Psychometrika (2026), **91**, 1191–1231
doi:10.1017/psy.2026.10115

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

A Robust Bootstrap-CUSUM Method for Detecting Test Speededness and Pinpointing Changepoints in Computerized Testing

Baoqun Chang[1], Jing Lu[1] and Jiwei Zhang[2]

[1]School of Mathematics and Statistics, Key Laboratory of Applied Statistics of MOE, Key Laboratory of Big Data Analysis of Jilin Province, Northeast Normal University, Changchun, Jilin, China; [2]Faculty of Education, Key Laboratory of Applied Statistics of MOE, Northeast Normal University, Changchun, Jilin, China

Corresponding authors: Jing Lu; Email: luj282@nenu.edu.cn; Jiwei Zhang; Email: zhangjw713@nenu.edu.cn

(Received 16 December 2025; revised 4 April 2026; accepted 17 April 2026)

Abstract

Test speededness, caused by time constraints, can impact examinees' performance, leading to decreased response accuracy, particularly toward the end of the test. Most existing methods for detecting test speededness rely on specific distributional assumptions for response times (RTs), such as the lognormal distribution, which may lead to incorrect statistical inference if the true data distribution deviates from these assumptions. This article proposes a novel Bootstrap-CUSUM method for detecting test speededness, which is robust to non-normality in log-RTs. By constructing a cumulative sum (CUSUM) person-fit statistic for log-RTs and using the multiplier bootstrap to estimate its empirical distribution, our method facilitates individual-level detection and changepoint estimation. We prove the theoretical consistency of the method under both null and alternative hypotheses. Simulation studies show that the Bootstrap-CUSUM method outperforms the likelihood ratio test, Wald test, and score test in terms of correct classification rate, true detection rate, and false positive rate, demonstrating superior robustness and adaptability across different data distributions. The real data analysis further demonstrates the practical utility of the proposed method for detecting test speededness.

Keywords: changepoint detection; CUSUM statistics; multiplier bootstrap; response times; test speededness

1. Introduction

In the context of computerized testing, response times (RTs) have become widely used as an important auxiliary information for detecting various types of aberrant test-taking behaviors. Researchers have explored the use of RTs to identify test speededness (Schnipke & Scrams, 1997; van decr Linden et al., 1999), item pre-knowledge (van der Linden & Guo, 2008), and low motivation or lack of effort (Wise & Kong, 2005). Among these, test speededness is particularly noteworthy, as it occurs when time limits impact examinees' performance, even though speed is not a construct that the test aims to measure (Shao et al., 2016). Test speededness can lead to a decrease in response accuracy, particularly toward the end of the test, or result in missing responses (Goegebeur et al., 2010). Additionally, it may distort the estimation of both item parameters and latent abilities (Douglas et al., 1998; Lu et al., 2024), thereby threatening the validity of the test results. Moreover, the presence of speededness can change or diminish

© The Author(s), 2026. Published by Cambridge University Press on behalf of Psychometric Society.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial licence (https://creativecommons.org/licenses/by-nc/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

the relationship between the test scores and other relevant factors (Cheng & Shao, 2022). To ensure the security and reliability of the test, it is therefore crucial to identify and address response patterns influenced by test speededness.

To address the challenges posed by test speededness, various methods have been proposed to detect test speededness using RTs. The first detection method is based on residual analysis. This method involves constructing a parametric model for RTs, such as the lognormal model proposed by van der Linden (2006). It then determines whether test speededness exists by comparing the differences between predicted RTs and observed RTs. Specifically, an examinee can be considered to exhibit test speededness if their observed RT significantly deviates from the model's predicted value, which is typically indicated by large residuals that are statistically different from zero and exceed a predefined threshold, such as several standard deviations (van der Linden & van Krimpen-Stoop, 2003).

The second method involves constructing a mixture hierarchical model or using a latent class model to detect test speededness, particularly rapid guessing behavior, which is a form of test speededness. Specifically, Wang and Xu (2015) adopted a mixture model to simultaneously consider RTs and response accuracy, aiming to distinguish between two primary behaviors that test-takers may exhibit during the test: solution behavior and rapid guessing behavior. Solution behavior typically occurs when test-takers carefully consider each question, whereas rapid guessing behavior typically arises under time pressure or lack of effort, resulting in hurriedly answering the remaining items, with their responses being close to random chance. In addition, Fox and Marianti (2016) proposed another method by constructing a latent growth model to account for variable working speed. The premise of this model is that a test-taker's response speed is not constant throughout the test. By introducing random effects to model each test-taker's speed process, this model allows test-takers to adjust response speed during the test, which may manifest as linear or quadratic growth. The goal of this method is to capture changes in response speed at different stages of the test.

The third method for detecting test speededness is changepoint analysis (CPA). This approach identifies significant shifts in RT patterns, particularly changes in a test taker's working speed. The main purpose of CPA is to determine whether a changepoint exists, which indicates a noticeable shift in working speed that is usually reflected in RTs. In the likelihood ratio test (LRT) version of CPA, the LRT compares the likelihoods of two models. The null hypothesis assumes constant speed (no changepoint), and the alternative hypothesis assumes a speed change at a certain point. A significant likelihood ratio indicates a change in speed, and the changepoint location can be estimated by selecting the point that maximizes the LRT statistic (Cheng & Shao, 2022; Shao et al., 2016; Sinharay, 2016). The Wald test directly compares the estimated working speeds before and after the changepoint to evaluate whether the difference is significant (Cheng & Shao, 2022; Sinharay, 2016). The score test evaluates whether working speed differs across stages by examining the ratio of the first derivative of the log likelihood function with respect to the speed parameter to the corresponding Fisher information (Sinharay, 2017). More recently, Lu et al. (2026) proposed three Schwarz information criterion (SIC)-based CPA procedures using responses only, RTs only, or both jointly, and showed that the joint procedure generally achieved the best overall performance for detecting test speededness, while the SIC-based CPA procedures were also computationally more efficient than the likelihood ratio and Wald tests.

Within the general CPA framework, the cumulative sum (CUSUM) procedure is one of the most widely used methods. Originally proposed by Page (1954), CUSUM has been extensively applied in statistical process control and CPA, where it has proven particularly effective for detecting small shifts in the mean of measured variables. Its sensitivity to small changes has made CUSUM an important tool in many statistical applications. Over time, both the theoretical foundations and practical implementations of CUSUM have been further developed in recent studies (Ciuperca, 2017; Gao et al., 2025; Sinharay, 2016; Wang & Ning, 2025; Yu & Cheng, 2022). In the field of educational and psychological measurement, the CUSUM method has also demonstrated strong potential when applied to item response data. For example, Sinharay (2016) proposed three CPA-based statistics and compared them with four CUSUM-based indices. Yu and Cheng (2022) later conducted a systematic evaluation of 12 CUSUM-based statistics and 3 CPA-based procedures for detecting test speededness.

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

In CUSUM-based changepoint detection, the choice of critical value is critical, as it directly affects the sensitivity and specificity of the test. Several methods have been proposed for calculating critical values to infer test speededness. For instance, Shao et al. (2016) derived critical values for the LRT statistic through data permutation, while Sinharay (2016) relied on asymptotic results (Andrews, 1993), and Worsley (1979) and Yu and Cheng (2022) employed Monte Carlo simulations. However, these traditional methods have notable limitations in practice. Permutation methods can be computationally intensive for long tests or large samples, and they require that the null hypothesis holds when estimating critical values; if the data actually contain a changepoint, the method may fail. In contrast, asymptotic methods and Monte Carlo simulations require a priori assumptions about the underlying data distribution, and critical values are estimated based on these assumed distributions. Asymptotic critical values may also be inaccurate when changepoints occur near boundaries (Andrews, 1993; Sinharay, 2016), while Monte Carlo simulations, though flexible, are less efficient near critical points, making it difficult to obtain precise thresholds (Singha et al., 2023).

To address these limitations, the Bootstrap method has been increasingly adopted to estimate the empirical distribution of the CUSUM statistic, thereby obtaining the critical value for the test. Unlike traditional methods, Bootstrap does not require any assumptions about the data distribution. Moreover, even when the null hypothesis is not strictly valid, the multiplier Bootstrap can still be used to estimate the empirical distribution of the CUSUM statistic, providing a robust and flexible way to determine critical values. For example, Yu and Chen (2021) applied the multiplier Bootstrap to a high-dimensional mean-change model, ensuring theoretical validity and consistency when detecting single or multiple changepoints. Similarly, Horowitz and Savin (2000) demonstrated that the multiplier Bootstrap can substantially reduce bias in finite samples, improving the accuracy of critical value estimation.

Building on existing research, our study proposes an innovative Bootstrap-CUSUM CPA method for detecting test speededness in examinees' RT data. We construct an individual-level CUSUM statistic based on RT data and employ the multiplier Bootstrap method to estimate its empirical distribution. This makes it possible to determine the critical value based on the empirical distribution, without relying on specific distributional assumptions for RTs or on the asymptotic normality of the CUSUM statistic. The procedure is applied to each examinee to obtain an individual-specific critical value, and the averaged critical value across examinees is then computed. When an examinee's CUSUM statistic exceeds this averaged critical value, they are identified as exhibiting test speededness, and the specific item at which this behavior begins can be pinpointed. Our method enables the simultaneous detection of test speededness for each examinee and the estimation of changepoint locations. This approach not only allows for a more detailed and flexible analysis of test speededness, but also provides a methodologically rigorous and empirically grounded tool for studying test-taking behavior.

Strengths of the Bootstrap-CUSUM method in detecting test speededness

Bootstrap-CUSUM offers several key advantages over residual analysis, mixture models, latent growth models, and CPA methods (including LRT, Wald, and score tests) when detecting test speededness. These advantages arise not only from its data-driven nature but also from its ability to effectively control false positives, provide accurate changepoint estimation, ensure theoretical consistency, and adapt to the complexities of real-world data.

Traditionally, residual analysis and CPA methods typically require a pre-specified RT distribution model (e.g., the lognormal model) or a joint response and RT model. While these methods allow for interpretability through model parameters, their performance is highly sensitive to the correctness of distributional and structural assumptions. If the true data distribution deviates from these assumptions, the statistical properties may be distorted, leading to inflated Type I error rates, reduced statistical power, or even misclassification of results (Berk, 1966; Box, 1953). Furthermore, mixture models and latent growth models are explicit modeling methods used to detect test speededness. Their primary advantage is their direct ability to model and interpret changes in response speed. However, these models

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

are heavily dependent on accurate model specification. For instance, the number of latent classes, the functional form of growth curves, and the covariance structure must either be pre-specified or determined through complex model selection procedures. Misspecification in any of these areas can lead to poor model fit, incorrect classifications, or biased test-taking behavior trajectory estimates (Nylund et al., 2007). Moreover, these methods often require large sample sizes for stable estimation; when the number of examinees or items is small, model convergence can be unstable, and parameter estimates may exhibit large variances.

Compared to the methods discussed above, the Bootstrap-CUSUM method offers several distinct advantages. First, the Bootstrap-CUSUM method does not require specifying a particular parametric distribution on RTs, although it does require the item time intensity parameters to be relatively homogeneous so that changes in observed RTs can interpreted as changes in speed. More importantly, its critical values are not derived from a parametric reference distribution or from the asymptotic normality of the CUSUM statistic. Instead, they are obtained through a multiplier bootstrap that directly approximates the empirical sampling distribution of the extreme-value statistic K_i conditional on the observed sequence. This allows the procedure to avoid relying on asymptotic normal approximations and to adapt, to some extent, to irregular tail behavior and data fluctuations in practice. Second, unlike traditional methods, it does not require explicit specification of latent speed classes or functional forms, which enhances its adaptability to diverse data characteristics. This flexibility allows the method to be applied to RT data under broader distributional conditions, including complex or misspecified distributions, making it a more robust tool in practical settings where distributional assumptions are often violated. Furthermore, the method avoids parameter estimation, which mitigates the risk of inference errors due to parameter misspecification. Traditional methods typically rely on parameter estimates based on predefined models, which can lead to inaccurate conclusions if the model assumptions are not fully satisfied. In contrast, the Bootstrap-CUSUM method is entirely data-driven: for each examinee, RTs (typically log-transformed to stabilize variance) are analyzed using a mean-shift changepoint model to compute the maximum CUSUM statistic. This method allows for a more adaptive and accurate representation of the data. Given that the null distribution of the CUSUM statistic is analytically intractable, the Bootstrap-CUSUM method employs the multiplier bootstrap to empirically approximate its distribution and estimate the critical value. This resampling-based approach avoids reliance on closed-form distributional assumptions and is less sensitive to non-normality in log-RTs. The ability to obtain critical values empirically further enhances the robustness of the method, particularly in complex real-world settings.

In addition, the proposed Bootstrap-CUSUM method offers several notable advantages in theoretical properties, the detection of test speededness, and changepoint estimation. First, in terms of the theoretical properties, the Bootstrap-CUSUM method provides separate consistency guarantees for its two inference steps. Under the null hypothesis, the multiplier bootstrap distribution converges to the true sampling distribution in large samples, thereby ensuring accurate control of the significance level. Under the alternative hypothesis, we establish that the changepoint estimator obtained by maximizing the CUSUM statistic converges in probability to the true changepoint location, thus providing rigorous theoretical guarantees for localization accuracy. Second, in terms of detecting test speededness, comprehensive performance comparisons based on simulation studies conducted under multiple distributional conditions, as well as varying test lengths and proportions of speededness, indicate that the Bootstrap-CUSUM method consistently achieves the highest correct classification rate (CCR), the lowest false positive rate (FPR), and a true detection rate (TDR) comparable to or superior to that of competing methods. This demonstrates a well-balanced performance characterized by high accuracy and low false alarm rates. Moreover, the Bootstrap-CUSUM method avoids reliance on complex latent structures (e.g., the number of classes or the form of growth trajectories), requiring only the observed RT sequence for analysis. This substantially reduces the impact of model misspecification on inference results, while its computation remains straightforward. This method not only maintains broad applicability across different distributional forms of RT data, but also accounts for heteroscedasticity in the error term. In contrast, traditional approaches, such as model-based residual analysis, mixture models, latent growth

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

models, as well as LRTs, score tests, and Wald tests, often require additional steps or adjustments to ensure that the data satisfy their underlying statistical assumptions. Finally, with respect to changepoint estimation, the Bootstrap-CUSUM method can also pinpoint the location of the changepoints. When examinee is identified as exhibiting test speededness, the location of the changepoint can be estimated by maximizing the test statistic. This feature enables the method not only to detect test speededness at individual level but also to accurately pinpoint the item location at which it occurs.

The remainder of this article is organized as follows. In Section 2, changepoint detection models are introduced. Section 3 introduces the proposed Bootstrap-CUSUM method, detailing the formulation of the CUSUM statistic, the computation of critical values via the multiplier bootstrap, and the estimation of changepoint locations in log-RTs. Section 4 establishes the consistency of the Bootstrap-CUSUM method for estimating both the critical value and the changepoint location under the null and alternative hypotheses. Section 5 presents four simulation studies assessing the detection performance of the proposed method under three RT distributions: multivariate normal, multivariate *t*, and contaminated Gaussian, including gradual and abrupt changepoint scenarios, heterogeneity in item time-intensity parameters, and different critical value selection strategies. Section 6 provides an empirical data analysis. Finally, Section 7 concludes with a summary of the findings, a discussion of strengths and limitations, and future research directions.

2. Changepoint analysis models

In this section, we introduce the generalized heteroskedastic mean-shift log-RT (GHMS-LRT) model for detecting test speededness. Before formally presenting this model, we first provide a brief review of the log-normal RT model.

2.1. Log-normal response time model

van der Linden (2006) proposed a lognormal model for RTs. Let T_{ij} represent the RT of examinee i ($i = 1, \ldots, n$) on item j ($j = 1, \ldots, J$). T_{ij} is assumed to follow a lognormal distribution:

$$\log(T_{ij}) = \beta_j - \tau_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N\left(0, \sigma_j^2\right), \tag{1}$$

where τ_i denotes the latent speed parameter of the ith examinee, with larger values corresponding to faster responses. The parameter β_j characterizes the time intensity of item j, where a greater β_j indicates that the item requires more time to complete. σ_j^2 quantifies the dispersion of the lognormal distribution.

2.2. Generalized heteroskedastic mean-shift log-RT model

Within the framework of van der Linden's (2006) lognormal RT model, the GHMS-LRT model is used to detect test speededness. The GHMS-LRT model extends the traditional lognormal RT model by embedding it into a heteroskedastic mean-shift structure, thereby enabling flexible distributional assumptions while explicitly modeling test speededness through changepoint detection.

For a given examinee i, the GHMS-LRT model is specified as

$$\log(T_{ij}) = \beta_j - \tau_i + \delta_i I(j > m_i) + \varepsilon_{ij}, \quad i = 1, \ldots, n; \quad j = 1, \ldots, J, \tag{2}$$

where $m_i \in \{2, 3, \ldots, J-2\}$ denotes the changepoint location for examinee i, and δ_i is an individual-level shift parameter indicating whether examinee i exhibits test speededness. If $\delta_i = 0$, the RT sequence contains no changepoint, indicating stable responding, and we set $m_i = J$. If $\delta_i < 0$, the mean log-RT decreases after m_i, implying acceleration and thus the presence of test speededness. In this case, the examinee is assumed to respond at a normal speed to items $1, \ldots, m_i$, and to exhibit test speededness from item $m_i + 1$ to item J.

Accordingly, the hypotheses for detecting test speededness in examinee i are formulated as

$$\mathcal{H}_0 : \delta_i = 0 \quad \text{(no test speededness)} \qquad \text{vs.} \qquad \mathcal{H}_1 : \delta_i < 0 \quad \text{(test speededness present)}. \tag{3}$$

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

Remark. In model (2), we assume that the item time intensity parameters β_j are relatively homogeneous across items. That is, item time intensities fluctuate only slightly to moderately around a common mean and do not, by themselves, induce noticeable changes in the RT sequence. In other words, although β_j may vary within a certain range across items, such variation should not be large enough to create essentially shift in response speed described in model (2), such as the two distinct speed levels between the first m_i items (i.e., τ_i) and the remaining $J - m_i$ items when $\delta_i < 0$ (i.e., $\tau_i - \delta_i$). Thus, when the time intensities β_j are relatively homogeneous, changepoint detection for the log RT sequence in the changepoint model (2) can be approximately reduced to detecting changes in the individual speed parameter τ_i.

This assumption is also consistent with common simulation designs in the existing literature, researchers typically generate β_j from distributions with small variance, yielding homogeneous time intensity parameters. For example, Lu et al. (2020) set the variance of β_j to 0.2. Cheng and Shao (2022), when constructing likelihood ratio and Wald tests based on RTs to detect test speededness, followed previous work and further fixed the variance strictly at $1/9 \approx 0.111$. In addition, Wang and Xu (2015) assess the performance of mixture model under different levels of time intensity parameters, they generated β_j from the three uniform distributions $U(-0.25,0.25)$, $U(0,0.5)$, and $U(0.25,0.75)$, each with variance $1/48 \approx 0.021$. In this study, to evaluate the effectiveness of the proposed method and the feasibility of the homogeneous assumption on β_j, we consider two fluctuation scenarios: Simulation Study I sets the variance of β_j to 0.2, whereas Simulation Study II follows Wang and Xu (2015) and draws β_j from the three uniform distributions described above, each with variance $1/48 \approx 0.021$.

Note that the relative homogeneity condition does not require β_j to be exactly equal across all items or to have an extremely small variance. Instead, the key requirement is that β_j should not exhibit pronounced ordered or segmented structural changes along the item sequence; otherwise, item characteristics may induce a systematic drift that resembles a changepoint, thereby confounding item-driven changes with speed-change signals. Before applying our method, users may conduct simple diagnostics of the data and test structure to assess whether structural heterogeneity in item time intensity is likely. Specifically, (1) checking whether item types are highly consistent (e.g., all multiple-choice or all constructed-response items); (2) checking whether item types or reading passages appear in "blocks" or in an obviously segmented order along the test; and (3) if a substantial risk of structural segmentation exists, considering stratified analyses by item type or adjusting for item-order effects.

The error term ε_{ij} is introduced to capture random variation in log-RTs. Our model does not assume any specific distribution for ε_{ij}; it only requires that $E(\varepsilon_{ij}) = 0$ and $Var(\varepsilon_{ij}) = \sigma_j$. Nevertheless, because RT data are typically positive, right-skewed, and heterogeneous across items, adopting flexible and robust distributional forms can better accommodate different data characteristics. Accordingly, we allow $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iJ})^{\mathrm{T}}$ to follow one of the following distributions:

1. *Multivariate normal distribution*: Appropriate when RTs are approximately symmetric after log-transformation, with moderate variability across items. Assume that the random error term $\boldsymbol{\varepsilon}_i \sim MVN(0, \boldsymbol{V})$, where $\boldsymbol{V} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_J^2)$.
2. *Multivariate t distribution*: Suitable for heavy-tailed RT data, where some items or examinees may produce unusually long RTs that deviate from the bulk of the data. Assume that the random error term $\boldsymbol{\varepsilon}_i$ follows a J-dimensional multivariate t-distribution with degrees of freedom $\nu > 0$ and a positive-definite scale matrix $\boldsymbol{V}$, denoted as $\boldsymbol{\varepsilon}_i \sim t_\nu(\mathbf{0}, \boldsymbol{V})$ with the probability density function

$$f(\boldsymbol{x}; \nu, \boldsymbol{V}) = \frac{\Gamma\left(\frac{\nu+J}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{J/2}\det(\boldsymbol{V})^{1/2}}\left(1 + \frac{\boldsymbol{x}^{\top}\boldsymbol{V}^{-1}\boldsymbol{x}}{\nu}\right)^{-\frac{\nu+J}{2}}. \tag{4}$$

When $\nu > 2$, the covariance matrix of $\boldsymbol{\varepsilon}_i$ exists and is given by

$$\boldsymbol{\Sigma} = \frac{\nu}{\nu - 2}\boldsymbol{V}.$$

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

In practical applications, it is rare for data to conform strictly to normality. Compared with the multivariate normal distribution, the multivariate t distribution has heavier tails, meaning that its probability mass decays more slowly in the extremes. This heavy-tailed property reduces the influence of aberrant or extreme observations on CPA, thereby improving the accuracy of changepoint detection. In educational and psychological assessments, the test length for an individual examinee is often finite (referred to as finite-sample). The degrees-of-freedom parameter of the multivariate t distribution can be flexibly tuned, enabling it to more accurately approximate the true data distribution while maintaining high robustness under conditions of finite-sample and distributional uncertainty.

3. *Contaminated Gaussian distribution*: A flexible mixture model designed to represent situations in which normal and aberrant observations coexist within the data. Let the random error term $\boldsymbol{\varepsilon}_i$ follow a contaminated Gaussian distribution, denoted as ctm-Gaussian$(\varphi, \nu, \boldsymbol{V})$, with contamination proportion $\varphi \in [0,1]$, degrees of freedom ν, and scale matrix $\boldsymbol{V}$. Its probability density function is given by

$$f(\boldsymbol{x};\varphi,\nu,\boldsymbol{V}) = (1-\varphi)\frac{\exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{V}^{-1}\boldsymbol{x}\right)}{(2\pi)^{J/2}\det(\boldsymbol{V})^{1/2}} + \varphi\frac{\Gamma\left(\frac{\nu+J}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{J/2}\det(\boldsymbol{V})^{1/2}}\left(1+\frac{\boldsymbol{x}^\top \boldsymbol{V}^{-1}\boldsymbol{x}}{\nu}\right)^{-\frac{\nu+J}{2}}. \quad (5)$$

Equation (5) shows that contaminated Gaussian distribution is a convex combination of the multivariate normal distribution $MVN(\mathbf{0}, \boldsymbol{V})$ and the multivariate t distribution $t_\nu(\boldsymbol{V})$, with φ governing the contamination proportion. The covariance matrix of $\boldsymbol{\varepsilon}_i$ is

$$\boldsymbol{\Sigma} = \left[(1-\varphi) + \varphi\frac{\nu}{\nu-2}\right]\boldsymbol{V}, \quad \nu > 2,$$

ensuring a finite covariance structure when $\nu > 2$.

The contaminated Gaussian distribution is an extension of the multivariate normal distribution that more effectively accommodates the non-standard scenarios commonly encountered in practice. It can be broadly applied across various statistical contexts, including clustering, detection of aberrant observations, and latent variable modeling (Punzo & Tortora, 2021; Tomarchio et al., 2022). This distribution is particularly well-suited to situations where data contain aberrant or noisy observations, allowing for a more accurate representation of the underlying data distribution.

It should be noted that the three distributions mentioned in this article for the error term, namely, the multivariate normal distribution, the multivariate t distribution, and the contaminated Gaussian distribution, are not intended as distributional assumptions for parametric inference. Instead, they are used to construct simulated data scenarios with different tail properties and variability, to systematically evaluate the performance and robustness of the proposed method under various data-generating mechanisms. In other words, the proposed method does not rely on any specific form of data distribution. These distributions are used solely for data generation in simulation studies and are not distributional assumptions required for applying the proposed method. Accordingly, the method does not require recovering these distributions. This distinguishes the proposed method from traditional parametric approaches and constitutes one of its main advantages.

3. Bootstrap-CUSUM method

To determine whether a changepoint exists in an examinee's log-RTs under the GHMS-LRT model (Equation (2)), this section introduces the CUSUM statistic to measure changes in response speed across the examinee's RT sequence and presents the implementation of the multiplier Bootstrap method for estimating the critical value for the test.

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

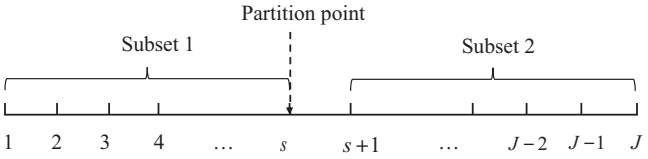


Figure 1. Response time segmentation for the *i*-th examinee, illustrating the partition between subset 1 and subset 2 at the item location *s* (i.e., partition point).

3.1. CUSUM statistic

To investigate potential changes in the RT characteristics of examinees, we partition the RT sequence of the *i*-th examinee at the *s*-th item location. The resulting subsets are defined as Subset 1 $(T_{i1}, T_{i2}, \ldots, T_{is})$ and Subset 2 $\left(T_{i(s+1)}, T_{i(s+2)}, \ldots, T_{iJ}\right)$, as illustrated in Figure 1.

If examinee *i* does not exhibit test speededness, then after partitioning at any location *s*, the means of the two resulting subsets should remain approximately equal. To conduct the hypothesis testing, we construct a CUSUM statistic. For examinee *i*, given the log RT sequence $\{\log T_{i1}, \log T_{i2}, \ldots, \log T_{iJ}\}$, define the cumulative deviation up to item location *s* as

$$\Delta_i(s) = \sum_{j=1}^{s} \left(\log(T_{ij}) - \overline{T_i}\right), \text{ where } \overline{T_i} = \frac{1}{J}\sum_{j=1}^{J} \log(T_{ij}). \tag{6}$$

Normalizing by segment length yields the standardized CUSUM statistic:

$$Z_i(s) = \sqrt{\frac{J}{s(J-s)}}\,\Delta_i(s). \tag{7}$$

Substituting the terms in Equation (7) by Equation (6) gives

$$Z_i(s) = \sqrt{\frac{J-s}{Js}} \sum_{j=1}^{s} \log(T_{ij}) - \sqrt{\frac{s}{J(J-s)}} \sum_{j=s+1}^{J} \log(T_{ij}), \tag{8}$$

which corresponds to Equation (8). Under $\mathcal{H}_0$, $Z_i(s)$ fluctuates around 0. A systematic departure of $Z_i(s)$ from 0 indicates a significant difference between the means of the two subsets, suggesting a structural change in the log-transformed RT sequence.

Since the exact changepoint location *s* is unknown, it is necessary to evaluate all possible locations $s \in \{2,3,\ldots,J-2\}$. A localized deviation in $Z_i(s)$ may occur due to random noise; however, if a true changepoint exists, the deviation will be maximized near the actual shift. Therefore, the test statistic for the *i*-th examinee is defined as the maximum absolute value of the CUSUM statistic:

$$K_i = \max_{s\in\{2,\ldots,J-2\}} |Z_i(s)|. \tag{9}$$

This definition is supported by several reasons. First, because the changepoint location is unknown, taking the maximum across all *s* ensures that the strongest evidence of departure is captured without assuming prior knowledge of the changepoint. Second, the use of the absolute value allows the method to detect both acceleration (faster responses) and deceleration (slower responses) in RTs. Finally, in changepoint detection theory, the maximum of standardized CUSUM statistics is recognized as a well-established and powerful method for identifying structural breaks (Csörgö & Horváth, 1997; Page, 1954).

3.2. Multiplier bootstrap method

The key issue is how to determine the magnitude of the statistic K_i and use it to assess whether an examinee exhibits test speededness. A common method is to set a critical value that distinguishes

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

normal examinees from aberrant ones. In hypothesis testing, several methods have been proposed for obtaining critical values. For example, Shao et al. (2016) suggested deriving critical values for the LRT statistic through data permutation; Sinharay (2016) recommended using the asymptotic critical values provided by Andrews (1993); and Worsley (1979) and Yu and Cheng (2022) employed Monte Carlo simulations to estimate critical values. However, these traditional methods have certain limitations in practice. Permutation methods are often computationally inefficient in the context of long tests; the applicability of asymptotic critical values may be restricted when the changepoint occurs at the beginning or the end of a test (Andrews, 1993; Sinharay, 2016); and the efficiency of MCMC methods can decrease significantly near critical points, resulting in slow simulations (Singha et al., 2023).

To overcome these drawbacks, this study introduces the multiplier Bootstrap method to estimate critical values. Compared with traditional methods, this method offers several notable advantages. First, it can provide more accurate finite-sample critical values, which is particularly important for complex statistical models or finite-sample data. For instance, Horowitz and Savin (2000) demonstrated that the multiplier Bootstrap method substantially reduces finite-sample bias in obtaining critical values for hypothesis testing, thereby improving test accuracy. Second, traditional statistical inference methods often rely on assumptions that the data follow standard distributions, such as the normal or t-distribution. In contrast, the multiplier bootstrap method employs resampling techniques to construct the empirical distribution of a statistic directly from the observed sample, thereby avoiding inference bias that may arise when the population distribution deviates from traditional assumptions. This approach is particularly valuable for handling complex or non-standard data distributions and has been successfully applied to various sophisticated models, including U-processes and multivariate time series (Chen & Kato, 2020; Kojadinovic & Yi, 2024). Therefore, compared with traditional methods, the multiplier Bootstrap method exhibits greater flexibility and accuracy in determining critical values and is particularly suitable for the purposes of this study.

Let $Y_{i1}, Y_{i2}, \ldots, Y_{iJ}$ be independent random variables from the standard normal distribution $N(0,1)$ and independent of $\log(T_{i1}), \log(T_{i2}), \ldots, \log(T_{iJ})$. Define the left and right means at the partition point

$$\bar{L}_{is}^{-} = \frac{1}{s}\sum_{j=1}^{s}\log(T_{ij}) \quad \text{and} \quad \bar{L}_{is}^{+} = \frac{1}{J-s}\sum_{j=s+1}^{J}\log(T_{ij}).$$

Therefore, to construct the test statistic under the null hypothesis, we follow the approach of Yu and Chen (2021) and introduce the following multiplier Bootstrap-CUSUM statistic:

$$Z_i^*(s) = \sqrt{\frac{J-s}{Js}}\sum_{j=1}^{s} Y_{ij}\Big[\log(T_{ij}) - \bar{L}_{is}^{-}\Big] - \sqrt{\frac{s}{J(J-s)}}\sum_{j=s+1}^{J} Y_{ij}\Big[\log(T_{ij}) - \bar{L}_{is}^{+}\Big], \tag{10}$$

and the corresponding maximum statistic is defined as $K_i^* = \max_{s\in\{2,\ldots,J-2\}}|Z_i^*(s)|$. The implementation of the multiplier Bootstrap for estimating the critical value of the CUSUM statistic proceeds is as follows:

Generate multipliers: For a given examinee i, in each bootstrap replication $b \in \{1,\ldots,B\}$, we generate J independent random variables $Y_{i1}^{(b)}, Y_{i2}^{(b)}, \ldots, Y_{iJ}^{(b)} \sim N(0,1)$. These serve as multipliers to re-weight the centered log-RTs.

Construct Bootstrap-CUSUM statistics: Using Equation (10), compute the Bootstrap-CUSUM statistic $Z_i^{*(b)}(s)$ for all partition points $s \in \{2,\ldots,J-2\}$ and obtain the corresponding maximum

$$K_i^{*(b)} = \max_{s\in\{2,\ldots,J-2\}}\left|Z_i^{*(b)}(s)\right|. \tag{11}$$

Repeat Bootstrap resampling: Repeat the above two steps B times, producing a sequence $\{K_i^{*(1)}, K_i^{*(2)}, \ldots, K_i^{*(B)}\}$ that approximates the sampling distribution of K_i under the null hypothesis of no test speededness.

Estimate empirical critical values: For each examinee i, the empirical $(1-\alpha)$ quantile of $\{K_i^{*(1)}, \ldots, K_i^{*(B)}\}$ is denoted as $q_i(1-\alpha)$. To obtain a general critical value across examinees, define the

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

average empirical quantile as

$$q^{*}(1-\alpha) = \frac{1}{n}\sum_{i=1}^{n} q_i(1-\alpha). \tag{12}$$

Decision rule: For examinee i, if $K_i > q^{*}(1-\alpha)$, then the null hypothesis of no test speededness is rejected, and examinee i is classified as exhibiting test speededness; otherwise, the examinee is regarded as responding at a normal pace.

Figure 2 shows the specific process of generating critical values.

4. Theoretical properties

This section provides theoretical support for the proposed method of first determining whether speededness exists and then detecting the changepoint. First, we demonstrate that the distribution of the maximum CUSUM statistic constructed via the multiplier Bootstrap can consistently approximate the sampling distribution of the original statistic under the null hypothesis, thereby ensuring that the selected critical values are statistically valid and capable of correctly controlling the type I error rate. Second, conditional on an examinee being identified as speeded, we further show that estimating the changepoint location by maximizing the CUSUM statistic is consistent in large samples, meaning that as the number of items J increases, the estimator converges to the true changepoint location.

4.1. Validity of multiplier bootstrap critical values

Under the null hypothesis $\mathcal{H}_0$ (no test speededness), our goal is to prove that the distribution of the maximum statistic K_i^{*} constructed via the multiplier Bootstrap consistently approximates the distribution of the original maximum statistic K_i. If this holds, then the empirical quantiles of K_i^{*} can serve as valid critical values for K_i, thereby guaranteeing correct control of the significance level in large samples.

Assumption 1. There exist constants $C > 0$ and $\underline{b} > 0$ such that

1. $|\log(T_{ij})| \le C$, that is, the log-RTs are bounded;
2. $\mathrm{Var}(\varepsilon_{ij}) = \sigma_j^2 \ge \underline{b}$, and under $\mathcal{H}_0$, we can write $\log(T_{ij}) = \beta_j - \tau_i + \varepsilon_{ij}$ with $\mathbb{E}(\varepsilon_{ij}) = 0$.

Assumption 1 prevents the CUSUM statistic from diverging by ensuring boundedness of the log-RTs, and it avoids degeneracy by requiring a strictly positive error variance. In practice, this is natural because examinees' RTs are finite in time-limited tests, and the variance of such data cannot be zero.

Theorem 1. *Suppose that $\mathcal{H}_0$ and Assumption 1 hold. If there exists $v > 0$ such that*

$$v^{-3}J^{-1/2}(\log J)^2 = o_p(1),$$

then

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}(K_i \le t) - \mathbb{P}(K_i^{*} \le t)\right| \to 0.$$

Theorem 1 establishes that the multiplier Bootstrap statistic K_i^{*} consistently reproduces the distribution of K_i under the null hypothesis. Consequently, the empirical $(1-\alpha)$ quantile of K_i^{*} can be used as a valid critical value $q(1-\alpha)$ for K_i. In other words, applying the multiplier Bootstrap critical values to determine whether an examinee exhibits test speededness guarantees correct control of the significance level in large samples, thereby avoiding systematic false alarms.

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

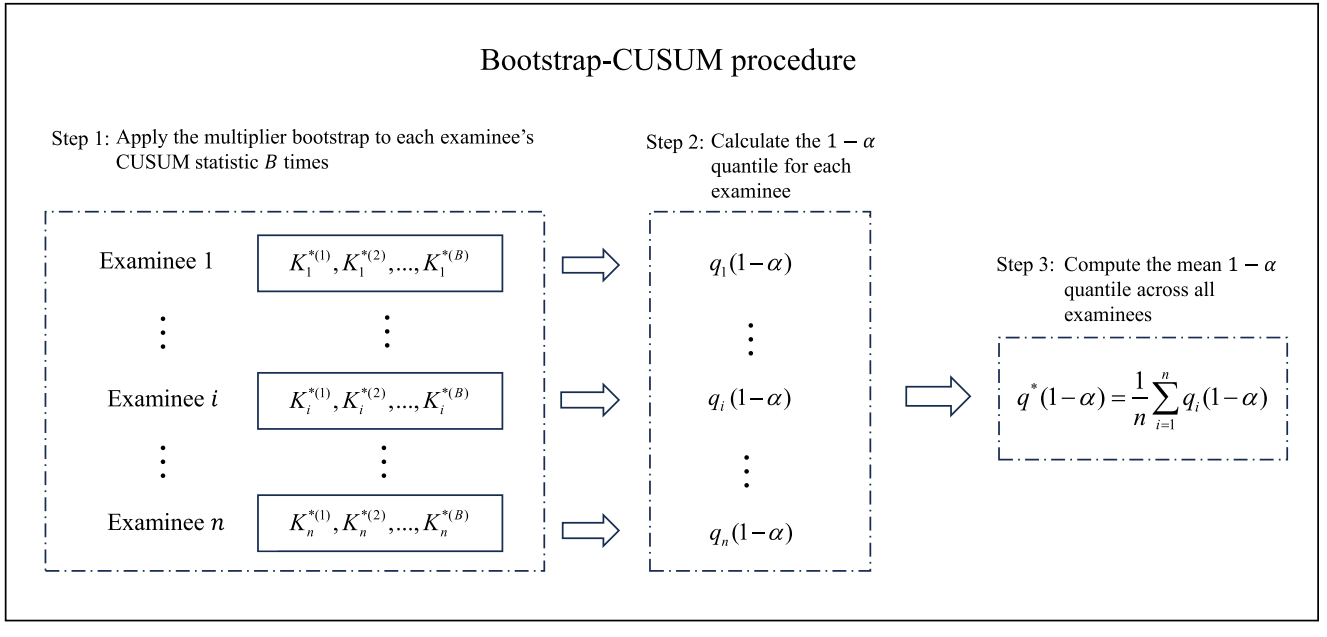


Figure 2. Bootstrap critical value generation flowchart.

4.2. Consistency of changepoint estimation

Conditional on an examinee being identified as speeded, the next task is to estimate the changepoint location for test speededness, that is, the index at which the acceleration in RTs begins. We aim to prove that the estimator

$$\hat{m}_i = \underset{s\in\{2,\ldots,J-2\}}{\arg\max} |Z_i(s)|$$

is consistent for the true changepoint location m_i^*. For ease of interpretation, we normalize the location indices by defining $\hat{\varrho}_i = \hat{m}_i/J$ and $\varrho_i^* = m_i^*/J$.

Assumption 2. For the log-RT sequence $\{\log(T_{ij}) : 1 \le j \le J\}$ of examinee i, there exists a constant $C > 0$ such that

$$Var\left\{\sum_{j=s}^{d}\log(T_{ij})\right\} \le C \tag{13}$$

holds uniformly for all $1 \le s \le d \le J$.

Assumption 2 ensures that the cumulative variability of log-RTs over any interval remains bounded, thereby preventing extreme local fluctuations from dominating the CUSUM process. Similar conditions appear in Kokoszka and Leipus (1998), who propose the more general requirement

$$Var\left\{\sum_{j=s}^{d}\log(T_{ij})\right\} \le C(d-s+1)^{\kappa}, \quad 0 \le \kappa < 2, \tag{14}$$

with Assumption 2 corresponding to the special case $\kappa = 0$.

Theorem 2. *Suppose that $\mathcal{H}_1$ (test speededness exists) and Assumption 2 hold. Then, for the i-th aberrant examinee and any $\varepsilon > 0$, we have*

$$\mathbb{P}\left(|\hat{\varrho}_i - \varrho_i^*| > \varepsilon\right) \le \frac{C\log J}{\varepsilon^2 J}.$$

Theorem 2 establishes that the estimated changepoint location $\hat{\varrho}_i$ is consistent for the true location ϱ_i^*, with the probability of deviation greater than ε shrinking at a near $1/J$ rate (up to a logarithmic factor). This guarantees that once an examinee is classified as test speededness, the changepoint estimate will converge to the true changepoint as the test length increases, thereby ensuring the reliability of the localization procedure. The detailed proofs of Theorems 1 and 2 are provided in the Supplementary Material.

5. Simulation studies

In this section, we conduct four simulation studies. Simulation Studies I and II are designed to evaluate the effectiveness and robustness of the proposed Bootstrap-CUSUM method in detecting test speededness under gradual and abrupt changepoint scenarios, respectively. Simulation Study III extends these baseline settings by introducing heterogeneity in the item time intensity parameters $\beta_j (j = 1,2,\ldots,J)$, to examine the performance of the method when the relative homogeneity condition is violated. Simulation Study IV further investigates how different critical value selection strategies affect the detection performance of the proposed procedure. Across all four studies, the method is examined under a range of conditions, including different data-generating mechanisms for test speededness, proportions of speeded examinees, and RT distributions, so as to reflect complex settings commonly encountered in educational and psychological assessments.

We also assess the method's performance under various RT distributions. Specifically, we model the random errors as multivariate normal, multivariate t, and contaminated Gaussian distributions,

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

allowing us to examine sensitivity to departures from normality. These scenarios reflect complex, realistic conditions where errors are often non-normal in real data, thereby providing a more comprehensive assessment of the Bootstrap-CUSUM method.

Additionally, by comparing the Bootstrap-CUSUM method with three commonly used CPA-based tests (LRT, Wald test, and score test), as well as the person-fit statistic χ_{pf} method proposed by Sinharay (2018) for RTs, the simulation study provides a benchmark for evaluating the strengths and limitations of the proposed method. It should be noted that the LRT, Wald test, score test, as well as the method of Sinharay (2018), all assume that the item parameters β_j and α_j are known, while the speed parameter τ_i needs to be estimated. Following Sinharay (2018), we use maximum likelihood estimation to estimate τ_i. Moreover, the method of Sinharay (2018) is designed only to detect aberrant behavior at individual level and does not provide an estimate of the changepoint location. Therefore, for the method of Sinharay (2018), we restrict the comparison to its detection performance at individual level. Detailed descriptions of the four comparative methods and the estimation of parameter τ_i are provided in ' Section 2.1 of the Supplementary Material.

5.1. Evaluation criteria

The following evaluation criteria were calculated to show the detection results.

Individual level: We use the CCR, TDR, and FPR to evaluate the performance of the methods in determining examinees whether exhibit changepoints. Specifically, CCR represents the correct classification rate for examinees with or without changepoints, TDR denotes the proportion of examinees with changepoints that were correctly flagged, and FPR was the ratio between incorrectly flagged examinees with changepoints and the total examinees without changepoints. Let $\gamma_i = 1$ indicate that the examinee i has changepoint (exhibiting test speededness), and $\gamma_i = 0$ indicate that the examinee does not have changepoint (exhibiting normal behavior); $\hat{\gamma}_i = 1$ and $\hat{\gamma}_i = 0$ indicate that the examinee flagged as speeding or normal behavior. The CCR, TDR, and FPR at individual level can be expressed as

$$CCR = \frac{\sum_{i=1}^{n} I\left(\{\gamma_i = 1\}\&\{\hat{\gamma}_i = 1\}\right) + \sum_{i=1}^{n} I\left(\{\gamma_i = 0\}\&\{\hat{\gamma}_i = 0\}\right)}{n},$$

$$TDR = \frac{\sum_{i=1}^{n} I\left(\{\gamma_i = 1\}\&\{\hat{\gamma}_i = 1\}\right)}{\sum_{i=1}^{n} I\{\gamma_i = 1\}},$$

$$FPR = \frac{\sum_{i=1}^{n} I\left(\{\gamma_i = 0\}\&\{\hat{\gamma}_i = 1\}\right)}{\sum_{i=1}^{n} I\{\gamma_i = 0\}}.$$

Changepoint estimation recovery: To evaluate the accuracy of the Bootstrap-CUSUM method in estimating the changepoint locations, we employ the mean and SD of the absolute lag for changepoint estimation (i.e., $Mean(|lag|)$ and $SD(|lag|)$). The absolute lag is defined as the absolute difference between the estimated and true changepoint locations, that is, $|lag_r| = |\hat{m}_r - m_r|$, where $\hat{m}_r$ and m_r denote the estimated and true locations of the rth changepoint, respectively. The $Mean(|lag|)$ and $SD(|lag|)$ can be represented as

$$Mean(|lag|) = \frac{1}{R}\sum_{r=1}^{R}|lag_r|,$$

$$SD(|lag|) = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left(|lag_r| - \frac{1}{R}\sum_{r=1}^{R}|lag_r|\right)^2},$$

where R represents the total number of true changepoints.

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

5.2. Simulation Study I

This simulation evaluates the effectiveness of the Bootstrap-CUSUM method in detecting gradual shifts in test speededness, compared to several CPA-based tests, that is, the LRT, Wald test, and score test, as well as the person-fit statistic χ_{pf} method proposed by Sinharay (2018) for RTs. We use the gradual change RT model (Cheng & Shao, 2022) to simulate subtle shifts in RTs, which allows to assess the method's sensitivity to small deviations in RTs. Additionally, we evaluate the robustness of the Bootstrap-CUSUM method under different RT distributional assumptions, that is, multivariate normal, multivariate *t*, and contaminated Gaussian distributions to reflect real-world scenarios.

5.2.1. Simulation design

We consider several varying factors, including the number of items, changepoint locations, speededness parameters, and different error distributions. We adopt the gradual change model of Cheng and Shao (2022) to generate RT data. Specifically, the RT of examinee i for item j, denoted as T_{ij}, is given by

$$\log(T_{ij}) = (\beta_j - \tau_i + \varepsilon_{ij}) \times \min\left(1, \left[1 - \left(\frac{j}{J} - \eta_i\right)\right]^{\lambda_i}\right), \quad i \in \{1, \ldots, n\} \text{ and } j \in \{1, \ldots, J\}, \tag{15}$$

where η_i $(0 \le \eta_i \le 1)$ indicates the stage of the test at which examinee i begins to speed up. For example, $\eta_i = 0.6$ means that examinee i speeds up during the last 40% of the test. The speededness rate parameter λ_i controls the rate at which the $\log(T_{ij})$ decreases after the speeding point. A larger λ_i results in a faster decrease in RT for examinee i.

We fixed the sample size at $n = 500$ and set the number of multiplier Bootstrap samples to $B = 1{,}000$. The number of items was $J = 20$, $J = 50$, and $J = 80$, with corresponding time limits of 50, 75, and 120 minutes, respectively. Following Shao et al. (2016), the parameter η_i was generated from a Beta distribution with mean values ($\text{Mean}(\eta_i) = 0.6$ and 0.7) and variance values ($\text{Var}(\eta_i) = 0.01$ and 0.04). To simulate more extreme forms of speededness, we further considered $\text{Mean}(\eta_i) = 0.5$, 0.6, 0.7, and 0.8, combined with variance parameters $\text{Var}(\eta_i) = 0.001$, 0.01, and 0.04. Accordingly, the changepoint location was defined as $m_i = \lceil J \times \eta_i \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function. The proportions of speededness were set to 10% and 30%. For the speededness rate parameter, we follow Shao et al. (2016) and generate $\lambda_i \sim logN(3.912, 1)$. The item and individual parameters follow the respective multivariate normal distributions

$$\begin{pmatrix} b_j \\ \beta_j \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 3.6 \end{pmatrix}, \begin{pmatrix} 1 & 0.15 \\ 0.15 & 0.2 \end{pmatrix}\right), \quad \begin{pmatrix} \theta_i \\ \tau_i \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.4 \\ 0.4 & 0.25 \end{pmatrix}\right),$$

where b_j and β_j represent the item difficulty and time intensity parameters, respectively, and θ_i and τ_i represent the individual ability and speed parameters.

Scenario 1: The random error term $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ})$ follows a multivariate normal distribution, that is, $\boldsymbol{\varepsilon}_i \sim MVN(0, \boldsymbol{V})$, where $\boldsymbol{V} = \text{diag}(\alpha_1^{-2}, \alpha_2^{-2}, \ldots, \alpha_J^{-2})$, and $\alpha_j \sim U(1.75, 3.25)$ (Lu et al., 2020, 2024).

Scenario 2: The random error term $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ})$ follows a multivariate *t* distribution with degrees of freedom v and covariance matrix $\boldsymbol{V}$, that is, $\boldsymbol{\varepsilon}_i \sim t_v(\boldsymbol{V})$, and the probability density function of $\boldsymbol{\varepsilon}_i$ is given in Equation (4). We take $v = 6$, $\boldsymbol{V} = \text{diag}(\alpha_1^{-2}, \alpha_2^{-2}, \ldots, \alpha_J^{-2})$, where $\alpha_j \sim U(1.75, 3.25)$.

Scenario 3: The random error term $\boldsymbol{\varepsilon}_i$ follows a contaminated Gaussian distribution, that is, $\boldsymbol{\varepsilon}_i \sim ctmGaussian(\varphi, v, \boldsymbol{V})$, with a density function as shown in Equation (5). We take $v = 2$, and $\boldsymbol{V}$ as in Scenario 2.

5.2.2. Results

To evaluate the effectiveness of critical value estimation for the proposed Bootstrap-CUSUM method, we generated RT data distributions under the null hypothesis ($\mathcal{H}_0$) for Scenarios 1–3 and employed the multiplier bootstrap method to estimate the distribution of the maximum CUSUM statistic K_i for each examinee. We visualized the recovery performance of the bootstrap approximations of significance

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

Table 1. Seventy-two simulation conditions with different proportions of speeded examinees, numbers of items, and the means and variances of η

Condition	Proportion of speeded examinees	Item	Mean(η)	Var(η)	Condition	Proportion of speeded examinees	Item	Mean(η)	Var(η)
1	10%	20	0.5	0.001	37	30%	20	0.5	0.001
2	10%	20	0.5	0.010	38	30%	20	0.5	0.010
3	10%	20	0.5	0.040	39	30%	20	0.5	0.040
4	10%	20	0.6	0.001	40	30%	20	0.6	0.001
5	10%	20	0.6	0.010	41	30%	20	0.6	0.010
6	10%	20	0.6	0.040	42	30%	20	0.6	0.040
7	10%	20	0.7	0.001	43	30%	20	0.7	0.001
8	10%	20	0.7	0.010	44	30%	20	0.7	0.010
9	10%	20	0.7	0.040	45	30%	20	0.7	0.040
10	10%	20	0.8	0.001	46	30%	20	0.8	0.001
11	10%	20	0.8	0.010	47	30%	20	0.8	0.010
12	10%	20	0.8	0.040	48	30%	20	0.8	0.040
13	10%	50	0.5	0.001	49	30%	50	0.5	0.001
14	10%	50	0.5	0.010	50	30%	50	0.5	0.010
15	10%	50	0.5	0.040	51	30%	50	0.5	0.040
16	10%	50	0.6	0.001	52	30%	50	0.6	0.001
17	10%	50	0.6	0.010	53	30%	50	0.6	0.010
18	10%	50	0.6	0.040	54	30%	50	0.6	0.040
19	10%	50	0.7	0.001	55	30%	50	0.7	0.001
20	10%	50	0.7	0.010	56	30%	50	0.7	0.010
21	10%	50	0.7	0.040	57	30%	50	0.7	0.040
22	10%	50	0.8	0.001	58	30%	50	0.8	0.001
23	10%	50	0.8	0.010	59	30%	50	0.8	0.010
24	10%	50	0.8	0.040	60	30%	50	0.8	0.040
25	10%	80	0.5	0.001	61	30%	80	0.5	0.001
26	10%	80	0.5	0.010	62	30%	80	0.5	0.010
27	10%	80	0.5	0.040	63	30%	80	0.5	0.040
28	10%	80	0.6	0.001	64	30%	80	0.6	0.001
29	10%	80	0.6	0.010	65	30%	80	0.6	0.010
30	10%	80	0.6	0.040	66	30%	80	0.6	0.040
31	10%	80	0.7	0.001	67	30%	80	0.7	0.001
32	10%	80	0.7	0.010	68	30%	80	0.7	0.010
33	10%	80	0.7	0.040	69	30%	80	0.7	0.040
34	10%	80	0.8	0.001	70	30%	80	0.8	0.001
35	10%	80	0.8	0.010	71	30%	80	0.8	0.010
36	10%	80	0.8	0.040	72	30%	80	0.8	0.040

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

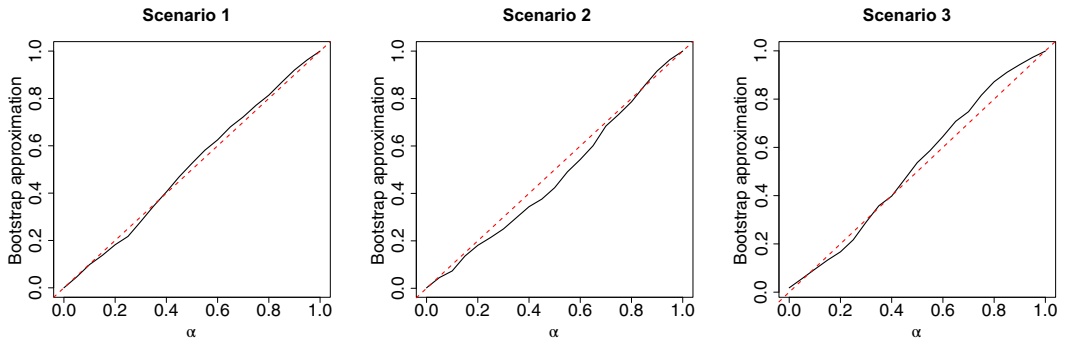


Figure 3. *Q–Q* plot comparing the true significance level α and bootstrap approximations of significance level α in Scenarios 1–3 under H_0 for J=80 and B=1,000.

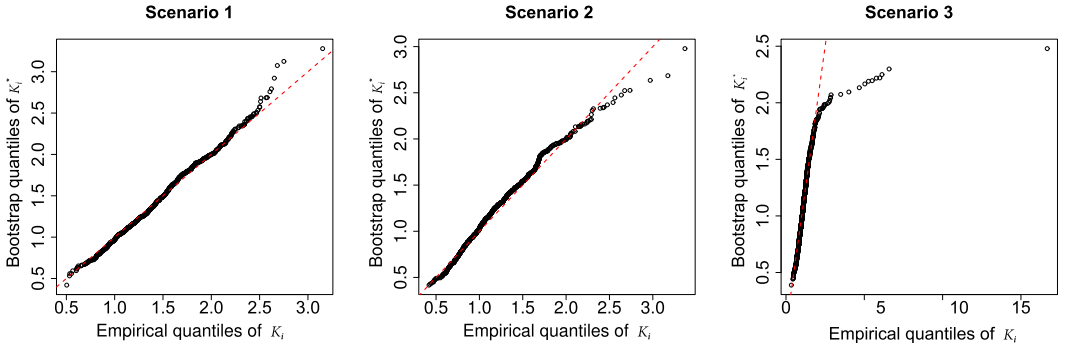


Figure 4. *Q–Q* plot comparing the empirical quantiles of K_i and bootstrap quantiles of K_i^* in Scenarios 1–3 under $\mathcal{H}_0$ for J=80 and B=1,000.

level α. Figure 3 shows the true α values on the horizontal axis against their bootstrap approximations on the vertical axis, with the red dashed line representing the $y = x$ diagonal. The closer the observed data points are to this diagonal, the more accurate the distribution estimation by the multiplier bootstrap method. From Figure 3, it is evident that across all three simulation scenarios, the bootstrap estimates increase with α while fluctuating around the diagonal, demonstrating good recovery performance. Figure 4 displays the quantile–quantile (Q–Q) plots comparing the statistics K_i and K_i^*. It should be noted that for the observed RT sequence of examinee i, $\log(t_{i1}), \ldots, \log(t_{iJ})$, we can compute only a single value of K_i. To approximate the distribution of K_i and assess how its quantiles relate to those of the Bootstrap statistic K_i^*, we repeatedly generate B simulated RT sequences using the true values of item parameters and examinee i's true speed parameter, and then compute B values of K_i to construct its empirical distribution. In Figure 4, the horizontal axis represents the empirical quantiles of K_i derived from B simulated RT sequences under the true parameter values, while the vertical axis shows the Bootstrap quantiles of K_i^* obtained from the B Bootstrap resamples of the observed RT sequence. It further demonstrates that the multiplier bootstrap method can accurately estimate the distribution of K_i across different distributional scenarios, with the highest estimation accuracy achieved in Scenario 1, followed by Scenarios 2 and 3. It should be noted that some deviations from the fitted line appear in the tail regions of all three subplots in Figure 4. This occurs due to the heavy-tailed nature of the maximum value distribution, and given the characteristics of the data generation process, minor deviations in the tails of extreme values are considered acceptable.

Figure 5 presents the line plot of the CCR at individual level, with different colors corresponding to different methods. It can be observed that the Bootstrap-CUSUM method consistently achieves CCR values closest to 1 across all three scenarios, clearly outperforming the other methods.

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

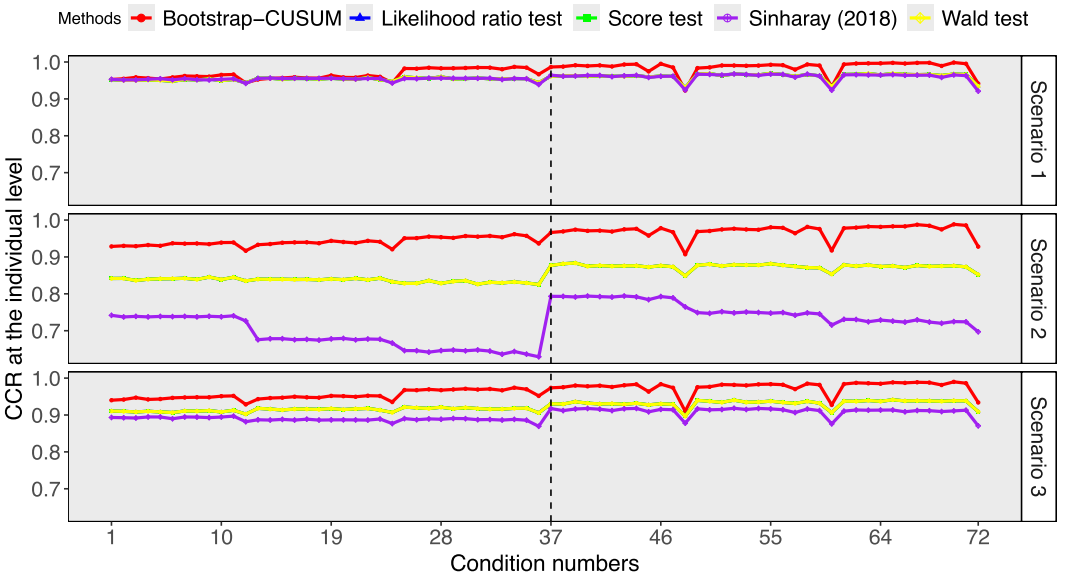


Figure 5. Line plot of CCRs for Scenarios 1–3 across different detection methods under different simulation conditions. The specific condition numbers in the *x*-axis are provided in Table 1.
Note: Sinharay (2018) refers to the person-fit statistic χ_{pf} method for RTs proposed by Sinharay (2018). The *x*-axis represents 72 conditions with different proportions of speeded examinees, numbers of items, and the medians and variances of η, while the *y*-axis shows the values of CCR. The two vertical dashed lines indicate the 37th condition: in the first 36 conditions, the proportion of speeded examinees is 10%, whereas in the latter 36 conditions, it is 30%.

Bootstrap-CUSUM exhibits stable and excellent performance under all three RT distribution assumptions, whereas the other methods perform well only in Scenario 1 under the normal distribution. Compared with Scenario 1, in Scenarios 2 and 3, where the data deviate from normality, the CCR of the remaining methods decreases substantially.

The LRT, score test, Wald test, and the method of Sinharay (2008) display an increasing trend in CCR after the 37th condition, whereas the Bootstrap-CUSUM method is much less affected by this change and consistently maintains the highest CCR. Since Conditions 1–36 correspond to speeded examinee proportion of 10% and Conditions 37–72 correspond to speeded examinee proportion of 30%, this comparison indicates that methods, such as the LRT, are more sensitive to variations in the proportion of speeded examinees, while the Bootstrap-CUSUM method is less influenced and thus demonstrates stronger stability. Overall, the Bootstrap-CUSUM method exhibits the best performance in terms of both the accuracy of test speededness identification and robustness across different RT distributions.

Additionally, Figure 5 shows that the LRT, score test, and Wald test yield nearly identical CCR values across different distributions and parameter settings, indicating high consistency among these three methods. This consistency arises due to their asymptotic equivalence under large sample sizes, as also confirmed in the simulation study by Cheng and Shao (2022). Consequently, in presenting the consistent findings, we focus on the LRT and omit separate results for the Wald and score tests due to page limit.

Figure 6 presents the FPR values for the different methods across various conditions. In Scenario 1, under the normal distribution, the Bootstrap-CUSUM method achieves the lowest FPR across all test lengths and proportions of speeded examinees. More importantly, in Scenarios 2 and 3, where the data follow a multivariate *t*-distribution and a contaminated Gaussian distribution, the FPR for the Bootstrap-CUSUM method remains around 0.05 or lower, while the FPR for the methods of Sinharay (2018) and the LRT reaches a minimum of approximately 0.1, with maximum values up to 0.4. This indicates that the Bootstrap-CUSUM method not only performs well under normality but also maintains a low FPR under non-normal distributions. In contrast, when the

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

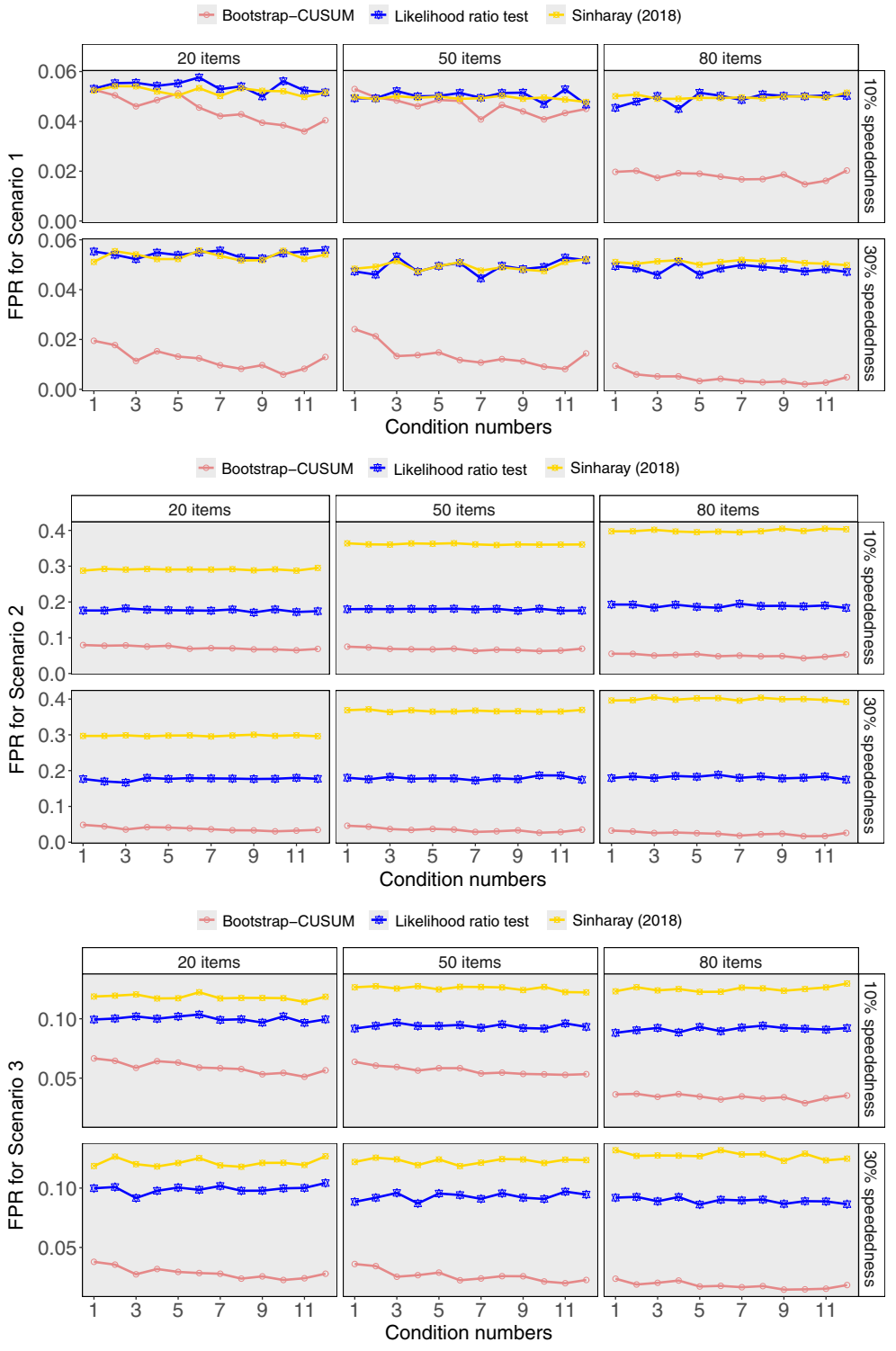


Figure 6. FPRs in Scenarios 1–3 for the three detection methods.
Note: 10% and 30% speededness indicate the proportions of speeded examinees.

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

distribution deviates even slightly from normality, the performance of Sinharay's (2018) method and the LRT significantly deteriorates. This limitation arises from the reliance of these methods on known distribution assumptions, which impacts their performance when the true distribution is unknown.

It is important to distinguish between the nominal significance level used in the proposed method and the FPR reported in the simulation results. In our procedure, the significance level $\alpha = 0.05$ is calibrated at the individual level through the multiplier bootstrap, so that the probability of falsely flagging a truly normal examinee is controlled at approximately 0.05 for the individual-level hypothesis test. By contrast, the FPR is not the Type I error rate of a single hypothesis test. Rather, it is a population-level classification measure, defined as the proportion of truly normal examinees who are incorrectly classified after all examinees are evaluated using an average critical value. As an empirical proportion aggregated over the all examinees, the FPR is not statistically equivalent to either the nominal significance level or the Type I error rate of an individual-level hypothesis test. The validity of the individual-level Type I error calibration has already been directly verified in Figure 3. Therefore, even if the individual-level hypothesis test is well controlled around $\alpha = 0.05$, it does not follow that the overall FPR must also be close to 0.05. In addition, the overall FPR may be further reduced because the proposed procedure uses a common averaged critical value across examinees, which can be more conservative than individual-specific critical values under heterogeneity. Mild conservatism may also arise from finite-sample bootstrap calibration of the upper tail of the maximum CUSUM statistic.

Table 2 presents the TDR values for the different methods under 10% and 30% speeded examinees across Scenarios 1–3. The results show that, under all conditions in these three scenarios, the TDRs of the Bootstrap-CUSUM method are close to 1, comparable to the other methods. However, when combined with the FPRs from Figure 6, it is clear that the Bootstrap-CUSUM method effectively minimizes the FPRs while maintaining high TDRs. This allows for more accurate classification between speeded and normal examinees. In contrast, although the other methods achieve relatively high TDRs for aberrant examinees, they are less effective in controlling the FPRs.

Figures 7 and 8 present line graphs of the mean and SD of the absolute lag across Scenarios 1–3. The results show that both the Bootstrap-CUSUM method and the LRT yield absolute lag values within the range of 0–1.2 items in all scenarios. Notably, the Bootstrap-CUSUM method exhibits lower mean absolute lag and smaller SD values in most conditions. This indicates that the Bootstrap-CUSUM method not only achieves greater precision in changepoint estimation but also demonstrates more stable detection performance.

5.3. Simulation Study II

This simulation is to evaluate the effectiveness of the Bootstrap-CUSUM method in detecting abrupt changes in test speededness and pinpointing changepoints across various data distributions. We adopt the RT generation manner of Wang and Xu (2015), which incorporates an abrupt-change model, to investigate the performance of the Bootstrap-CUSUM method to identify rapid and distinct shifts in test speededness.

5.3.1. Simulation design

The number of examinees is fixed at $n = 500$. The following settings were applied across all scenarios: (1) the number of test items is set to $J = 20$, $J = 50$, and $J = 80$, corresponding to time limits of 40, 100, and 160 minutes, respectively and (2) following the RT data generation manner of Wang and Xu (2015), the time intensity parameter β_j is simulated from three different uniform distributions to represent different levels of speededness: $\beta_j \sim U(-0.25, 0.25)$ (low speededness), $\beta_j \sim U(0, 0.5)$ (medium speededness), and $\beta_j \sim U(0.25, 0.75)$ (high speededness); The speed parameter τ_i is generated from a normal distribution with a mean of 0 and a variance of 0.25. The distribution of RTs resulting from test speededness C_{ij} follows a lognormal distribution

$$log(C_{ij}) \sim N(-2, 0.5^2),$$

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

Table 2. TDRs in Scenarios 1–3 for the three detection methods

Item	$M(\eta)$	$V(\eta)$	Scenario 1						Scenario 2						Scenario 3					
			10%	30%	10%	30%	10%	30%	10%	30%	10%	30%	10%	30%	10%	30%	10%	30%	10%	30%
			B-C		LRT		Sinharay (2018)		B-C		LRT		Sinharay (2018)		B-C		LRT		Sinharay (2018)	
20	0.5	0.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.010	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.040	0.999	0.997	1.000	0.999	0.999	0.999	0.999	0.996	0.999	0.999	0.999	0.999	0.998	0.997	1.000	0.999	1.000	1.000
	0.6	0.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.010	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.040	0.992	0.989	0.998	0.999	0.997	0.997	0.992	0.986	0.996	0.999	0.996	0.997	0.994	0.986	0.997	0.998	0.998	0.997
	0.7	0.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.010	0.998	0.999	1.000	1.000	0.999	1.000	0.998	0.999	1.000	1.000	0.999	1.000	0.999	0.999	1.000	1.000	1.000	1.000
		0.040	0.959	0.938	0.991	0.987	0.984	0.981	0.955	0.937	0.988	0.986	0.983	0.979	0.959	0.938	0.989	0.986	0.986	0.977
	0.8	0.001	0.997	0.998	1.000	1.000	0.999	1.000	0.995	0.998	0.999	1.000	0.998	1.000	0.999	0.998	1.000	1.000	1.000	1.000
		0.010	0.986	0.970	0.997	0.998	0.993	0.993	0.978	0.966	0.996	0.996	0.990	0.992	0.979	0.969	0.997	0.996	0.996	0.992
		0.040	0.798	0.773	0.900	0.899	0.884	0.874	0.787	0.772	0.917	0.908	0.919	0.906	0.798	0.766	0.919	0.899	0.886	0.888

(Continued)

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

Table 2. Continued

Item	$M(\eta)$	$V(\eta)$	Scenario 1						Scenario 2						Scenario 3					
			10%	30%	10%	30%	10%	30%	10%	30%	10%	30%	10%	30%	10%	30%	10%	30%	10%	30%
			B-C		LRT		Sinharay (2018)		B-C		LRT		Sinharay (2018)		B-C		LRT		Sinharay (2018)	
50	0.5	0.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.010	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.040	0.999	0.999	1.000	1.000	1.000	0.999	1.000	0.999	1.000	1.000	0.999	0.999	1.000	0.999	1.000	1.000	1.000	1.000
	0.6	0.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.010	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.040	0.993	0.995	0.997	0.999	0.995	0.997	0.993	0.994	0.998	0.999	0.995	0.997	0.993	0.995	0.998	0.999	0.996	0.997
	0.7	0.001	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.010	1.000	1.000	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.040	0.974	0.959	0.991	0.983	0.977	0.974	0.969	0.959	0.989	0.985	0.981	0.980	0.975	0.962	0.992	0.984	0.978	0.976
	0.8	0.001	0.999	0.999	0.999	1.000	0.999	1.000	0.999	0.999	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000	0.998	1.000
		0.010	0.989	0.988	0.996	0.997	0.989	0.991	0.986	0.985	0.995	0.997	0.992	0.992	0.989	0.986	0.994	0.997	0.989	0.992
		0.040	0.843	0.815	0.894	0.901	0.862	0.869	0.836	0.813	0.908	0.915	0.902	0.904	0.837	0.815	0.908	0.909	0.868	0.875

(Continued)

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

Table 2. Continued

Item	$M(\eta)$	$V(\eta)$	Scenario 1						Scenario 2						Scenario 3					
			10%	30%	10%	30%	10%	30%	10%	30%	10%	30%	10%	30%	10%	30%	10%	30%	10%	30%
			B-C		LRT		Sinharay (2018)		B-C		LRT		Sinharay (2018)		B-C		LRT		Sinharay (2018)	
80	0.5	0.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.010	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.040	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	0.999
	0.6	0.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.010	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.040	0.998	0.996	0.999	0.999	0.999	0.997	0.998	0.997	0.999	1.000	0.997	0.995	0.998	0.997	0.999	0.999	0.998	0.996
	0.7	0.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.010	0.999	1.000	1.000	1.000	1.000	0.999	0.999	1.000	1.000	1.000	0.999	0.999	0.999	1.000	1.000	1.000	1.000	0.999
		0.040	0.973	0.973	0.991	0.989	0.973	0.980	0.973	0.971	0.990	0.992	0.980	0.983	0.974	0.972	0.987	0.990	0.970	0.980
	0.8	0.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.010	0.994	0.991	0.999	0.998	0.992	0.993	0.991	0.991	0.996	0.999	0.988	0.994	0.994	0.989	0.998	0.999	0.991	0.993
		0.040	0.854	0.821	0.894	0.891	0.856	0.856	0.847	0.825	0.900	0.912	0.911	0.895	0.841	0.827	0.889	0.898	0.862	0.860

Note: Sinharay (2018) refers to the person-fit statistic χ_{pf} method for RTs proposed by Sinharay (2018). 10% and 30% represent the proportions of speeded examinees; B-C refers to the Bootstrap-CUSUM method, and LRT denotes the likelihood ratio test method.

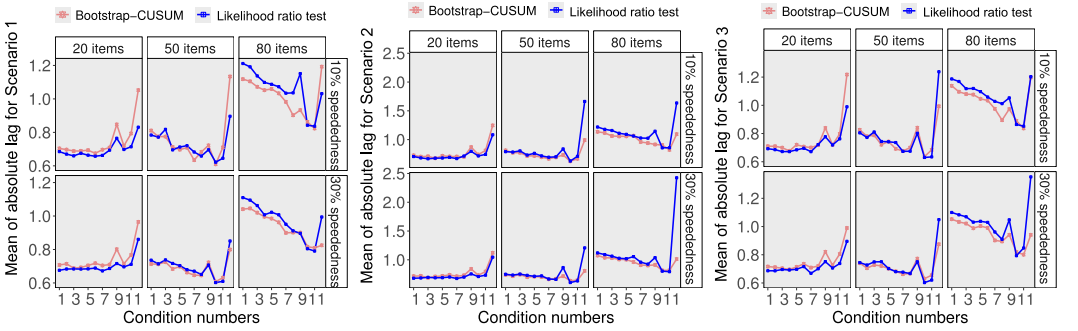


Figure 7. Mean of absolute lag in Scenarios 1–3 for the Bootstrap-CUSUM and likelihood ratio test methods.

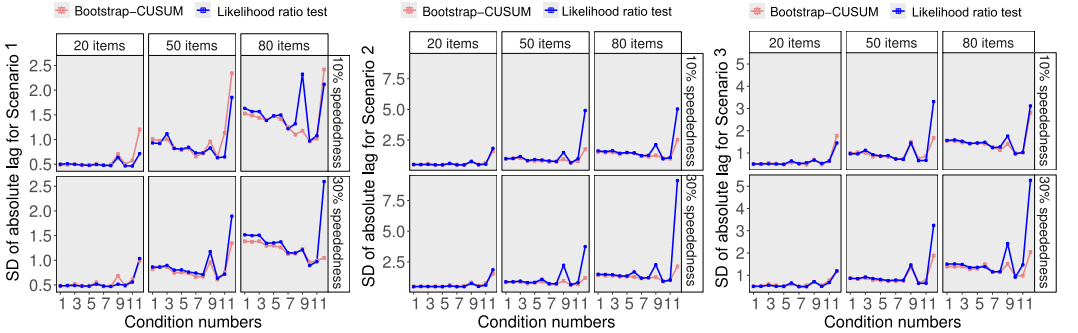


Figure 8. SD of absolute lag in Scenarios 1–3 for the Bootstrap-CUSUM and likelihood ratio test methods.

where the mean and variance of the lognormal distribution are chosen to simulate the natural variability in RTs under normal test conditions.

We assume that examinees speed up from the item at the end of the test due to time limit, resulting in a single changepoint for each examinee. Following the data generation manner of Wang and Xu (2015), we first assumed that there was no time limit and generated all examinees' RTs according to Equation (1). Then, based on each examinee's total RT, we classified them into two categories: (1) normal examinees whose total RT was less than or equal to the time limit and (2) examinees whose total RT exceeded the time limit, labeled as exhibiting test speededness. We replaced T_{ij} with C_{ij}, starting from the last item and moving backward until the total RT was less than or equal to the time limit. The item location j of the last unreplaced item for each examinee labeled as examinee's speed changepoint, that is, the changepoint location of test speededness.

For the random error term ε_i, we considered the following three different types of distributions:

Scenario 1: The random error term ε_i follows a multivariate normal distribution, that is, $\boldsymbol{\varepsilon}_i \sim MVN(0, \boldsymbol{V})$, where $\boldsymbol{V} = \sigma_j^2 \mathbf{I}_J$, with $\sigma_j = 0.5$ and $\mathbf{I}_J$ denoting the J-dimensional identity matrix.

Scenario 2: The random error term $\boldsymbol{\varepsilon}_i$ follows a multivariate t-distribution, that is, $\boldsymbol{\varepsilon}_i \sim t_v(\boldsymbol{V})$, where $v = 6$ and $\boldsymbol{V} = \sigma_j^2 \mathbf{I}_J$ with $\sigma_j = 0.5$.

Scenario 3: The random error term $\boldsymbol{\varepsilon}_i$ follows a contaminated Gaussian distribution, that is, $\boldsymbol{\varepsilon}_i \sim ctm\text{-}Gaussian(\varphi, v, \boldsymbol{V})$, its density function is given in Equation (5). In this case, $v = 6$ and $\boldsymbol{V}$ was specified identically to Scenario 2.

5.3.2. Results

Tables 3 and 4 and Table A1 in the Supplementary Material present the performance of three methods, the Bootstrap-CUSUM method, the LRT, and the method of Sinharay (2018), in detecting test speed-

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

Table 3. Detection results of test speededness in Scenario 1 for the three methods

Item	β	Method	CCR	TDR	FPR	$Mean(\|lag\|)$	$SD(\|lag\|)$
20	$U(-0.25,0.25)$	B-C	0.969	1.000	0.036	0.014	0.084
		LRT	0.953	1.000	0.053	0.009	0.055
		Sinharay (2018)	0.959	1.000	0.047	-	-
	$U(0,0.5)$	B-C	0.990	1.000	0.014	0.014	0.098
		LRT	0.964	1.000	0.049	0.009	0.073
		Sinharay (2018)	0.966	1.000	0.046	-	-
	$U(0.25,0.75)$	B-C	0.998	1.000	0.003	0.006	0.066
		LRT	0.975	1.000	0.044	0.005	0.051
		Sinharay (2018)	0.975	1.000	0.044	-	-
50	$U(-0.25,0.25)$	B-C	0.981	1.000	0.022	0.005	0.039
		LRT	0.958	1.000	0.048	0.007	0.043
		Sinharay (2018)	0.960	0.999	0.045	-	-
	$U(0,0.5)$	B-C	0.995	1.000	0.007	0.007	0.064
		LRT	0.965	1.000	0.048	0.006	0.059
		Sinharay (2018)	0.965	1.000	0.048	-	-
	$U(0.25,0.75)$	B-C	0.999	1.000	0.001	0.006	0.065
		LRT	0.973	1.000	0.049	0.005	0.059
		Sinharay (2018)	0.973	1.000	0.048	-	-
80	$U(-0.25,0.25)$	B-C	0.981	1.000	0.022	0.008	0.052
		LRT	0.954	1.000	0.053	0.008	0.053
		Sinharay (2018)	0.958	1.000	0.048	-	-
	$U(0,0.5)$	B-C	0.993	1.000	0.009	0.007	0.066
		LRT	0.961	1.000	0.053	0.009	0.070
		Sinharay (2018)	0.964	1.000	0.048	-	-
	$U(0.25,0.75)$	B-C	0.999	1.000	0.001	0.006	0.066
		LRT	0.969	1.000	0.055	0.006	0.061
		Sinharay (2018)	0.975	0.999	0.045	-	-

Note: B-C refers to the Bootstrap-CUSUM method, LRT denotes the likelihood ratio test method, and Sinharay 2018 refers to the person-fit statistic χ_{pf} method for RTs proposed by Sinharay (2018).

edness and estimating changepoint locations under different RT distribution conditions. Since the LRT, Wald test, and score test yield almost identical results across all evaluation criteria, as also confirmed in Simulation Study I, only the results of the LRT are reported in the subsequent analysis. As shown in Table 3, the proposed Bootstrap-CUSUM method outperforms the other methods across all evaluation criteria. At individual level, compared with the LRT and the method of Sinharay (2018), the Bootstrap-CUSUM method achieves the highest CCRs and TDRs, as well as the lowest FPR. Specifically, under all conditions, its CCR reaches up to 99.9%, and its TDR attains 100%. In addition, with respect to the recovery of changepoint estimation, that is, $Mean(|lag|)$ and $SD(|lag|)$, the Bootstrap-CUSUM method exhibits minimal estimation bias, with values as low as 0.5%, further confirming its high precision in

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

Table 4. Detection results of test speededness in Scenario 2 for the three detection methods

Item	β	Method	CCR	TDR	FPR	*Mean*(|*lag*|)	*SD*(|*lag*|)
20	$U(-0.25,0.25)$	B-C	0.939	0.997	0.071	0.154	0.638
		LRT	0.868	0.997	0.157	0.160	0.689
		Sinharay (2018)	0.775	0.996	0.267	-	-
	$U(0,0.5)$	B-C	0.968	0.997	0.044	0.101	0.532
		LRT	0.899	0.998	0.142	0.092	0.488
		Sinharay (2018)	0.824	0.997	0.249	-	-
	$U(0.25,0.75)$	B-C	0.988	0.997	0.019	0.059	0.382
		LRT	0.930	0.998	0.129	0.059	0.402
		Sinharay (2018)	0.875	0.998	0.232	-	-
50	$U(-0.25,0.25)$	B-C	0.945	0.999	0.065	0.150	0.843
		LRT	0.860	1.000	0.167	0.213	1.301
		Sinharay (2018)	0.703	0.998	0.353	-	-
	$U(0,0.5)$	B-C	0.972	0.999	0.039	0.096	0.586
		LRT	0.892	0.999	0.155	0.103	0.735
		Sinharay (2018)	0.767	0.996	0.333	-	-
	$U(0.25,0.75)$	B-C	0.989	0.999	0.019	0.067	0.492
		LRT	0.926	1.000	0.142	0.058	0.385
		Sinharay (2018)	0.836	0.998	0.313	-	-
80	$U(-0.25,0.25)$	B-C	0.944	1.000	0.067	0.225	1.565
		LRT	0.853	1.000	0.177	0.284	1.982
		Sinharay (2018)	0.680	0.996	0.383	-	-
	$U(0,0.5)$	B-C	0.973	1.000	0.039	0.118	0.850
		LRT	0.888	1.000	0.161	0.117	0.783
		Sinharay (2018)	0.744	0.998	0.366	-	-
	$U(0.25,0.75)$	B-C	0.991	1.000	0.016	0.051	0.402
		LRT	0.926	1.000	0.142	0.071	0.667
		Sinharay (2018)	0.822	0.996	0.340	-	-

Note: B-C refers to the Bootstrap-CUSUM method, LRT denotes the likelihood ratio test method, and Sinharay 2018 refers to the person-fit statistic χ_{pf} method for RTs proposed by Sinharay (2018).

estimating changepoint locations. The results also show that the performance of all methods improves as the time intensity parameter (β_j) increases and as the number of items grows.

Table 4 and Table A1 in the Supplementary Material show that the Bootstrap-CUSUM method not only performs well under the normal distribution assumption as shown in Table 3, but also maintains stable performance when the RT distribution follows a multivariate t distribution or a contaminated normal distribution. In contrast, the performance of the LRT and the method proposed by Sinharay (2018) is clearly affected by the RT distributions. Specifically, under non-normal distribution conditions, both methods exhibit a noticeable decrease in CCR, accompanied by a significant increase in FPR.

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

Table 5. Mixture distribution settings of three heterogeneous item time intensity parameter β_j under the framework of Simulation Study I (gradual-change scenario) and the framework of Simulation Study II (abrupt-change scenario)

Mixture	Gradual-change scenario		Abrupt-change scenario	
	Distribution 1	Distribution 2	Distribution 1	Distribution 2
Mix 1	$N(3.6,0.25)$	$N(4,0.16)$	$U(-0.25,0.25)$	$U(0.25,0.75)$
Mix 2	$N(3.2,0.09)$	$N(4,0.25)$	$U(-0.5,0)$	$U(0.75,1)$
Mix 3	$N(3.6,0.25)$	$N(4,0.16)$	$U(-0.25,0.25)$	$U(0.25,0.75)$

Note: The difference between Mix 1 and Mix 2 lies in the degree of parameter heterogeneity; the difference between Mix 1 and Mix 3 lies in whether the parameter distributions are generated with an ordered structure.

5.4. Simulation Study III

Simulation Study III extends Simulation Study I (gradual-change scenario) and Simulation Study II (abrupt-change scenario) by further incorporating heterogeneity in the item time intensity parameters $\beta_j (j = 1,2,\ldots,J)$. The aim is to systematically examine the performance and robustness of the proposed Bootstrap-CUSUM method in detecting test speededness under two types of item time intensity heterogeneity, namely, random heterogeneity and structural heterogeneity.

5.4.1. Simulation design

To examine the effect of heterogeneity in the time intensity parameters, we consider three heterogeneous scenarios for generating β_j. In the first two scenarios, denoted as Mix 1 and Mix 2, random heterogeneity is introduced by randomly assigning $\frac{I}{2}$ of the items to have β_j generated from Distribution 1 and the remaining $\frac{I}{2}$ from Distribution 2, thereby forming a heterogeneous mixture distribution. The two scenarios differ in the magnitude of the discrepancy between the two component distributions. To further investigate the impact of structural heterogeneity with a fixed ordered pattern, we consider a third scenario, Mix 3. Under this setting, the time intensity parameters for items $1,\ldots,\frac{I}{2}$ are generated from Distribution 1, whereas those for items $\frac{I}{2}+1,\ldots,J$ are generated from Distribution 2. That is, $\beta_j \sim$ Distribution 1 for $j = 1,2,\ldots,\frac{I}{2}$, and $\beta_j \sim$ Distribution 2 for $j = \frac{I}{2}+1,\frac{I}{2}+2,\ldots,J$. The heterogeneous mixture distributions used to generate β_j are summarized in Table 5.

Simulation Study III is conducted within the general frameworks of Simulation Studies I and II. Therefore, except for the specification of the time intensity parameters β_j, all other design settings remain the same as those in the corresponding baseline studies. The purpose of Simulation Study III is to evaluate how heterogeneity in β_j affects the performance of the proposed Bootstrap-CUSUM method.

Because changepoint location is not the primary focus of this study, under the framework of Simulation Study I (gradual-change scenario), we consider the representative setting $M(\eta) = 0.7$, together with $V(\eta) = 0.01$ and 0.04. These settings provide a relatively stringent evaluation, since a later changepoint combined with a larger variance makes detection more difficult. All remaining parameter settings and data generation procedures are identical to those used in Simulation Study I. Under the framework of Simulation Study II (abrupt-change scenario), only the heterogeneous specification of β_j is modified (see Table 5), while all other parameter settings and data generation mechanisms remain unchanged.

5.4.2. Result

Tables 6 and 7 present the results of Simulation Study III for the proposed Bootstrap-CUSUM method. Table 6, which corresponds to the gradual-change scenario, reveals two main patterns. First, across most parameter settings, Mix 1 shows the best detection performance, whereas Mix 3 performs the worst. Second, as the heterogeneity in β_j becomes more pronounced from Mix 1 to Mix 3, the CCR

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

Table 6. Detection results of test speededness for Bootstrap-CUSUM method in Simulation Study III under the framework of Simulation Study I (gradual-change scenario)

Proportion of speeded	Item	$V(\eta)$	Mixture	Scenario 1				Scenario 2				Scenario 3			
examinees				CCR	TDR	FPR	*Mean*(|*lag*|)	CCR	TDR	FPR	*Mean*(|*lag*|)	CCR	TDR	FPR	*Mean*(|*lag*|)
10%	20	0.01	Mix 1	0.980	1.000	0.022	0.738	0.961	1.000	0.043	0.735	0.969	0.999	0.034	0.755
			Mix 2	0.988	1.000	0.014	0.703	0.969	0.999	0.034	0.722	0.977	1.000	0.026	0.745
			Mix 3	0.847	0.999	0.170	0.679	0.857	0.999	0.159	0.676	0.855	0.997	0.161	0.683
		0.04	Mix 1	0.975	0.951	0.023	0.808	0.950	0.943	0.049	0.802	0.962	0.950	0.037	0.816
			Mix 2	0.969	0.935	0.027	0.805	0.952	0.934	0.046	0.837	0.960	0.943	0.038	0.847
			Mix 3	0.885	0.958	0.123	0.791	0.886	0.958	0.123	0.905	0.881	0.953	0.127	0.817
	50	0.01	Mix 1	0.959	0.999	0.047	0.664	0.943	0.998	0.064	0.662	0.955	1.000	0.050	0.665
			Mix 2	0.956	0.999	0.049	0.671	0.944	0.999	0.063	0.700	0.950	1.000	0.056	0.666
			Mix 3	0.738	0.999	0.296	0.704	0.773	0.999	0.258	0.717	0.749	0.999	0.284	0.732
		0.04	Mix 1	0.983	0.982	0.017	0.736	0.960	0.977	0.042	0.685	0.970	0.977	0.031	0.782
			Mix 2	0.970	0.964	0.029	0.694	0.955	0.963	0.046	0.679	0.962	0.965	0.038	0.741
			Mix 3	0.737	0.971	0.292	0.814	0.775	0.968	0.250	0.836	0.760	0.971	0.267	0.840
	80	0.01	Mix 1	0.989	0.999	0.012	0.846	0.967	0.999	0.037	0.894	0.977	0.998	0.025	0.895
			Mix 2	0.950	1.000	0.056	0.900	0.939	1.000	0.069	0.935	0.946	1.000	0.061	0.872
			Mix 3	0.592	0.998	0.459	1.034	0.632	0.999	0.416	1.075	0.602	0.999	0.449	1.122
		0.04	Mix 1	0.969	0.971	0.031	0.886	0.952	0.968	0.050	0.906	0.961	0.967	0.040	0.866
			Mix 2	0.953	0.972	0.050	0.874	0.944	0.967	0.059	0.929	0.945	0.969	0.058	0.894
			Mix 3	0.576	0.979	0.475	1.235	0.625	0.983	0.420	1.472	0.598	0.982	0.451	1.400

(Continued)

Table 6. Continued

Proportion of speeded	Item	$V(\eta)$	Mixture	Scenario 1				Scenario 2				Scenario 3			
examinees				CCR	TDR	FPR	*Mean*(|*lag*|)	CCR	TDR	FPR	*Mean*(|*lag*|)	CCR	TDR	FPR	*Mean*(|*lag*|)
30%	20	0.01	Mix 1	0.997	0.998	0.004	0.718	0.982	0.998	0.025	0.729	0.986	0.998	0.019	0.723
			Mix 2	0.990	0.997	0.012	0.708	0.979	0.995	0.028	0.729	0.983	0.997	0.023	0.713
			Mix 3	0.975	0.998	0.036	0.671	0.963	0.998	0.053	0.669	0.969	0.998	0.044	0.672
		0.04	Mix 1	0.979	0.941	0.004	0.805	0.968	0.939	0.019	0.831	0.970	0.940	0.017	0.818
			Mix 2	0.973	0.942	0.014	0.856	0.963	0.940	0.028	0.884	0.964	0.938	0.025	0.861
			Mix 3	0.930	0.930	0.070	0.761	0.925	0.924	0.074	0.796	0.931	0.928	0.068	0.801
	50	0.01	Mix 1	0.996	0.999	0.005	0.656	0.987	1.000	0.020	0.671	0.991	0.999	0.012	0.667
			Mix 2	0.993	0.999	0.009	0.672	0.981	0.999	0.028	0.696	0.987	0.999	0.018	0.669
			Mix 3	0.964	0.999	0.054	0.696	0.960	0.999	0.059	0.689	0.962	1.000	0.056	0.684
		0.04	Mix 1	0.985	0.966	0.007	0.716	0.971	0.963	0.025	0.796	0.978	0.965	0.015	0.684
			Mix 2	0.980	0.953	0.007	0.698	0.970	0.953	0.022	0.744	0.975	0.953	0.014	0.748
			Mix 3	0.880	0.961	0.159	0.736	0.887	0.960	0.150	0.800	0.885	0.961	0.153	0.764
	80	0.01	Mix 1	0.996	1.000	0.007	0.912	0.983	1.000	0.025	0.923	0.989	1.000	0.016	0.919
			Mix 2	0.991	1.000	0.014	0.928	0.981	0.999	0.028	0.945	0.984	1.000	0.024	0.952
			Mix 3	0.809	1.000	0.285	0.993	0.826	1.000	0.262	0.993	0.814	1.000	0.278	0.998
		0.04	Mix 1	0.986	0.966	0.005	0.917	0.974	0.966	0.021	0.923	0.978	0.966	0.016	0.900
			Mix 2	0.980	0.966	0.014	0.951	0.966	0.962	0.032	0.938	0.972	0.965	0.025	0.927
			Mix 3	0.844	0.978	0.220	1.248	0.850	0.976	0.211	1.177	0.852	0.976	0.208	1.138

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

Table 7. Detection results of test speededness for Bootstrap-CUSUM method in Simulation Study III under the framework of Simulation Study II (abrupt-change scenario)

Item	Mixture	Scenario 1				Scenario 2				Scenario 3			
		CCR	TDR	FPR	*Mean*(|*lag*|)	CCR	TDR	FPR	*Mean*(|*lag*|)	CCR	TDR	FPR	*Mean*(|*lag*|)
20	Mix 1	0.982	1.000	0.024	0.012	0.966	0.996	0.046	0.098	0.979	0.998	0.028	0.026
	Mix 2	0.983	0.998	0.026	0.047	0.976	0.989	0.033	0.126	0.982	0.996	0.028	0.075
	Mix 3	0.878	1.000	0.165	0.003	0.903	0.999	0.138	0.095	0.881	0.999	0.161	0.015
50	Mix 1	0.994	1.000	0.009	0.015	0.975	0.998	0.036	0.138	0.989	1.000	0.015	0.036
	Mix 2	0.989	0.999	0.018	0.049	0.984	0.998	0.028	0.169	0.990	0.999	0.017	0.072
	Mix 3	0.669	1.000	0.460	0.002	0.765	0.999	0.346	0.099	0.693	1.000	0.430	0.024
80	Mix 1	0.994	1.000	0.008	0.011	0.977	1.000	0.033	0.129	0.991	1.000	0.013	0.044
	Mix 2	0.992	0.999	0.013	0.050	0.983	0.998	0.030	0.170	0.990	0.998	0.017	0.063
	Mix 3	0.471	1.000	0.732	0.004	0.571	1.000	0.627	0.141	0.492	1.000	0.710	0.036

tends to decrease, while the FPR tends to increase. A further comparison between Mix 1 and Mix 2 shows that the differences in CCR, TDR, FPR, and $Mean(|lag|)$ are generally small. This suggests that when heterogeneity in the time intensity parameters β_j is randomly distributed across items, as in Mix 1 and Mix 2, its impact on the detection performance of the proposed method is limited. In other words, the Bootstrap-CUSUM procedure appears reasonably robust to random heterogeneity in item time intensity.

However, when the heterogeneity follows a fixed ordered structure, as in Mix 3, detection performance deteriorates substantially. In this case, the first and second halves of the test differ systematically in their time intensity parameters, creating an ordered structural shift in the RT sequence. This pattern can interfere with the CUSUM-based changepoint statistic, which is designed to detect changes in the sequence mean under a relative homogeneity condition. When the item time intensity parameters themselves exhibit stepwise or trend-like structural variation along the test sequence, the resulting RT sequence may exhibit systematic shifts that resemble genuine speed changes. As a consequence, the procedure may incorrectly interpret such structural heterogeneity as evidence of test speededness, leading to an increase in FPR and a reduction in TDR. Therefore, in practical applications, we suggest to conduct preliminary diagnostic analyses to assess whether pronounced structural heterogeneity is present before applying the proposed method.

A similar pattern is observed in Table 7, which reports the results under the abrupt-change scenario. The main difference is that the absolute lag bias of the changepoint estimator, measured by $Mean(|lag|)$, is substantially smaller in Table 7 than in Table 6. This result is consistent with the findings from Simulation Studies I and II, namely, that changepoint localization is generally more accurate under abrupt-change scenarios than under gradual-change scenarios.

In summary, the results clearly distinguish between two types of heterogeneity. Random heterogeneity in the time intensity parameters, as represented by Mix 1 and Mix 2, has little effect on overall detection performance and does not lead to systematic inflation of the FPR. In contrast, structural heterogeneity, as represented by Mix 3, introduces systematic patterns along the item sequence that confound with the CUSUM-based detection mechanism. These findings suggest that the proposed method does not require strict homogeneity of the time intensity parameters; instead, it requires that the parameter sequence not contain pronounced ordered structural discontinuities in order to maintain reliable control of false positives.

5.5. Simulation Study IV

To investigate the impact of different critical value selection strategies on the performance of the proposed Bootstrap-CUSUM method in detecting examinee speeded behavior, we further designed Simulation Study IV based on Simulation Study I (gradual-change scenario) and Simulation Study II (abrupt-change scenario).

5.5.1. Simulation design

As illustrated in Figure 2, suppose there are n examinees, and for each examinee, B multiplier Bootstrap samples are generated, resulting in a total of $n \times B$ statistics. Under the significance level $\alpha = 0.05$, we consider the following three cases to compare the effects of different critical value selection strategies:

Case 1: Per-person critical values. Each examinee is tested using an individual-level critical value, that is, the critical value for the ith examinee is $q_i(1-\alpha)$.

Case 2: Averaged critical value. All examinees share a common averaged critical value $q^*(1-\alpha)$, which is the thresholding strategy used in this article. Please see Figure 2 for details of its computation.

Case 3: Pooled critical value. The 95% quantile of all Bootstrap statistics from all examinees is used as a common critical value. Denote $K^* = \{K_1^{*(1)}, \ldots, K_1^{*(B)}, K_2^{*(1)}, \ldots, K_2^{*(B)}, \ldots, K_n^{*(1)}, \ldots, K_n^{*(B)}\}$, then the pooled critical value is the 95% quantile of K^*.

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

In this simulation study, except for the differences in critical value selection strategies and the setting of time intensity parameters, all other parameters, such as the number of examinees, remain consistent with those in Simulation Study III. Specifically, under the framework of Simulation Study I (gradual-change scenario), the time intensity parameters are the same as those in Simulation Study I. Under the framework of Simulation Study II (abrupt-change scenario), lower time intensity parameters, that is, $\beta_j \sim U(-0.25, 0.25)$, are adopted, because lower time intensity makes the detection of changepoints more difficult.

5.5.2. Result

Tables 8 and 9 present the detection results of test speededness for the Bootstrap-CUSUM method in Simulation Study IV. Table 8, which corresponds to the gradual-change scenario setting of Simulation Study I, shows that all three critical-value selection strategies achieve generally good detection performance. Among them, the per-person critical values method (Case 1) performs worst overall, whereas the averaged critical values method (Case 2) and the pooled critical values method (Case 3) produce very similar results across most conditions. More specifically, Case 2 tends to yield a slightly higher TDR than Case 3, whereas Case 3 tends to produce a slightly lower FPR than Case 2; as a result, the overall difference in CCR between these two strategies is small. This similarity becomes even more evident as the number of items increases or the proportion of speeded examinees becomes larger, suggesting that Cases 2 and 3 have comparable stability in overall classification performance. Furthermore, Table 9 shows that under the abrupt-change scenario of Simulation Study II follow a pattern similar to that observed under the gradual-change scenario of Simulation Study I, indicating that the above conclusions are stable across different simulation conditions.

To further compare Cases 2 and 3, Tables 10 and 11 report their corresponding critical values under each condition. Across all settings, the critical values in Case 2 are consistently lower than those in Case 3, indicating that Case 3 adopts a more conservative decision rule. This pattern has a straightforward statistical explanation. Case 2 is based on the average of individual-level quantiles, whereas Case 3 uses the overall 95% quantile calculated from all Bootstrap samples combined. Because the latter is taken from the pooled empirical distribution and lies further in the right tail, it is generally larger than the averaged quantile in Case 2. This difference also explains the performance pattern observed in Tables 8 and 9: the averaged critical value strategy tends to yield a higher TDR, whereas the pooled critical value strategy tends to produce a lower FPR, although the overall difference in classification performance between the two strategies remains limited.

It should also be noted that, although per-person critical values are theoretically appealing because they can achieve significance control at the individual level, they have several practical limitations. First, using different critical values for different examinees makes it difficult to implement a unified operational decision rule and may introduce seemingly unfairness in high stakes settings. Second, these critical values are more sensitive to individual random fluctuations, which may reduce classification stability. Third, the simulation results show that their overall detection performance is weaker than that of the two unified-threshold strategies. Therefore, although the per-person critical value approach is theoretically feasible, it is not well suited for practical use as a common decision rule.

The pooled critical value approach implicitly assumes that the null distributions of the test statistics are the same across examinees. When individual heterogeneity is present, pooling all Bootstrap samples produces a quantile from a mixture distribution rather than from a common homogeneous distribution. Such a pooled quantile is typically larger than the average of the individual quantiles and therefore leads to a more conservative decision threshold. As the degree of heterogeneity increases, this conservatism may become more pronounced. In extreme cases, the pooled critical value may be disproportionately influenced by individuals with larger variability, which can reduce the sensitivity of the procedure to genuine test speededness.

In contrast, the averaged critical value strategy adopted in this article first calibrates the critical values at the individual level and then aggregates them into a common threshold at the population level.

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

Table 8. Detection results of test speededness for Bootstrap-CUSUM method in Simulation Study IV under the framework of Simulation Study I (gradual-change scenario)

Proportion of speeded	Item	$V(\eta)$	Case	Scenario 1				Scenario 2				Scenario 3			
examinees				CCR	TDR	FPR	*Mean*(|*lag*|)	CCR	TDR	FPR	*Mean*(|*lag*|)	CCR	TDR	FPR	*Mean*(|*lag*|)
10%	20	0.01	Case 1	0.919	0.997	0.090	0.709	0.923	0.998	0.085	0.701	0.918	0.997	0.091	0.719
			Case 2	0.961	0.998	0.043	0.709	0.937	0.998	0.070	0.701	0.948	0.999	0.058	0.718
			Case 3	0.986	0.998	0.015	0.709	0.970	0.998	0.033	0.701	0.976	0.999	0.026	0.718
		0.04	Case 1	0.876	0.915	0.128	0.753	0.881	0.918	0.123	0.757	0.877	0.910	0.127	0.771
			Case 2	0.936	0.958	0.066	0.815	0.919	0.956	0.085	0.815	0.930	0.965	0.074	0.867
			Case 3	0.973	0.952	0.024	0.810	0.961	0.947	0.037	0.802	0.967	0.950	0.031	0.820
	50	0.01	Case 1	0.941	1.000	0.066	0.652	0.941	0.999	0.066	0.684	0.941	1.000	0.066	0.693
			Case 2	0.972	1.000	0.031	0.652	0.951	0.999	0.055	0.684	0.963	1.000	0.042	0.693
			Case 3	0.989	1.000	0.013	0.652	0.975	0.999	0.028	0.684	0.982	1.000	0.021	0.693
		0.04	Case 1	0.937	0.954	0.065	0.726	0.934	0.950	0.068	0.763	0.938	0.953	0.064	0.790
			Case 2	0.966	0.962	0.033	0.751	0.946	0.958	0.055	0.832	0.955	0.959	0.045	0.754
			Case 3	0.981	0.957	0.016	0.689	0.970	0.951	0.027	0.723	0.975	0.954	0.023	0.723
	80	0.01	Case 1	0.963	1.000	0.041	0.896	0.962	1.000	0.042	0.922	0.963	1.000	0.042	0.906
			Case 2	0.980	1.000	0.022	0.896	0.958	1.000	0.047	0.922	0.967	1.000	0.037	0.906
			Case 3	0.989	1.000	0.012	0.896	0.980	1.000	0.022	0.922	0.981	1.000	0.022	0.906
		0.04	Case 1	0.975	0.960	0.023	0.912	0.973	0.960	0.025	0.928	0.975	0.961	0.023	0.945
			Case 2	0.986	0.960	0.010	0.895	0.960	0.959	0.040	0.913	0.973	0.965	0.026	1.135
			Case 3	0.991	0.959	0.005	0.896	0.978	0.954	0.019	0.898	0.981	0.957	0.016	0.939

(Continued)

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

Table 8. Continued

Proportion of speeded examinees	Item	$V(\eta)$	Case	Scenario 1				Scenario 2				Scenario 3			
				CCR	TDR	FPR	*Mean*(|*lag*|)	CCR	TDR	FPR	*Mean*(|*lag*|)	CCR	TDR	FPR	*Mean*(|*lag*|)
30%	20	0.01	Case 1	0.936	0.997	0.090	0.709	0.933	0.996	0.094	0.726	0.935	0.994	0.091	0.714
			Case 2	0.985	0.997	0.020	0.709	0.972	0.997	0.039	0.730	0.976	0.997	0.033	0.731
			Case 3	0.994	0.995	0.007	0.709	0.987	0.994	0.016	0.721	0.989	0.995	0.014	0.729
		0.04	Case 1	0.897	0.901	0.105	0.757	0.902	0.902	0.099	0.807	0.898	0.904	0.105	0.793
			Case 2	0.972	0.942	0.016	0.804	0.957	0.940	0.036	0.848	0.960	0.939	0.031	0.830
			Case 3	0.975	0.923	0.002	0.770	0.966	0.914	0.012	0.807	0.967	0.916	0.011	0.782
	50	0.01	Case 1	0.976	0.999	0.035	0.633	0.975	1.000	0.036	0.640	0.977	0.999	0.034	0.647
			Case 2	0.999	0.999	0.002	0.633	0.986	0.999	0.020	0.640	0.989	0.999	0.015	0.646
			Case 3	0.999	0.999	0.000	0.632	0.994	0.998	0.007	0.639	0.994	0.999	0.009	0.646
		0.04	Case 1	0.962	0.954	0.035	0.711	0.960	0.951	0.036	0.743	0.962	0.955	0.035	0.741
			Case 2	0.986	0.961	0.003	0.688	0.971	0.960	0.023	0.728	0.977	0.961	0.016	0.766
			Case 3	0.985	0.952	0.000	0.685	0.978	0.949	0.008	0.701	0.979	0.950	0.008	0.736
	80	0.01	Case 1	0.959	0.999	0.059	0.919	0.961	1.000	0.057	0.936	0.960	0.999	0.058	0.926
			Case 2	0.994	0.999	0.008	0.919	0.980	0.999	0.029	0.932	0.984	0.999	0.023	0.926
			Case 3	0.998	0.999	0.003	0.919	0.992	0.999	0.011	0.931	0.993	0.999	0.010	0.925
		0.04	Case 1	0.953	0.968	0.053	0.919	0.954	0.966	0.051	0.933	0.954	0.965	0.052	0.954
			Case 2	0.986	0.965	0.005	0.904	0.970	0.966	0.028	0.926	0.974	0.964	0.021	0.927
			Case 3	0.987	0.959	0.001	0.883	0.980	0.959	0.010	0.894	0.981	0.959	0.009	0.925

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

Table 9. Detection results of test speededness for Bootstrap-CUSUM method in Simulation Study IV under the framework of Simulation Study II (abrupt-change scenario)

Item	Case	Scenario 1				Scenario 2				Scenario 3			
		CCR	TDR	FPR	*Mean*(|*lag*|)	CCR	TDR	FPR	*Mean*(|*lag*|)	CCR	TDR	FPR	*Mean*(|*lag*|)
20	Case 1	0.924	0.998	0.086	0.014	0.919	0.960	0.088	0.118	0.924	0.991	0.086	0.028
	Case 2	0.973	1.000	0.031	0.014	0.944	0.997	0.066	0.154	0.965	0.999	0.040	0.029
	Case 3	0.992	1.000	0.009	0.014	0.979	0.995	0.024	0.151	0.987	0.999	0.015	0.029
50	Case 1	0.945	0.998	0.063	0.005	0.945	0.986	0.063	0.144	0.946	0.996	0.061	0.023
	Case 2	0.981	1.000	0.022	0.005	0.945	0.999	0.065	0.150	0.972	1.000	0.032	0.027
	Case 3	0.995	1.000	0.006	0.005	0.982	0.999	0.021	0.150	0.990	1.000	0.012	0.027
80	Case 1	0.950	1.000	0.057	0.008	0.950	0.990	0.058	0.182	0.949	0.996	0.059	0.033
	Case 2	0.981	1.000	0.022	0.008	0.944	1.000	0.067	0.225	0.972	1.000	0.032	0.034
	Case 3	0.995	1.000	0.006	0.008	0.985	0.999	0.018	0.225	0.990	1.000	0.012	0.034

Table 10. Averaged critical values (Case 2) and pooled critical values (Case 3) in Simulation Study IV under the framework of Simulation Study I (gradual-change scenario)

Item	$V(\eta)$	Case	10% speeded examinees			30% speeded examinees		
			Scenario 1	Scenario 2	Scenario 3	Scenario 1	Scenario 2	Scenario 3
20	0.01	Case 2	1.704	1.824	1.763	1.980	2.087	2.039
		Case 3	1.878	2.085	2.004	2.300	2.437	2.408
	0.04	Case 2	1.649	1.767	1.707	1.928	2.038	1.981
		Case 3	1.821	2.029	1.953	2.278	2.420	2.378
50	0.01	Case 2	1.898	2.027	1.946	2.192	2.300	2.241
		Case 3	2.043	2.280	2.156	2.505	2.641	2.592
	0.04	Case 2	1.883	2.005	1.932	2.186	2.294	2.229
		Case 3	2.020	2.246	2.137	2.484	2.632	2.569
80	0.01	Case 2	1.952	2.074	2.006	2.203	2.314	2.250
		Case 3	2.090	2.327	2.229	2.521	2.674	2.616
	0.04	Case 2	1.970	2.092	2.023	2.225	2.338	2.275
		Case 3	2.099	2.331	2.234	2.545	2.700	2.646

Table 11. Averaged critical values (Case 2) and pooled critical values (Case 3) in Simulation Study IV under the framework of Simulation Study II (abrupt-change scenario)

Item	Case	Scenario 1	Scenario 2	Scenario 3
20	Case 2	1.527	1.710	1.560
	Case 3	1.742	2.100	1.817
50	Case 2	1.667	1.869	1.709
	Case 3	1.866	2.292	1.968
80	Case 2	1.736	1.936	1.772
	Case 3	1.940	2.375	2.037

This design avoids the excessive conservatism that can arise from mixture distribution quantiles, while still preserving a unified decision rule for practical use. The simulation results show that this strategy achieves higher, or at least comparable, detection power while maintaining acceptable FPR control, thereby providing a more desirable balance between statistical validity and practical applicability.

6. Empirical example

6.1. Data description

The proposed Bootstrap-CUSUM method was used to evaluate its practical efficacy for an operational dataset from the 2012–2013 10th-grade English/Language Arts (ELA) test. This assessment is relatively low-stakes, and its item development followed standard principles of standardized testing, namely, maintaining consistency in item format, response mode, and task requirements within a single test form, in order to reduce systematic time variation induced by changes in item type or administration procedures. The dataset includes 1,776 students, each with responses and RTs for 35 items. All 35 items are of the same format, multiple-choice, and are scored dichotomously (correct/incorrect). Therefore,

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

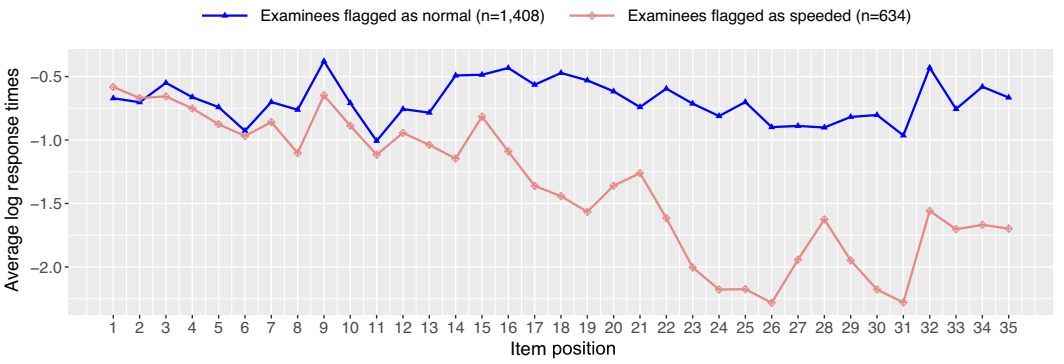


Figure 9. Line chart of the item-level average log RTs for 1,408 examinees flagged as normal and 634 examinees flagged as speeded.

it is clear that all 35 items are of the same item type (i.e., multiple-choice questions) and are consistent in terms of item format, which satisfies the assumption of relative homogeneity of item time intensity parameters for the proposed method. Previous analyses revealed that the distribution of RTs across all items and examinees is bimodal, implying the coexistence of at least two behavioral patterns: solution behavior and aberrant behavior (Wang et al., 2018). For data preprocessing, we excluded examinees with RTs equal to zero and those with a total log-transformed RT less than −2 (equivalent to a total RT shorter than 4.736 minutes). Such cases are likely to reflect unmotivated test-taking behavior across the entire assessment. After cleaning, the final sample size consisted of 1,682 examinees.

6.2. Results

We applied the Bootstrap-CUSUM method to detect test speededness among 1,682 examinees. The significance level was set at $\alpha = 0.05$, with $B = 1{,}000$ bootstrap samples. Specifically, a critical value was first computed for each examinee, and the average of these values was then used as the final decision threshold. The following findings are obtained. First, the Bootstrap-CUSUM method identified 634 examinees as exhibiting test speededness, while the remaining 1,048 examinees showing normal test-taking behavior. Second, the estimated changepoints were primarily located in the middle-to-late portion of the test, particularly between Items 15 and 28. Among all items, item 21 was associated with the largest number of examinees showing a changepoint (68 examinees), followed by item 22 (63 examinees). This pattern suggests that test speededness tends to emerge during the middle and later stages of the test.

To provide an intuitive comparison of RT patterns between examinees classified as normal and those classified as speeded, Figure 9 displays the item-level average log-RTs for the two groups identified by the detection procedure. The average log-RTs of the normal group exhibit only minor fluctuations across items, with a variance of 0.026, and remain stable overall. In contrast, the average log-RTs of speeded group show a clear downward trend, becoming especially pronounced after item 15. This provides the empirical support for the effectiveness of the proposed Bootstrap-CUSUM method in detecting test speededness.

Figure 10 presents the number of examinees exhibiting test speededness across the 35 items after the changepoint. The results show that, as the item location increases, the frequency of examinees identified as exhibiting test speededness also gradually increases, suggesting that more examinees display test speededness behavior in the later stages of the test. Figure 11 presents the log-RTs for three randomly selected examinees (the 55th, 229th, and 243rd), where the black asterisks "*" indicate the estimated changepoint locations. Noticeable changes in log-RTs occur before and after these estimated changepoints, demonstrating that the proposed Bootstrap-CUSUM method can effectively identify examinees exhibiting test speededness.

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

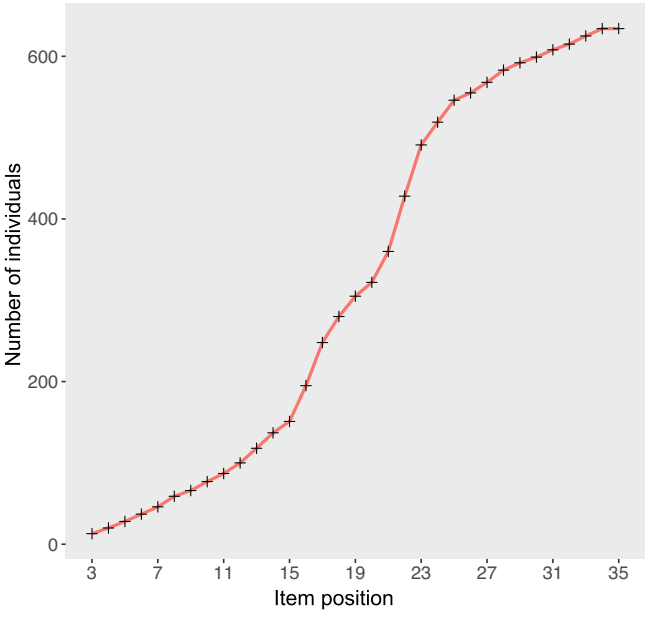


Figure 10. Number of examinees identified as exhibiting test speededness for each of the 35 items.

7. Conclusion

This article proposes a Bootstrap-CUSUM CPA method for detecting test speededness using examinees' RT data. Specifically, we construct an individual-level CUSUM statistic based on log-RTs data and employ the multiplier Bootstrap method to estimate its empirical distribution. This makes it possible to determine the critical value based on the empirical distribution, without relying on specific distributional assumptions for log-RTs or on the asymptotic normality of the CUSUM statistic. The procedure is applied to each examinee to obtain an individual-specific critical value, and the averaged critical value across examinees is then computed. When an examinee's CUSUM statistic exceeds this averaged critical value, they are identified as exhibiting test speededness; otherwise, the examinee is regarded as responding normally. For those examinees identified as speeded, the changepoint, that is, the item at which speeded responding begins, is further estimated by maximizing the CUSUM statistic. In theory, under the null hypothesis, we establish the consistency between the multiplier bootstrap CUSUM statistic and the original CUSUM statistic. Under the alternative hypothesis, we establish that the changepoint estimator obtained by maximizing the CUSUM statistic is consistent with the true changepoint. Based on two simulation studies, we find that the proposed method performs well across multiple evaluation criteria and achieves clear advantages over several commonly used methods for detecting test speededness. The analyses of real data further confirm the effectiveness of the proposed method.

Compared with the LRT, Wald test, score test, and the RTs-based aberrant behavior detection method proposed by Sinharay (2018), the proposed method has the following advantages. First, at individual level, it achieves the best performance in terms of CCR, TDR, and FPR, thereby effectively distinguishing between normal and speeded examinees. Second, the method can more accurately estimate the changepoint locations, as reflected in the smaller mean and SD of absolute lag in changepoint estimation. Third, our method is robust to non-normality in log-RTs. It can conduct testing and threshold calibration under weaker distributional conditions, thereby offering greater applicability and robustness in the presence of skewness, heavy tails, and related deviations from normality. In contrast, the LRT, Wald test, score test, and the method of Sinharay (2018) show a noticeable increase in FPR and a significant

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

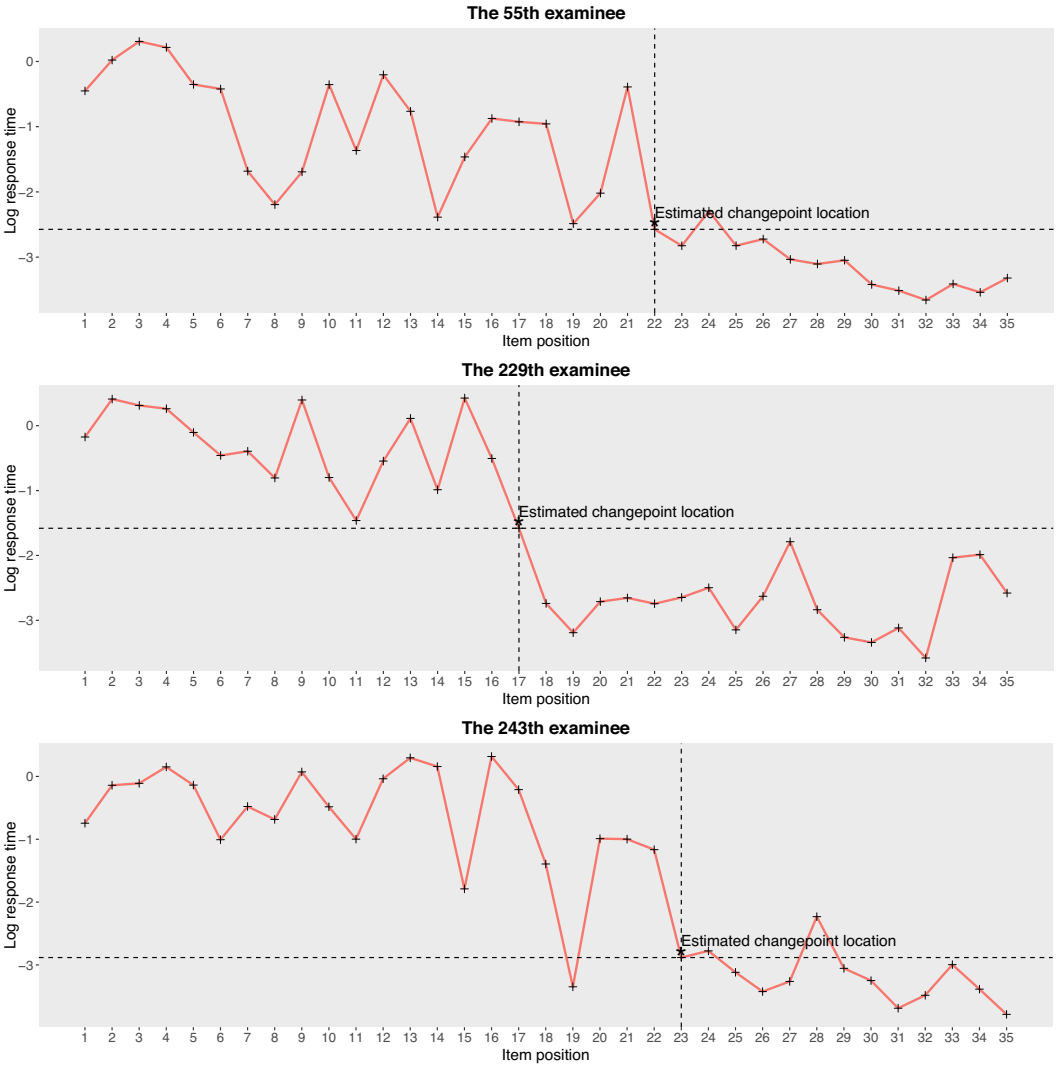


Figure 11. Changepoint estimated locations for the 55th, 229th, and 243rd examinees.

decline in CCR when the distribution deviates from normality. Finally, this study not only demonstrates the effectiveness of the method through simulation and empirical analyses but also systematically investigates its theoretical properties, thereby providing solid theoretical support for its validity and large-sample statistical properties.

Although this study has obtained encouraging results in both theoretical properties and simulation validation, several issues remain worthy of further investigation. First, in real testing contexts, examinees' response sequences may involve multiple types of aberrant behaviors. The present study focuses on test speededness, future study could investigate the performance of identifying the other aberrant behaviors, such as warm-up effects and cheating behavior.

Second, the proposed method is currently designed primarily for detecting test speededness characterized by a single changepoint. In real testing settings, however, examinees may exhibit more complex, multi-stage speededness patterns, such as speeding up, returning to normal performance, and then speeding up again. Such response behavior may produce multiple changepoints within a single response

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

sequence, which is not fully accommodated by the current framework. A possible extension for future research is to first use a BIC-based model selection procedure to estimate the number of changepoints for each examinee and then combine binary segmentation with the proposed Bootstrap-CUSUM procedure to locate them more precisely. Such an extension would improve the applicability of the method in more realistic testing contexts.

Third, although the proposed method outperforms all comparison methods in detection accuracy, it computationally more demanding. For example, in a testing scenario with 50 items under Simulation Study I, our method requires approximately 16–19 seconds, whereas the LRT, Wald, and score tests each take about 1.8 seconds. Detailed runtime comparisons are provided in the Supplementary Material. This difference is expected because the proposed method performs multiplier bootstrap calibration separately for each examinee, with B resampling iterations required for every examinee. As a result, it is naturally slower than conventional methods that compute test statistics directly. Nevertheless, from a practical perspective, the overall runtime remains within an acceptable range, especially given the substantial improvement in detection accuracy. In this sense, the method trades a moderate increase in computational cost for more reliable statistical inference. When faster computation is needed, the procedure may be further accelerated by moderately reducing B while retaining acceptable accuracy, or by adopting early-stopping or adaptive-B strategies. In addition, because the bootstrap computations for different examinees are independent, the procedure is naturally parallel and can be substantially accelerated in multi-core CPU or cluster computing environments.

Fourth, the proposed method is developed under an independence assumption across items. However, in real assessment settings, RT sequences may exhibit local dependence, for example, due to shared influences, such as reading speed, short-term fatigue, or continuity in interface operations. The method can accommodate mild dependence through bootstrap calibration, which empirically absorbs such effects. Nevertheless, when item correlations become strong, the detection performance of the Bootstrap-CUSUM method may deteriorate. Developing changepoint inference procedures that are more robust to strong item dependence therefore remains an important direction for future research.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psy.2026.10115.

Data availability statement. The real data set is subject to a confidentiality agreement with the data provider and therefore-cannot be made publicly available.

Funding statement. This research was supported by the National Social Science Fund of China on Statistics (Grant No. 25CTJ030).

Ethical standards. The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

References

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, *61*(4), 821–856.

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, *37*(1), 51–58.

Box, G. E. (1953). Non-normality and tests on variances. *Biometrika*, *40*(3–4), 318–335.

Chen, X., & Kato, K. (2020). Jackknife multiplier bootstrap: Finite sample approximations to the U-process supremum with applications. *Probability Theory and Related Fields*, *176*(3), 1097–1163.

Cheng, Y., & Shao, C. (2022). Application of change point analysis of response time data to detect test speededness. *Educational and Psychological Measurement*, *82*(5), 1031–1062.

Ciuperca, G. (2017). Real time change-point detection in a nonlinear quantile model. *Sequential Analysis*, *36*(1), 87–110.

Csörgö, M., & Horváth, L. (1997). *Limit theorems in change-point analysis*. John Wiley & Sons, Inc.

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics*, *23*(2), 129–151.

Fox, J.-P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, *51*(4), 540–553.

Gao, M., Yang, W., Li, X., & Yao, M. (2025). The consistency for CUSUM estimator of mean change-point model based on association. *Communications in Statistics-Theory and Methods*, *54*(6), 1836–1867.

Goegebeur, Y., De Boeck, P., & Molenberghs, G. (2010). Person fit for test speededness: Normal curvatures, likelihood ratio tests and empirical Bayes estimates *methodology*. *European Journal of Research Methods for the Behavioral and Social Sciences*, *6*(1), 31–36.

Horowitz, J. L., & Savin, N. (2000). Empirically relevant critical values for hypothesis tests: A bootstrap approach. *Journal of Econometrics*, *95*(2), 375–389.

Kojadinovic, I., & Yi, B. (2024). Resampling techniques for a class of smooth, possibly data-adaptive empirical copulas. *Journal of Statistical Planning and Inference*, *231*, 106132.

Kokoszka, P., & Leipus, R. (1998). Change-point in the mean of dependent observations. *Statistics & Probability Letters*, *40*(4), 385–393.

Lu, J., Wang, C., Zhang, J., & Liu, Z. (2026). Detecting test speededness using responses and/or response times: Change point analysis approaches based on schwarz information criterion. *Psychometrika*, Published online, 1–30. https://doi.org/10.1017/psy.2026.10094.

Lu, J., Wang, C., Zhang, J., & Tao, J. (2020). A mixture model for responses and response times with a higher-order ability structure to detect rapid guessing behaviour. *British Journal of Mathematical and Statistical Psychology*, *73*(2), 261–288.

Lu, J., Wang, C., Zhang, J., & Wang, X. (2024). A sequential Bayesian changepoint detection procedure for aberrant behaviours in computerized testing. *British Journal of Mathematical and Statistical Psychology*, *77*(1), 31–54.

Nylund, K. L., Asparouhov, T., & Muthen, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *41*(1–2), 100–115.

Punzo, A., & Tortora, C. (2021). Multiple scaled contaminated normal distribution and its application in clustering. *Statistical Modelling*, *21*(4), 332–358.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232.

Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika*, *81*, 1118–1141.

Singha, A., Chakrabarti, D., & Arora, V. (2023). Conditional normalizing flow for Markov chain Monte Carlo sampling in the critical region of lattice field theory. *Physical Review D*, *107*(1), 14512.

Sinharay, S. (2016). Person fit analysis in computerized adaptive testing using tests for a change point. *Journal of Educational and Behavioral Statistics*, *41*(5), 521–549.

Sinharay, S. (2017). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, *42*(1), 46–68.

Sinharay, S. (2018). A new person-fit statistic for the lognormal model for response times. *Journal of Educational Measurement*, *55*(4), 457–476.

Tomarchio, S. D., Gallaugher, M. P., Punzo, A., & McNicholas, P. D. (2022). Mixtures of matrix-variate contaminated normal distributions. *Journal of Computational and Graphical Statistics*, *31*(2), 413–421.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365–384.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 195–210.

van der Linden, W. J., & van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*(2), 251–265.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477.

Wang, C., Xu, G., & Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*(1), 223–254.

Wang, P., & Ning, W. (2025). Nonparametric CUSUM change-point detection procedures based on modified empirical likelihood. *Computational Statistics*, 1–31.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.

Worsley, K. J. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association*, *74*(366a), 365–367.

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press

Yu, M., & Chen, X. (2021). Finite sample change point inference and identification for high-dimensional mean vectors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *83*(2), 247–270.
Yu, X., & Cheng, Y. (2022). A comprehensive review and comparison of CUSUM and change-point-analysis methods to detect test speededness. *Multivariate Behavioral Research*, *57*(1), 112–133.

Cite this article: Chang, B., Lu, J. and Zhang, J. (2026). A Robust Bootstrap-CUSUM Method for Detecting Test Speededness and Pinpointing Changepoints in Computerized Testing. *Psychometrika*, 1191–1231. https://doi.org/10.1017/psy.2026.10115

https://doi.org/10.1017/psy.2026.10115 Published online by Cambridge University Press