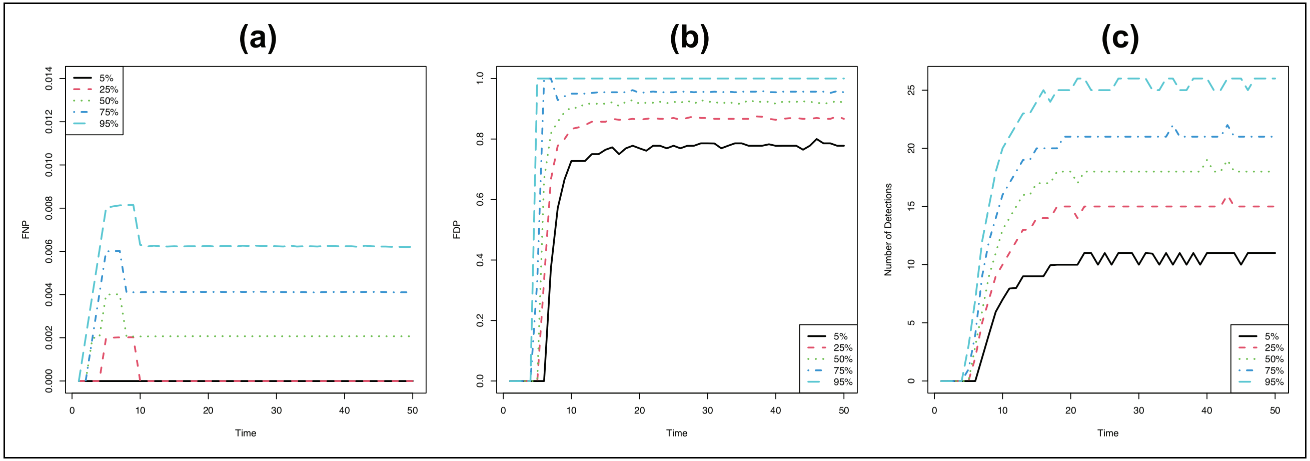

FIGURE 6. *The performance of Algorithm 2 under model misspecification. The plots can be interpreted the same as those in Figure 3.*

5.2. Study II: Educational Testing With Time-Varying Population

5.2.1. Simulation setting. We now evaluate the proposed method under an IRT setting that mimics operational tests. The setting is almost the same as above, except for the way the monitoring statistics are obtained. More precisely, prechange item response data are simulated using the 2PL model introduced in Section 4.2. Each item k is assumed to be associated with item parameters β_{k0} and β_{k1}, where the discrimination parameter β_{k1} is generated from a uniform distribution $[1, 1.5]$, and the easiness parameter β_{k0} is generated from a uniform distribution $[-2, 2]$. We assume that the item parameters are known in our sequential decision procedure, because in practice these parameters can usually be accurately precalibrated, before their operational use. At each time t, the number of examinees N_t is generated from a uniform distribution over the set $\{1001, 1002, ..., 3000\}$. Each examinee n at time t is associated with an ability parameter θ_{tn}, generated from normal distribution $N(m_t, 1)$, where the population mean m_t is generated from a uniform distribution over the interval $[-0.5, 0.5]$. The postchange item response data are simulated using the mixture model described in Section 4.2. The leakage proportion π_k is generated from a uniform distribution over the interval $[0.05, 0.1]$.

The simulated item pool originally contains $|S_1| = 500$ items. During the process, once a subset of items are detected, they will be removed, and the same number of new items will be added to maintain the size of the item pool. In addition, we add new items to S_t when necessary to ensure that S_t always contains at least five new items that have not been exposed before. This is because, we include at least five new items in S_t^* for the estimation of m_t, the mean of the population ability distribution at time t. The rest of the simulation setting is the same as that of Study I.

We construct an SIR statistic for each data stream using the method introduced in Section 4.2. Both the prechange and postchange distributions are approximated by normal distributions. We point out that the signal under the current setting is stronger than that under Study 1. In particular, the 25%, 50%, and 75% quantiles of the empirical distribution for $\hat{\mu}_{kt}(\pi_k)$ are 2.3, 3.9, and 5.1, respectively. Two cases are considered. In the first case (Case I), both parameters ρ_k and leakage proportions π_k are treated as known. We run Algorithm 1 with the prechange distribution being standard normal and the postchange distribution being estimated as in Equation 11. In the second case (Case II), both ρ_k and π_k are treated as unknown, which is usually the case in practice. We assume that we only know these parameters lying in the intervals $(0, 0.1]$ and $[0.05, 0.1]$, respectively. We run Algorithm 2, with the prechange distribution being standard normal and the postchange distribution being estimated as in Equation 11.

5.2.2. Results. The results for Case I are presented in Figure 7. As we can see, the proposed method still performs reasonably well under this setting. Specifically, the median FNP is slightly larger than the targeted level 0.01 after 20 time points, but it never exceeds 0.013. This slight overshoot is likely due to the normal approximation and the estimation of the population distribution at each time point. The FDP is quite small, with the median FDP always being zero. This is because, the signal of the change points is quite strong, given the sample sizes and the leakage proportions. Moreover, the number of detections remains low overtime. In fact, the median number of detection is always below 3.

The results for Case II are given in Figure 8. Similar to the results of Algorithm 2 as in Study I, the FNPs in Case II are smaller than those in Case I, meaning that Algorithm 2 again makes more conservative decisions. Specifically, the median FNP is always below the 0.004 level. The FDP values are acceptable but are much larger than those in Case I. Specifically, the median FDP is always below 0.73. Finally, the numbers of detections are still reasonably low, with the median number of detections always below 7. It would be affordable for the testing program to review detected items when detection size is of this scale.

6. Concluding Remarks

In this article, we provide a compound change detection framework for sequential item quality control, one of the most important problems in educational testing. A Bayesian change point model is proposed in both general and specific forms. Compound decision rules are proposed when the Bayesian change point model is completely known and when only partial information is available about the model. Theoretical properties of these decision rules are established and their empirical performance is evaluated by simulations under various settings. Our simulation studies show that the proposed method reasonably controls the proportion of postchange items among the undetected ones without making too many false detections of prechange items, suggesting that the proposed method may be applicable to operational tests for their quality control. Our simulation results also show that the proposed method is quite robust against model misspecification. More specifically, although our method is developed under a change point model assuming independent monitoring statistics, it still performs well when there exists weak positive dependence among the monitoring statistics.

We clarify that the proposed method controls local FNR but does not control local FDR. The local FDR in each step of our procedure is a result of the signal of the data and the threshold we use for local FNR. Given a threshold for our local FNR and an IRT setting, there is no simple way to characterize the relationship between the sample size and the local FDR, as the signal in the data depends on many different factors, not only the sample size but also the number of test takers to whom each postchange item is leaked, population of test takers, the number of items in a test, and the characteristics of the items in the pool (e.g., the

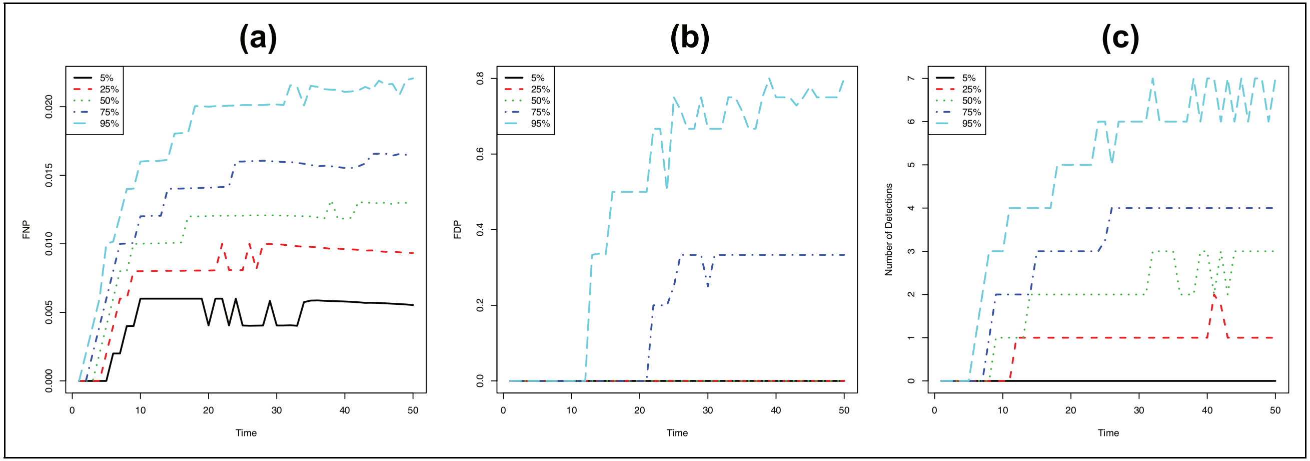

FIGURE 7. *The performance of Algorithm 1 when data are generated from an item response theory model under a setting that mimics operational tests, and the monitoring statistics are constructed based on item response data. The plots can be interpreted the same as those in Figure 3.*

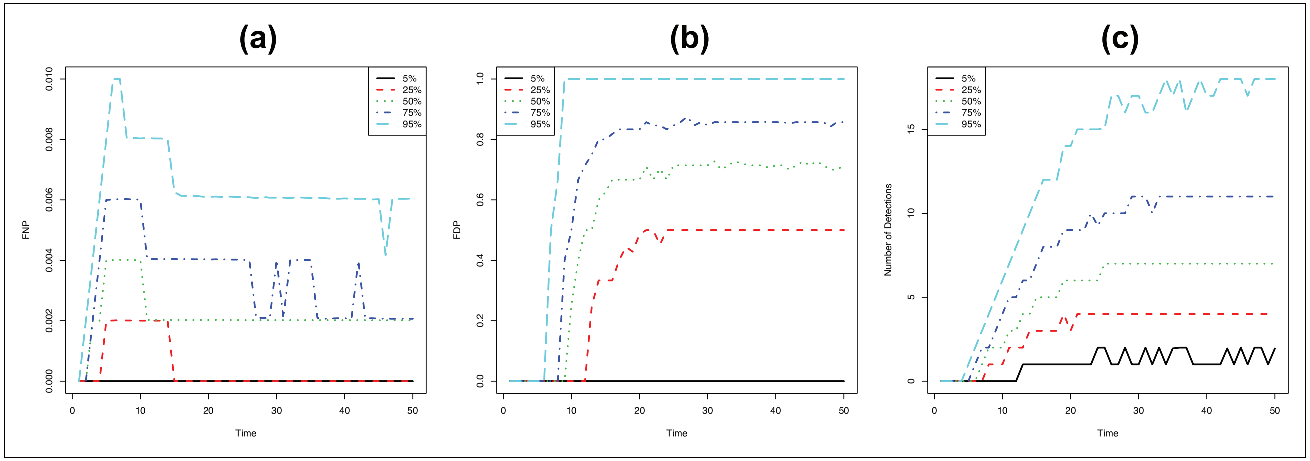

FIGURE 8. *The performance of Algorithm 2 when data are generated from an item response theory model under a setting that mimics operational tests, and the monitoring statistics are constructed based on item response data. The plots can be interpreted the same as those in Figure 3.*

distributions of the discrimination and difficulty parameters). For a real test, we would suggest to run simulation studies to understand how the local FDR depends on these factors, given the parameters of the items in the pool and the target level for local FNR.

One limitation of the current work lies in the simulation study. As discussed in Section 2.2, the proposed method can be generalized to continuous tests, for which item pool monitoring may be more important than frequent but noncontinuous tests. The current simulation only considers settings for frequent but noncontinuous tests. Its results do not imply the performance of the proposed method under settings for continuous tests, where the monitoring statistic for each stream is likely obtained from small and varying sample sizes. Simulation studies under real continuous test settings will be conducted in future research.

Another limitation is that the specific model with independent geometric change points may not be flexible enough. For example, the change points are likely correlated, driven by some events such as the leakage of a set of items to the public or the change of curriculum that may affect multiple items. A challenge from removing the independent geometric distribution assumption is that the posterior probabilities $P(\tau_k < t|\mathcal{F}_t)$ typically do not have an analytic form. Several questions are worth future investigation. First, can we still control the compound risk using a misspecified independent geometric model? Second, can we develop methods for approximating these posterior probabilities, such as variational approximation and certain Bayesian filters (e.g., Chapter 10, Bishop, 2006)?

In practice, there is always unknown information in the Bayesian change point model, especially the distribution of change points and the postchange distribution. The current solution is to take a conservative approach that makes decision under essentially the worst-case model. This approach guarantees the control of the compound risk for a finite sample, but there might be a sacrifice in making more false detections of prechange items. An alternative is to take an online estimation approach that estimates the unknown model parameters sequentially together with the sequential change detection process. The unknown parameters can be treated as random variables and estimated by a Bayesian approach or as fixed parameters and estimated by a likelihood-based approach. Similar to the multiarmed bandit problem (e.g., Robbins, 1952; Sutton & Barto, 2018), this approach also faces an exploration–exploitation trade-off dilemma, that is, the trade-off between the efforts to achieve a more accurate estimation and to make better decision. This problem is left for future investigation.

Many sequential change detection applications involve multiple data streams, such as the detection of customer behavior change in e-commence, the detection of changed sensors in engineering, and the detection of abnormal change in stocks. For such problems, it may be more sensible to control FDR- or FNR-type compound risks than individual risks for single data streams. Although this article focuses on item quality control in educational testing, the proposed methodological framework is very general and applicable to many other multi-stream change detection problems in different fields. In fact, the proposed procedure can be easily modified to control local FDR.

Acknowledgment

We thank the the editor, associate editor, and three referees for their useful suggestions. Any opinions expressed in this publication are those of the authors and not necessarily of Educational Testing Service.

ORCID iD

Yi-Hsuan Lee https://orcid.org/0000-0001-5637-9900

Note

1. Theoretically, the proposed Bayesian change detection framework can allow $\tau_k = 0$ to occur with a positive probability. We make the assumption that $\tau_k > 0$ because the major application of the proposed method considers the detection of item compromization that is due to previous exposure of the items (see, e.g., Veerkamp & Glas, 2000).

References

American Educational Research Association, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bartroff, J. (2018). Multiple hypothesis tests controlling generalized error rates for sequential data. *Statistica Sinica*, *28*, 363–398.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Addison-Wesley.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Chan, H. P. (2017). Optimal sequential detection in multi-stream data. *The Annals of Statistics*, *45*, 2736–2763.

Chen, H. (2019). Sequential change-point detection based on nearest neighbors. *The Annals of Statistics*, *47*, 1381–1407.

Chen, H., & Zhang, N. (2015). Graph-based change-point detection. *The Annals of Statistics*, *43*, 139–176.

Chen, J., Zhang, W., & Poor, H. V. (2020). A false discovery rate oriented approach to parallel sequential change detection problems. *IEEE Transactions on Signal Processing*, *68*, 1823–1836.

Chen, Y., & Li, X. (in press). Compound sequential change point detection in multiple data streams. *Statistica Sinica*.

Chen, Y., Lu, Y., & Moustaki, I. (in press). Detection of two-way outliers in multivariate data and application to cheating detection in educational tests. *Annals of Applied Statistics*.

Choe, E. M., Zhang, J., & Chang, H.-H. (2018). Sequential detection of compromised items using response times in computerized adaptive testing. *Psychometrika*, *83*, 650–673.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, *99*, 96–104.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, *23*, 1–22.

Efron, B. (2012). *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press.

Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, *96*, 1151–1160.

Florescu, I. (2014). *Probability and stochastic processes*. John Wiley & Sons.

Lee, Y.-H., & Haberman, S. J. (2021). Studying score stability with a harmonic regression family: A comparison of three approaches to adjustment of examinee-specific demographic data. *Journal of Educational Measurement*, *58*, 54–82.

Lee, Y.-H., & Lewis, C. (2021). Monitoring item performance with CUSUM statistics in continuous testing. *Journal of Educational and Behavioral Statistics*, *46*(5), 611–648.

Lee, Y.-H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika*, *78*, 557–575.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erbaum Associates.

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

Mei, Y. (2010). Efficient scalable schemes for monitoring a large number of data streams. *Biometrika*, *97*, 419–433.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *41*, 100–115.

Qian, J., & Li, S. (2021). Model adequacy checking for applying harmonic regression to assessment quality control. *ETS Research Report Series*, *2021*, 1–26.

Qu, Y., Huo, Y., Chan, E., & Shotts, M. (2017). Evaluating the stability of test score means for the TOEIC speaking and writing tests. *ETS Research Report Series*, *2017*, 1–8.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In J. Neyman (Ed.), *Proceedings of the second Berkeley symposium on mathematical statistics and probability* (pp. 131–149). University of California Press.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*, 527–535.

Robbins, H. (1956). An empirical bayes approach to statistics. In J. Neyman (Ed.), *Proceedings of the third Berkeley symposium on mathematical statistics and probability* (pp. 157–163). University of California Press.

Roberts, S. (1966). A comparison of some control chart procedures. *Technometrics*, *8*, 411–430.

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. Van Nostrand.

Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, *8*, 22–46.

Song, Y., & Fellouris, G. (2019). Sequential multiple testing with generalized error control: An asymptotic optimality theory. *The Annals of Statistics*, *47*, 1776–1803.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Veerkamp, W. J., & Glas, C. A. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, *25*, 373–389.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.

Xie, Y., & Siegmund, D. (2013). Sequential multi-sensor change-point detection. *The Annals of Statistics*, *41*, 670–692.

Yan, D., Von Davier, A. A., & Lewis, C. (2016). *Computerized multistage testing: Theory and applications*. CRC Press.

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods. *The Annals of Statistics*, *31*, 379–390.

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.

Authors

YUNXIAO CHEN is an assisant professor of statistics at London School of Economics and Political Science, London, UK, WC2A 2AE; email: y.chen186@lse.ac.uk. His research interests include large-scale item response data analysis, measurement and predictive modeling based on dynamic behavioral data, and sequential design of dynamic systems with applications to educational assessment and learning.

YI-HSUAN LEE is a principal research scientist at Educational Testing Service, 660 Rosedale Road, MS 12T, Princeton, NJ 08541; email: ylee@ets.org. Her primary research interests include analysis of timing and process data, test security, quality control of assessment, item response theory, and equating and linking.

XIAOOU LI is an assistant professor of statistics at the University of Minnesota, Minneapolis, MN 55455; email: lix-x1766@umn.edu. Her research interest includes latent variable models, sequential analysis, psychometrics, and applied probability.

Manuscript received September 23, 2020
First revision received April 14, 2021
Second revision received September 27, 2021
Accepted October 22, 2021