Psychometrika (2026), **91**, 965–994
doi:10.1017/psy.2026.10094

PSYCHOMETRIC SOCIETY

APPLICATION AND CASE STUDIES - ORIGINAL

Detecting Test Speededness Using Responses and/or Response Times: Change Point Analysis Approaches Based on Schwarz Information Criterion

Jing Lu[1], Chun Wang[2], Jiwei Zhang[3] and Zefeng Liu[1,4]

[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]Colledge of Education, University of Washington, Seattle, Washington, USA; [3]Faculty of Education, Key Laboratory of Applied Statistics of MOE, Northeast Normal University, Changchun, Jilin, China; [4]ShenYang Railway Experimental High School, Shenyang, China

Corresponding author: Jiwei Zhang; Email: zhangjw713@nenu.edu.cn

(Received 14 March 2025; revised 3 February 2026; accepted 4 February 2026)

Abstract

Change point analysis (CPA) detects structural shifts in a response sequence by partitioning it into segments with different statistical properties. This paper proposes three CPA approaches based on the Schwarz information criterion (SIC; hereafter SIC-CPA): response data only, response time (RT) data only, and the combination of response and RT data, to detect the prevalent test speededness in time-limit tests. To comprehensively investigate the efficiency and accuracy of the proposed approaches, six simulation studies were conducted under diverse conditions. Simulation results demonstrate that SIC-CPA can effectively enhance the power of change point detection and reduce Type I errors, while improving computational efficiency compared to the likelihood ratio and Wald tests. Moreover, the SIC-CPA combining response and RT data outperforms the SIC-CPA based solely on RTs, and the latter is substantially superior to the SIC-CPA based solely on responses. In addition, SIC-CPA accurately identifies two change points in RT patterns, corresponding to early warm-up and later test speededness. Using an iterative detect–clean–recalibrate procedure, SIC-CPA achieves more reliable Type I error control than likelihood ratio and Wald tests when item parameters are estimated from contaminated data. A real data analysis was conducted to show the application of the proposed approaches.

Keywords: change point detection; information criterion; responses and/or response times; test speededness

1. Introduction

In educational and psychological testing, pure power tests and pure speed tests aim to measure examinees' *effective* abilities and *effective* speeds (Meijer & Sotaridona, 2006; van der Linden, 2007), respectively, under ideal testing environments. However, for the convenience of test administration, time constraints are typically imposed on the tests, resulting in what is known as time-limit tests. The degree to which the time limit of a test affects the performance of examinees is referred to as test speededness. This manifests when examinees, within the time limit, are unable to devote sufficient consideration or effort toward the end of the test (Bejar, 1985). Van der Linden (2011) argued that test

© 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.2026.10094 Published online by Cambridge University Press

speededness was the cumulative outcome of the interaction among the cognitive speed of the examinee, the workload of the items, and the test's time limit. Various definitions of test speededness exist, but they all share a common aspect: Test speededness can inadvertently affect the test scores when the time limit influences the examinee's performance, even though speed is not a core aspect of the construct that the test seeks to measure (Evans & Reilly, 1972; Schnipke & Scrams, 1997; Shao et al., 2016). Such instances of speededness can induce a decrease in response accuracy, lead to missing responses (Goegebeur et al., 2010), bias the item and ability estimates (Oshima, 1994), and potentially compromise the validity of the test scores by distorting or diminishing the correlation between test scores and other interested variables. This ultimately impacts the interpretation and use of test scores in decision-making processes.

Due to the necessity of time limits, it is almost impossible to entirely eliminate test speededness. Consequently, it is essential to identify response patterns affected by speededness to maintain the reliability and validity of tests. More specifically, it becomes critical to pinpoint the locations and magnitudes of test speededness at the individual level. Over several decades, various approaches have been developed to measuring and detecting test speededness, and they can be broadly classified into three categories. The first category is non-model-based approaches, such as rules of thumb and descriptive statistics, which provide basic guidelines and analytic techniques for identifying and assessing speededness. However, these methods have limitations in capturing individual differences in test speededness. To overcome this issue, the second approach known as model-based methods is proposed (e.g., Lu & Sireci, 2007; Schnipke & Scrams, 1997; Wise & Kong, 2005). One such method is the hybrid modeling framework proposed by Yamamoto (1995). This framework assumes different models before and after the speeded point, with the same speeded location for examinees (i.e., speededness homogeneity). Boughton and Yamamoto (2007) extended this approach to the hybrid Rasch model, allowing for different speeded locations among examinees (i.e., speededness heterogeneity). Another model-based method is the two-class mixture Rasch model proposed by Bolt et al. (2002). This model assumes a single switch point, beyond which examinees engage in rapid guessing behavior. Wang and Xu (2015) proposed a mixture hierarchical model to distinguish solution behavior from rapid guessing behavior based on responses and RTs, which allows to consider multiple switch points for each examinee. However, these models assumed that the probabilities of correct response after the switch points are fixed or suddenly decrease. More flexible models are needed to model the gradual switch among test-taking behaviors. Wollack and Cohen (2004) addressed this by introducing a gradual change model, where the probabilities of correct responses decline gradually after the switch point.

The third category involves detecting aberrant patterns of responses and/or RTs at individual level. Person-fit indices based on responses (e.g., Meijer & Sijtsma, 2001) and/or RTs (Marianti et al., 2014; van der Linden & van Krimpen-Stoop, 2003) are introduced. Key to this approach is that regular behavior should be adequately fitted by a psychometric model, otherwise, it will induce false detection rate due to blurred classifications of normal and aberrant behaviors. Another popular method for detecting test speededness is the cumulative sum (CUSUM) (e.g., Armstrong & Shi, 2009; Egberink et al., 2010; Meijer, 2002; Tendeiro & Meijer, 2012; van Krimpen-Stoop & Meijer, 2001). However, the CUSUM procedures have higher power only when the underlying statistical model before and after the change point are known. Change point analysis (CPA) is another approach used to detect test speededness (Sinharay, 2017). Shao et al. (2016) implemented a CPA based on the likelihood ratio test for detecting test speededness using responses only, thereby pinpointing the locations where examinees begin to speed up. Cheng and Shao (2022) extended this method, focusing solely on RTs, and utilized both likelihood ratio test and Wald test to detect test speedness. Sinharay (2016) introduced three statistics based on CPA using only responses. Subsequently, Yu and Cheng (2020) conducted a comprehensive comparison between these three change-point analysis methods and twelve CUSUM methods based solely on responses. However, the above methods primarily consider either response data or RT data, lacking integration of both types of data. Even though several studies focus on responses and RTs to detect aberrancy, they are either limited to identifying cheating behavior from item preknowledge (Demirkaya, 2022; Demirkaya et al., 2023; Sinharay & Johnson, 2020) or unable to differentiate between

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

specific types of aberrancies (Gorney et al., 2024). Recently, Lu et al. (2024) proposed a real-time two-stage detection procedure to detect rapid guessing and cheating behaviors, in which Stage I identifies the aberrant RTs and then Stage II determines the type of aberrant behavior. This approach is optimal for immediate real-time monitoring and swift change detection in computer-based testing.

In contrast, likelihood ratio and Wald tests are more appropriate for post-hoc analysis on existing datasets. They involve testing a single hypothesis between two models, that is, with and without the specific change point, at a time. Specifically, null hypothesis assumes no change point (regarded as one model), and alternative hypothesis assumes a change point at a specific item location (regarded as the other model). These two methods determine whether to reject the null hypothesis based on the ratio between these two models.

In this article, we aim to provide a more effective CPA method compared to the likelihood ratio and Wald tests. The proposed CPA approaches based on Schwarz information criterion (SIC) (abbreviated as SIC-CPA) rely on the idea of model selection in statistics. SIC is preferred over other information criteria such as Akaike information criterion because it introduces a stronger penalty for model complexity, leading to more parsimonious models. This is particularly beneficial for detecting change points in noisy data, where overly complex models might capture random variations instead of true changes. Additionally, SIC has strong asymptotic properties (Chen et al., 2000), ensuring that as the sample size increases, it is more likely to select the true model, thereby offering a robust and reliable framework in detecting structural changes within data.

The principle of the SIC-CPA method is as follows: it fits models to all possible change point locations, calculates the SIC value for each item location, and then the location with the smallest SIC is selected as the most possible change point. This comprehensiveness of the SIC-CPA method ensures identifying true change points more accurately. Moreover, the SIC-CPA method is computationally straightforward, thereby enhancing the computational efficiency of change point detection. More importantly, for test speededness detection with just one change point, the proposed SIC-CPA approaches take the model complexity into account when making change point decision, penalizing the model based on the number of parameters. In contrast, the likelihood ratio and Wald tests depend solely on the model fit without directly considering its complexity. Consequently, in data with subtle variations, the likelihood ratio and Wald tests might falsely pinpoint change points, leading to higher Type-I errors. In such cases, the proposed SIC-CPA approach tends to favor models without a change point, reducing the Type-I errors but possibly increasing Type-II errors. However, if the change point has a significant effect and substantially improves the model fit, the SIC-CPA approach can efficiently detect it, thereby reducing Type-II errors. Furthermore, given that the SIC-CPA approach introduces penalties for model complexity, it is more resistant to random fluctuations in the data, preventing these fluctuations from being mistaken for actual change points. All three approaches are implemented conditional on the available item parameter values, which are needed to evaluate model fit and to obtain the person parameter estimates used in the test statistics. In simulation studies, item parameters are either fixed at their true values or estimated through calibration to reflect applied testing scenarios.

In the context of aberrant behavior detection, the mixture approach often integrates both responses and RTs to differentiate between normal and aberrant behaviors. Such joint mixture models can be estimated using either Bayesian or frequentist strategies, and the computational burden depends on model specifications, estimation algorithms, and whether item parameters are treated as known. In contrast, the main advantage of the SIC-CPA method lies in its relatively simple detection-oriented and lightweight feature: when item parameters are treated as known, change points can be identified via SIC-based model comparison without needing to estimate the additional latent parameters typically involved in joint mixture models (e.g., class membership and/or person-by-item latent indicators indexing normal versus aberrant behavior). Importantly, both the SIC-CPA approach and joint mixture models typically involve estimating examinees' ability and speed parameters. The key difference is that joint mixture models additionally model latent behavioral structure along with all model parameters simultaneously, which typically requires more complex model specification and estimation and, in some

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

cases, involve high-dimensional data matrices (e.g., test-taking indicators at the person-by-item level). Besides, the mixture model usually requires that both response data and RT data exhibit aberrancies on the same set of items. If aberrant RTs seem normal while response data remains aberrant, the mixture model may fail to accurately identify the aberrant behavior. This is because the assumption of the mixture model, which expects aberrancies to occur in both dimensions at once, is violated in such cases. In contrast, the SIC method is capable of accurately identifying speeded behavior even when there is significant overlap between the aberrant and normal RT distributions, as long as there is a clear difference in the response data (see simulation study III for detailed analysis).

The structure of this article is as follows. First, we introduce three CPA approaches based on SIC, which focus on response data only, RT data only, and the combination of both, respectively. Second, the critical value and significance level of SIC-CPA method are given to determine whether test speededness occurs at the specific item location. Third, we conduct simulation studies to evaluate the effectiveness of the proposed SIC-CPA method in detecting test speededness. Fourth, a real data set is employed to show the application of the proposed SIC-CPA method. Finally, we conclude by discussing the strengths and limitations of this article and outline potential future research directions.

2. Method

2.1. Schwarz information criterion to detect change points

Information criteria have been extensively utilized in model selection for change point problems (Chen & Gupta, 1997; Yao, 1988). SIC is a well-established method for detecting change points, and Chen and Gupta (1997) showed that it is a highly effective method for detecting change points and is marginally superior to likelihood ratio procedures and Bayesian methods.

SIC is a statistical method for model selection that was first proposed by Schwarz (1978). It is more commonly known as the Bayesian information criterion (BIC) and reflects a trade-off between model fit and complexity, which is typically determined by the number of parameters.
Specifically, SIC[1] is mathematically equivalent to BIC and is calculated as

$$\mathrm{SIC} = -2\log L\left(\widehat{\theta}\right) + p \times \log(n), \tag{1}$$

where $L\left(\widehat{\theta}\right)$ is the maximum likelihood, $p \times \log(n)$ is called as the penalty parameter, p is the number of parameters, and n is the sample size. In change point detection, SIC is used to identify the optimal model that fits the data with the fewest number of change points.

2.2. CPA based on SIC for item responses

Throughout Section 2.2, we treat item parameters as given; this is a prerequisite for computing the likelihood and the SIC values, and the setting in which item parameters are unknown is studied in simulation study V. Given a test length of J items to be answered by the ith examinee, it is assumed that the ith examinee will answer the jth item with latent ability $\theta_{i,j}$, where $j = 1, 2, \ldots, J$. Therefore, the following hypothesis testing is considered. The null hypothesis is

$$H_0^{\theta} : \theta_{i,1} = \theta_{i,2} = \cdots = \theta_{i,J} \triangleq \theta_i, \tag{2}$$

against the alternative hypothesis

$$H_1^{\theta} : \theta_{i,1} = \cdots = \theta_{i,k} \triangleq \theta_i \neq \theta_{i,k+1} = \cdots = \theta_{i,J} \triangleq \theta_i'. \tag{3}$$

This alternative hypothesis assumes that change point occurs between item k and $k+1$.

[1] Although the term *Bayesian Information Criterion (BIC)* is now more common, we retain the notation "SIC" (*Schwarz Information Criterion*) to maintain consistency with earlier change point detection studies (e.g., Chen et al., 2000; Yao, 1988).

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

As an illustration, the traditional two parameter logistic model (2PLM; Birnbaum, 1968) is used as the item response theory (IRT) model, that is,

$$P_j(\theta_i) = p(y_{ij} = 1 \mid \theta_i, a_j, b_j) = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}. \tag{4}$$

In Eq. (4), y_{ij} represents the response of the ith examinee answering the jth item, where $i = 1, \ldots, N$ and $j = 1, \ldots J$. The correct response probability is expressed as $P_j(\theta_i)$, and the corresponding incorrect response probability is $Q_j(\theta_i) = 1 - P_j(\theta_i)$ .a_j is the discrimination parameter of the jth item, and b_j is the difficulty parameter of the jth item. θ_i denotes the latent ability of the ith examinee.

Therefore, under H_0^{θ}, the maximum likelihood function based on the 2PL model is

$$L_{i0}(\boldsymbol{y}_i \mid \hat{\theta}_i) = \prod_{j=1}^{J} \left\{ \frac{\exp[a_j(\widehat{\theta}_i - b_j)]}{1 + \exp[a_j(\hat{\theta}_i - b_j)]} \right\}^{y_{ij}} \left\{ \frac{1}{1 + \exp[a_j(\widehat{\theta}_i - b_j)]} \right\}^{1-y_{ij}}, \tag{5}$$

where $\boldsymbol{y}_i$ denotes the vector of item responses for the ith examinee, that is, $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})'$. $\widehat{\theta}_i$ is the maximum likelihood estimate (MLE) of the latent ability obtained from the responses based on all items of the test. Therefore, the SIC (Chen & Gupta, 1997; Schwarz, 1978) under H_0^{θ}, that is, $\mathrm{SIC}_{\theta}(J)$, can be expressed as

$$\begin{aligned} \mathrm{SIC}_{\theta_i}(J) &= -2\left[\log L_{i0}(\boldsymbol{y}_i \mid \widehat{\theta}_i)\right] + \log(J) \\ &= -2\sum_{j=1}^{J}\left[y_{ij}\log P_j(\widehat{\theta}_i) + (1 - y_{ij})\log Q_j(\widehat{\theta}_i)\right] + \log(J), \end{aligned} \tag{6}$$

where $P_j(\hat{\theta})_i)$ and $Q_j(\widehat{\theta}_i)$ are the values of $P_j(\theta)$ and $Q_j(\theta)$ by plugging in $\widehat{\theta}_i$, respectively.
Here, the item parameters are available, and only one latent ability parameter is needed to be estimated for the examinee i under the null hypothesis. Therefore, the number of parameters included in the penalty term is 1 (i.e., $p = 1$), and the number of observations included in the penalty term is J (i.e.,$n = J$). Note that when item parameters are unknown, the penalty term should include both item and person parameters. As a result, the effective penalty becomes substantially larger, which may make the criterion more conservative in selecting a change point model.

Under H_1^{θ}, the corresponding maximum likelihood function can be obtained as

$$\begin{aligned} L_{i1}(\boldsymbol{y}_i \mid \hat{\theta}_i, \widehat{\theta}_i') = {} & \prod_{j=1}^{k} \left\{ \frac{\exp[a_j(\hat{\theta}_i - b_j)]}{1 + \exp[a_j(\hat{\theta}_i - b_j)]} \right\}^{y_{ij}} \left\{ \frac{1}{1 + \exp[a_j(\hat{\theta}_i - b_j)]} \right\}^{1-y_{ij}} \\ & \times \prod_{j=k+1}^{J} \left\{ \frac{\exp[a_j(\hat{\theta}_i' - b_j)]}{1 + \exp[a_j(\hat{\theta}_i' - b_j)]} \right\}^{y_{ij}} \left\{ \frac{1}{1 + \exp[a_j(\hat{\theta}_i' - b_j)]} \right\}^{1-y_{ij}}, \end{aligned} \tag{7}$$

where $\widehat{\theta}_i$ and $\widehat{\theta}_i'$ are the MLEs of the latent ability obtained from the responses based on the k and $J - k$ items in the test, respectively. Therefore, the SIC under H_1^{θ}, that is,$\mathrm{SIC}_{\theta}(k)$, for $k = 2, \ldots, J - 1$, can be expressed as

$$\begin{aligned} \mathrm{SIC}_{\theta_i}(k) &= -2\left[\log L_{i1}(\boldsymbol{y}_i \mid \widehat{\theta}_i, \widehat{\theta}_i')\right] + 2\log(J) \\ &= -2\sum_{j=1}^{k}\left[y_{ij}\log P_j(\widehat{\theta}_i) + (1 - y_{ij})\log Q_j(\widehat{\theta}_i)\right] \\ &\quad -2\sum_{j=k+1}^{J}\left[y_{ij}\log P_j(\widehat{\theta}_i') + (1 - y_{ij})\log Q_j(\widehat{\theta}_i')\right] + 2\log(J), \end{aligned} \tag{8}$$

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

where the number of parameters included in the penalty term is 2 (i.e., $p = 2$). According to the minimum information criterion principle, H_0 is rejected if

$$\mathrm{SIC}_{\theta_i}(J) > \min_{2\le k\le J-1} \mathrm{SIC}_{\theta_i}(k), \tag{9}$$

and the estimated change point position denoted by $\widehat{k}$ such that

$$\mathrm{SIC}_{\theta_i}\left(\widehat{k}\right) = \min_{2\le k\le J-1} \mathrm{SIC}_{\theta_i}(k). \tag{10}$$

Remark 1. Although the SIC method can be applied to detect change points at any position in theory, it is suggested to be used when the ability parameters are accurately estimated. In other words, there needs to be a sufficient number of items available both before and after the change point.

Remark 2. The SIC method can be extended to identify multiple change points by performing a series of hypothesis testing for a given response sequence. The process involves the following steps:

Step 1: The null hypothesis, as defined by Eq. (2), is tested against the alternative hypothesis given by Eq. (3) to determine if there is a change point. If H_0^{θ} is not rejected, then no change point is present, and the process stops. If H_0^{θ} is rejected, then the estimated position of the change point, denoted by $\widehat{k}$, is determined by Eq. (10), where $\widehat{k}$ is the location of the single change point at this stage. The process then moves to Step 2.

Step 2: The SIC method is used to test the two subsequences before and after the change point detected by Step 1 separately for a change.

Step 3: Repeat the process until no further subsequences have change points.

Step 4: The collection of change point locations found by Steps 1–3 is denoted by $\left(\widehat{k}_1, \widehat{k}_2, \ldots, \widehat{k}_M\right)$, and the estimated total number of change points is M. Note that the calculation of SIC-CPA incorporates a penalty for model complexity, effectively preventing the selection of overly complex models with multiple change points.

Although SIC-CPA method can detect multiple change points, this study mainly focuses on detecting test speededness, and in common scenarios, examinees exhibit a single change point at the end of the test due to factors such as test fatigue, motivation less, test disengagement behavior, etc. Further investigation of the SIC-CPA method for detecting two change points can be found in simulation study IV.

2.3. CPA based on SIC for item response times

Similarly, considering a test with J items, examinee i answers item j with a speed parameter of $\tau_{i,j}$, where $j = 1, 2, \ldots, J$. The hypothesis testing problem is considered as follows: The null hypothesis is

$$H_0^{\tau} : \tau_{i,1} = \tau_{i,2} = \cdots = \tau_{i,J} \triangleq \tau_i, \tag{11}$$

and the alternative hypothesis is

$$H_1^{\tau} : \tau_{i,1} = \cdots \tau_{i,m} \triangleq \tau_i \neq \tau_{i,m+1} = \cdots = \tau_{i,J} \triangleq \tau_i'. \tag{12}$$

In this study, the traditional log-normal model (van der Linden, 2006) is used to fit the RT data, that is,

$$\log(t_{ij}) \mid \tau_i \sim N\left(\beta_j - \tau_i, \alpha_j^{-2}\right), \tag{13}$$

where α_j and β_j are time dispersion and time intensity parameters for item j, respectively. τ_i is the speed parameter of examinee i. Therefore, under H_0^{τ}, the maximum likelihood function of observing a RT pattern $\boldsymbol{t}_i$ for the ith examinee can be obtained as

$$L_{i0}\left(\boldsymbol{t}_i \mid \widehat{\tau}_i\right) = \prod_{j=1}^{J} \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{\alpha_j^2\left[\log(t_{ij}) - (\beta_j - \widehat{\tau}_i)\right]^2}{2}\right\}, \tag{14}$$

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

where $\boldsymbol{t}_i = (t_{i1}, t_{i2}, \ldots, t_{iJ})'$ is a RT pattern that captures the RTs of examinee i answering all items of the test. Therefore, the SIC under H_0^τ, that is, $\mathrm{SIC}_\tau(J)$, is

$$\begin{aligned}\mathrm{SIC}_{\tau_i}(J) &= -2\left[\log L_{i0}\left(\boldsymbol{t}_i \mid \widehat{\tau}_i\right)\right] + \log(J) \\ &= -2\left\{\sum_{j=1}^{J} \log\left(\frac{\alpha_j}{t_{ij}\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{j=1}^{J} \alpha_j^2\left[\log t_{ij} - (\beta_j - \widehat{\tau}_i)\right]^2\right\} + \log(J). \end{aligned} \tag{15}$$

Here, we assume that the item parameters are known, and only one latent speed parameter is needed to be estimated for examinee i under the null hypothesis. Therefore, the number of parameters included in the penalty term is 1 (i.e., $p = 1$), and the number of observations is J (i.e., J is number of all RT observations for examinee i).

Under H_1^τ, the corresponding maximum likelihood function is

$$\begin{aligned} L_{i1}\left(\boldsymbol{y}_i \mid \widehat{\tau}_i, \hat{\tau}_i'\right) = &\prod_{j=1}^{m}\left[\frac{\alpha_j}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{\alpha_j^2\left[\log t_{ij} - (\beta_j - \widehat{\tau}_i)\right]^2}{2}\right\}\right] \\ &\times \prod_{j=m+1}^{J}\left[\frac{\alpha_j}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{\alpha_j^2\left[\log t_{ij} - \left(\beta_j - \hat{\tau}_i'\right)\right]^2}{2}\right\}\right], \end{aligned} \tag{16}$$

where $\widehat{\tau}_i$ and $\hat{\tau}_i'$ are the MLEs of the speed parameter obtained by calculating the m and $J - m$ RTs, respectively. Therefore, the SIC under H_1^τ, $\mathrm{SIC}_\tau(m)$, for $m = 2, \ldots, J-1$, is obtained as

$$\begin{aligned}\mathrm{SIC}_{\tau_i}(m) &= -2\left[\log L_{i1}\left(\boldsymbol{t}_i \mid \widehat{\tau}_i, \hat{\tau}_i'\right)\right] + 2\log(J) \\ &= -2\left\{\sum_{j=1}^{m} \log\left(\frac{\alpha_j}{t_{ij}\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{j=1}^{m} \alpha_j^2\left[\log t_{ij} - (\beta_j - \widehat{\tau}_i)\right]^2\right\} \\ &\quad -2\left\{\sum_{j=m+1}^{J} \log\left(\frac{\alpha_j}{t_{ij}\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{j=m+1}^{J} \alpha_j^2\left[\log t_{ij} - \left(\beta_j - \hat{\tau}_i'\right)\right]^2\right\} + 2\log(J), \end{aligned} \tag{17}$$

where the number of parameters included in the penalty term is 2 (i.e., $p = 2$). Based on the minimum information criterion principle, we reject H_0^τ if

$$\mathrm{SIC}_{\tau_i}(J) > \min_{2 \le m \le J-1} \mathrm{SIC}_{\tau_i}(m), \tag{18}$$

and the estimated change point position denoted by $\widehat{m}$ such that

$$\mathrm{SIC}_{\tau_i}\left(\widehat{m}\right) = \min_{2 \le m \le J-1} \mathrm{SIC}_{\tau_i}(m). \tag{19}$$

2.4. CPA based on SIC for item responses and response times

A modified SIC method is proposed to detect test speededness by combining both response and RT data. The key to the hypothesis testing is to examine whether the ability and speed parameters change simultaneously during the responding process. We consider the following hypothesis testing problem: The null hypothesis is

$$H_0 : \theta_{i,1} = \theta_{i,2} = \cdots = \theta_{i,J} \triangleq \theta_i, \tau_{i,1} = \tau_{i,2} = \cdots = \tau_{i,J} \triangleq \tau_i, \tag{20}$$

and the alternative hypothesis is

$$\begin{aligned} H_1 : &\theta_{i,1} = \cdots = \theta_{i,r} \triangleq \theta_i \neq \theta_{i,r+1} = \cdots = \theta_{i,J} \triangleq \theta_i', \\ &\tau_{i,1} = \cdots = \tau_{i,r} \triangleq \tau_i \neq \tau_{i,r+1} = \cdots = \tau_{i,J} \triangleq \tau_i'. \end{aligned} \tag{21}$$

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

Here, van der Linden's (2006) joint model for responses and RTs is considered. Therefore, under H_0, the maximum likelihood function of observing response and RT pattern $\left(\boldsymbol{y_i}, \boldsymbol{t_i}\right)$ for examinee i is

$$L_{i0}\left(\boldsymbol{y_i}, \boldsymbol{t_i} \mid \hat{\theta}_i, \hat{\tau}_i\right) = \prod_{j=1}^{J} \left\{\frac{\exp\left[a_j\left(\hat{\theta}_i - b_j\right)\right]}{1+\exp\left[a_j\left(\hat{\theta}_i - b_j\right)\right]}\right\}^{y_{ij}} \left\{\frac{1}{1+\exp\left[a_j\left(\hat{\theta}_i - b_j\right)\right]}\right\}^{1-y_{ij}} \times \prod_{j=1}^{J} \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{\alpha_j^2\left[\log t_{ij} - \left(\beta_j - \hat{\tau}_i\right)\right]^2}{2}\right\}, \tag{22}$$

where $\boldsymbol{y_i} = \left(y_{i1}, y_{i2}, \ldots, y_{ij}, \ldots y_{iJ}\right)$ is the response pattern of examinee i and $\boldsymbol{t_i} = \left(t_{i1}, t_{i2}, \ldots, t_{iJ}\right)'$ is the RT pattern of examinee i. Therefore, the SIC under H_0, that is, $\mathrm{SIC}_{\theta,\tau}(J)$, is

$$\mathrm{SIC}_{\theta_i,\tau_i}(J) = -2\left[\log L_{i0}\left(\boldsymbol{y_i}, \boldsymbol{t}_i \mid \hat{\theta}_i, \widehat{\tau}_i\right)\right] + 2\log(J). \tag{23}$$

Under H_1, the corresponding maximum likelihood function is

$$\begin{aligned} L_{i1}\left(\boldsymbol{y_i}, \boldsymbol{t_i} \mid \hat{\theta}_i, \hat{\theta}_i', \hat{\tau}_i, \hat{\tau}_i'\right) = & \prod_{j=1}^{r} \left\{\frac{\exp\left[a_j\left(\hat{\theta}_i - b_j\right)\right]}{1+\exp\left[a_j\left(\hat{\theta}_i - b_j\right)\right]}\right\}^{y_{ij}} \left\{\frac{1}{1+\exp\left[a_j\left(\hat{\theta}_i - b_j\right)\right]}\right\}^{1-y_{ij}} \\ & \times \prod_{j=r+1}^{J} \left\{\frac{\exp\left[a_j\left(\hat{\theta}_i' - b_j\right)\right]}{1+\exp\left[a_j\left(\hat{\theta}_i' - b_j\right)\right]}\right\}^{y_{ij}} \left\{\frac{1}{1+\exp\left[a_j\left(\hat{\theta}_i' - b_j\right)\right]}\right\}^{1-y_{ij}} \\ & \times \prod_{j=1}^{r} \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{\alpha_j^2\left[\log t_{ij} - \left(\beta_j - \hat{\tau}_i\right)\right]^2}{2}\right\} \\ & \times \prod_{j=r+1}^{J} \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{\alpha_j^2\left[\log t_{ij} - \left(\beta_j - \hat{\tau}_i'\right)\right]^2}{2}\right\}, \end{aligned} \tag{24}$$

where $\widehat{\theta}_i$ and $\widehat{\tau}_i$ are the MLEs of the ability and speed parameters obtained from the first r responses and RTs, respectively, $\hat{\theta}_i'$ and $\hat{\tau}_i'$ are the corresponding MLEs obtained from the remaining $J-r$ responses and RTs, respectively. Therefore, the SIC under H_1, $\mathrm{SIC}_{\theta,\tau}(r)$, for $r = 2, \ldots, J-1$, is obtained as

$$\mathrm{SIC}_{\theta_i,\tau_i}(r) = -2\left[\log L_{i1}\left(\boldsymbol{y_i}, \boldsymbol{t_i} \mid \hat{\theta}_i, \hat{\theta}_i', \hat{\tau}_i, \hat{\tau}_i'\right)\right] + 4\log(J), \tag{25}$$

where the number of parameters included in the penalty term is 4 (i.e., $p = 4$). Based on the minimum information criterion principle, we reject H_0 if

$$\mathrm{SIC}_{\theta_i,\tau_i}(J) > \min_{2\le r\le J-1} \mathrm{SIC}_{\theta_i,\tau_i}(r), \tag{26}$$

and the estimated change point position denoted by $\widehat{r}$ such that

$$\mathrm{SIC}_{\theta_i,\tau_i}(\widehat{r}) = \min_{2\le r\le J-1} \mathrm{SIC}_{\theta_i,\tau_i}(r). \tag{27}$$

2.5. Critical value and significance level of CPA based SIC method

According to Gupta and Chen (1996), when the difference between SICs is small, it may be difficult to determine whether the change point actually exists or is simply caused by data fluctuation. In our study, the fluctuation of responses and/or RTs may induce subtle differences in SICs, which may result in examinees without change points being incorrectly identified as speeded (i.e., exhibiting changes in ability and/or speed). For example, as verified based solely on responses in our simulation study I, among the examinees without change points, 75% of the $\mathrm{SIC}_{\theta}(J) - \min_{2\le k\le J-1} \mathrm{SIC}_{\theta}(k)$ is less than critical value, which results in an extremely high Type-I error. To solve this problem, Gupta and Chen (1996)

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

introduced the significance level α and its associated critical value c_α, where $c_\alpha \geq 0$. Therefore, we accept H_0 if

$$\mathrm{SIC}(J) < \min_{2\leq k\leq J-1} \mathrm{SIC}_\theta(k) + c_\alpha, \tag{28}$$

where

$$\alpha = P\left[\mathrm{SIC}(J) > \min_{2\leq k\leq J-1} SIC_\theta(k) + c_\alpha \mid H_0\right]. \tag{29}$$

Unfortunately, as the (joint) distribution of our models (i.e., 2PL model, log-normal RT model, or van der Linden's (2006) joint response and RT model) is not normal, it is not feasible to obtain a closed form of the exact critical value c_α. Usually, permutation (Shao et al., 2016), bootstrap, and simulation (Armstrong & Shi, 2009; Cheng & Shao, 2022) methods are commonly used to determine the critical values in detecting aberrancy. However, permutation and bootstrap methods are computationally expensive. Therefore, in this study, empirical critical values were used via Monte Carlo simulations following the procedure outlined by Worsley (1979) and further adopted in Cheng and Shao (2022). Specifically, we generated 10,000 data under the null hypothesis—each dataset representing individual responses, RTs, or a combination of both, depending on the data type used in the SIC-CPA method. For each dataset, we computed the SIC values using estimated ability and/or speed parameters, while treating the item parameters as known. As a result, these 10,000 datasets consist solely of regular data, without any aberrancies. Note that the resulting critical values are determined by the simulation settings, such as test length, item parameters, and the distribution of person parameters, rather than by the proportion of speeded examinees. For the significance levels of 0.05, 0.01, and 0.001 in a one-sided test, the 500th, 100th, and 10th largest values were selected as critical values, respectively. For each test length condition, 1000 replicated processes were performed, and the average critical value of 1000 times was used as the final critical value for each SIC-CPA approach.

3. Simulation studies

Six simulation studies were carried out to investigate the performance of the proposed SIC-CPA approaches, which employed distinct data, including responses, RTs, and a combination of response and RT data, respectively. Specifically, simulation study I compared the proposed SIC-CPA method with Shao et al.'s (2016) method, based solely on response data. In simulation study II, we compared the proposed SIC-CPA method and Cheng and Shao's (2022) method, based solely on RT data. Simulation study III compared the performance of the proposed three SIC-CPA methods; in this case, the data were generated from the modified gradual change joint model for responses and RTs. To manipulate test speededness, examinees with one change point were considered in simulation studies I, II, and III. Simulation study IV was conducted to evaluate the performance of the SIC-CPA method based on RTs with two change points, simultaneously considering the warm-up effect and test speededness. Simulation study V includes an item parameter estimation stage and considers two cases: (1) non-iterative calibration-detection procedure and (2) iterative detect–clean–recalibrate procedure. Under each case, we compare the SIC-CPA with the likelihood ratio and Wald tests. In addition, simulation study VI was conducted to evaluate the performance of the SIC-CPA, likelihood ratio test, and Wald methods when there are no speeded responses and/or response times (RTs) in the data. Due to page limit, simulation study VI can be found in the Supplementary Material. Each simulation condition was conducted 100 replications.

3.1. Data generation

Concerning the generation models for regular and speeded responses and/or RTs, as claimed in Gorney and Wollack's (2022) paper, "All response accuracy (RA) models are formulated as mixture extensions

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

of the IRT model. All RT models are formulated as mixture extensions of the lognormal RT model developed by van der Linden (2006)." The unique difference between these generation models (for example, van der Linden & van Krimpen-Stoop, 2003; Wang & Xu, 2015; Wollack & Cohen, 2004) is the degree of change in the responses/RTs or ability/speed parameter, and thus they can be classified as abrupt and gradual change models. We adopted the gradual change model proposed by Suh et al., (2012) for responses, along with its extensions for RTs developed by Cheng and Shao (2022), as well as our modified gradual change joint model for responses and RTs to generate simulation data. These models were chosen because they offer a more realistic representation and can capture the complexity of real data. Furthermore, employing more sophisticated models for data generation allows for a more comprehensive evaluation of the robustness and flexibility of the proposed approaches (Cheng & Shao, 2022; Luecht & Ackerman, 2018).

Next, we introduce a modified gradual change joint model for responses and RTs:

$$P_j^*(\theta_i) = p(y_{ij} = 1 \mid \theta_i, a_j, b_j) \times \min\left(1, \left[1 - \left(\frac{j}{J} - \eta_i\right)\right]\right)^{\lambda_i}, \tag{30}$$

$$\log(t_{ij}) = (\beta_j - \tau_i + \varepsilon_{ij}) \times \min\left(1, \left[1 - \left(\frac{j}{J} - \eta_i\right)\right]\right)^{\lambda_i}, \varepsilon_{ij} \sim N\left(0, \alpha_j^{-2}\right), \tag{31}$$

where

$$\begin{pmatrix} \theta_i \\ \tau_i \end{pmatrix} \sim \text{MVN}\left(\boldsymbol{\mu_p} = \begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix}, \boldsymbol{\Sigma_p} = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}\right), \tag{32}$$

$$\begin{pmatrix} b_j \\ \beta_j \end{pmatrix} \sim \text{MVN}\left(\boldsymbol{\mu_I} = \begin{pmatrix} \mu_b \\ \mu_\beta \end{pmatrix}, \boldsymbol{\Sigma_I} = \begin{pmatrix} \sigma_b^2 & \sigma_{b\beta} \\ \sigma_{b\beta} & \sigma_\beta^2 \end{pmatrix}\right), \tag{33}$$

and $p(y_{ij} = 1 \mid \theta_i, a_j, b_j)$ follows the 2PL model in Eq. (4), η_i (where $0 < \eta_i < 1$) is the stage of the test that test speededness occurs on examinee i, and $\eta_i \times J$ is the location of change point. For instance, if $\eta_i = 0.7$, it indicates that test speededness occurs at 70% of a test, and if there are 50 items in the test (i.e., $J = 50$), the change point would be item 35. The parameter λ_i (where $\lambda_i \geq 0$) is the speededness rate which controls how fast the correct response probability decreases as the test proceeds beyond η_i, at the same time, λ_i determines the rate at which less time is spent on items after η_i. When λ_i increases, it leads to faster declines of $P_j^*(\theta_i)$ and $\log(t_{ij})$ due to test speededness. For example, take two examinees with different λ and the same η as an example, when $\frac{j}{J} > \eta$ we have $\min\left(1, \left[1 - \left(\frac{j}{J} - \eta\right)\right]\right)^{\lambda_i} = \left[1 - \left(\frac{j}{J} - \eta\right)\right]^{\lambda_i}$, and if $\lambda_1 > \lambda_2$, then $\left[1 - \left(\frac{j}{J} - \eta_i\right)\right]^{\lambda_1} < \left[1 - \left(\frac{j}{J} - \eta_i\right)\right]^{\lambda_2}$, which means that $P_j^*(\theta_1) < P_j^*(\theta_2)$ and $\log(t_{1j}) < \log(t_{2j})$. This modified gradual change joint model is reduced to van der Linden's (2006) hierarchical response and RT model when $\eta_i = 0$ and $\lambda_i = 0$, which means the test speededness does not happen, and hence the responses and RTs are from the regular behavior. Note that Suh et al.'s (2012) gradual change model for responses and the extended gradual model for RTs used in Cheng and Shao (2022) are both special cases of this modified gradual change model for responses and RTs.

Table 1 displays the manipulated factors and their varied levels used in all four simulation studies. Here, 500 and 1,000 examinees were considered, and the proportions of speeded examinees were 10% and 30%, respectively. Three test lengths, 20 (short test), 50 (moderate test) and 80 (long test), were considered, as used in Zhu et al. (2023). Following the data generation scheme in Shao et al. (2016) and Cheng and Shao (2022), η was simulated from Beta distribution, and we set the median of η to 0.5, 0.6, 0.7, 0.8 and 0.9, and the variance of η was set to 0.001, 0.01, and 0.04. Because the detection results are consistent across sample sizes, we only show the simulation results for 500 examinees. In general, 90 (i.e., 3 (test lengths)×2 (speededness proportions)×5 (median of η)×3 (variance of η)) baseline simulation conditions were conducted. We implemented the SIC-CPA, Shao et al. (2016), and Cheng and Shao (2022) approaches using the R programming language (R Core Team, 2022) on a computer equipped with Intel(R) Core(TM) i7-7700HQ CPU at 2.80 GHz, 8 GB RAM. Supplementary Table S1 shows the

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

Table 1. Fixed factors/parameters in four simulation studies.

Factor	Values
Number of examinees	500
Test length	20, 50, 80
Proportion of examinees with test speededness	10%, 30%
η	Median (0.5, 0.6, 0.7, 0.8, 0.9)× Variance (0.001, 0.01, 0.04)

average values of critical values and the standard deviations (SD) for the SIC-CPA, likelihood ratio test, and Wald statistics across 1,000 replicated processes in simulation studies I, II, and III at the significant level of 0.01.

3.2. Evaluation criteria

We evaluate the performance of the proposed approaches from three aspects. First, concerning the detection of speeded examinees, we focus on the correct classification rate (CCR), power and Type I errors. CCR is the ratio of correctly classified examinees (including normal and speeded examinees) to the total number of examinees, power is the proportion of correctly identified speeded examinees among all identified speeded examinees, and Type I error rate is the proportion of normal examinees incorrectly classified as speeded examinees. Second, we assess the performance of the proposed method in pinpointing change points. We use "lag," which is equal to *estimated location of change point* minus *true location of change point*, to show the recoveries of estimating change points. Because "positive lags" and "negative lags" offset each other, the "average absolute lag" (AL_mean) and its standard deviation (AL_SD) are used to compare the performance of different methods in locating change points. Third, the computational time is given to show the efficiency of the different approaches.

3.3. Simulation study I

This simulation study evaluates the performance of the proposed SIC-CPA approach based solely on the responses. Meanwhile, we compare the effectiveness of the SIC-CPA approach with Shao et al.'s (2016) CPA method, hereafter referred to as Shao-CPA. The speeded responses were generated from Suh et al.'s (2012) gradual change model, that is, Eq. (31), and the non-speeded responses were generated from the 2PL model in Eq. (4). For the model parameters, we simulate $a_j \sim \log N(0, 0.5)$, $b_j \sim N(0,1)$, $\theta_i \sim N(0,1)$, and $\lambda_i \sim \log N(3.912, 1)$. We sampled λ_i values from $\log N(3.912, 1)$ following the previous studies (Goegebeur et al., 2008; Suh et al, 2012), which generated various test speededness patterns. Specifically, this choice represents a reasonable range of performance that accounts for both typical and extreme test speededness, ensuring that the model can simulate the full spectrum of examinee behaviors: the majority of examinees experience a gradual decline in performance, whereas a minority exhibits more extreme speededness near the end of the test. Note that, in Shao et al.'s (2016) simulation study, a permutation method was used to generate the null distribution for the Shao-CPA test statistics, with the false discovery rate set to 0.2. To ensure a fair comparison between their approach and our SIC-CPA approach, we used Monte Carlo simulations to determine the critical values for these two methods. Due to consistent trends in the SIC-CPA method with different critical values and page limit, we present the significance level of 0.01 for these two approaches.

Supplementary Tables S2–S4 present the detection results of Shao-CPA and SIC-CPA methods with test lengths of 20, 50, and 80, respectively. Both methods maintain the type I error rates around 0.01 across all conditions. For the test with 20 items (Supplementary Table S2), the performance of both methods in terms of CCR and power is similar. However, SIC-CPA performs better under high speeded

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

Table 2. Average running time of Shao-CPA and SIC-CPA approaches when sample size is 500 in simulation study I.

Test length	Shao-CPA (seconds)	SIC-CPA (seconds)
20	2.5	0.9
50	5.6	2.8
80	8.9	5.5

examinee proportions (30%) and larger values of the median and variance of η. Specifically, SIC-CPA achieves lower AL_mean and AL_SD, which indicates more accurate change point estimation. When the median of η is 0.8, the power of both methods drops below 0.43 when the proportion of speeded examinees is 30%. This decline can be attributed to the fact that the change point locations for speeded examinees approached the end of the test, resulting in fewer speeded responses, which in turn diminishes the performance of identifying speeded examinees and pinpointing the change points.

For the test with 50 (or 80) items in Supplementary Table S3 (or S4), as the median of η increases, both CCRs and powers of these two methods decrease in most cases. When the variance of η increases, indicating greater variations in change point locations, both AL_mean and AL_SD increase, indicating a decrease in change point estimation accuracy. As the proportion of speeded examinees increases, the performance of these two methods declines, and specifically, CCRs and powers decrease. Meanwhile, it becomes more difficult to accurately identify change point locations, leading to higher AL_mean and AL_SD values.

As the test length increases, Shao-CPA shows an increase in both AL_mean and AL_SD. This is because Shao-CPA struggles to accurately locate change points with longer tests, which is consistent with the findings of Shao et al. (2016). However, as the test length grows, the performance of detecting speeded examinees improves, resulting in consistently higher CCRs and power for both two methods. In particular, when the median of η is 0.8, the power of these two approaches remains above 0.56 under the condition of 30% speeded examinees, as shown in Supplementary Table S4. Furthermore, the advantages of SIC-CPA become more evident, especially under high values of median and variance of η. This suggests that SIC-CPA is more robust in handling data variability and it provides more accurate and stable estimates of change points. Therefore, SIC-CPA proves to be a more reliable and efficient approach, particularly in scenarios involving higher speeded examinee proportion and greater variability in the data.

Table 2 shows the average running time of Shao-CPA and SIC-CPA when the sample size is 500. It is evident that the SIC-CPA method operates quickly, indicating a higher computational efficiency. Note that under each simulation condition, the average running time reported in this article indicates the average time for a single replication, and the running time does not include the Monte Carlo simulations used to obtain the empirical critical value.

3.4. Simulation study II

Simulation study II was conducted to assess the performance of the proposed SIC-CPA approach utilizing only RTs, and to compare its performance with Cheng and Shao's (2022) CPA method. The speeded RTs were generated from the extended gradual change model for RTs used in Cheng and Shao (2022), that is, Eq. (32). For the model parameters, $\alpha_j \sim U(1.75, 3.25)$, $\beta_j \sim U(3.6, 0.11)$, $\tau_i \sim N(0, 0.25)$, and $\lambda_i \sim \log N(3.912, 1)$. In addition, the time limits for the test lengths of 20, 50, and 80 were 30, 75, and 120 min, respectively. 90 simulation conditions were conducted, please see Table 1.

Due to the consistency of results obtained from both the Wald test and likelihood ratio test, aligning with the findings from Cheng and Shao (2022), this simulation study only presents the results of the Wald test. Supplementary Tables S5–S7 display the detection results of SIC-CPA method and the Wald test under the significance level of 0.01. The last column in these two tables indicates the proportion of

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

Table 3. Average running time of SIC-CPA approach, Wald test, and likelihood ratio test when sample size is 500 in simulation study II.

Test length	SIC-CPA (seconds)	Wald test (seconds)	Likelihood ratio test (seconds)
20	0.2	0.3	0.6
50	0.5	0.4	1.1
80	0.7	0.5	1.9

examinees who did not complete the test within the time limit, ranging from 4% to 8%. This suggests that the majority of speeded examinees can complete the test on time. Consistent with Cheng and Shao (2022), this article labels examinees who did not complete the test as speeded without performing any statistical analysis.

Overall, the results of SIC-CPA and the Wald test are nearly identical, with SIC-CPA having slightly lower Type I errors. Both methods achieve high CCRs exceeding 0.9 except under the condition with 30% speeded examinee proportion and $\eta_{\text{median}} = 0.9$, $\eta_{\text{var}} = 0.04$. However, when the proportion of speeded examinees is 30%, detection becomes more challenging because the change points are located closer to the end of the test. Under the condition of 30% speeded examinee proportion with $\eta_{\text{median}} = 0.9$, $\eta_{\text{var}} = 0.04$, the power of both methods decrease substantially.

Regarding the accuracy of change point estimation, when the median of η is 0.9, both AL_mean and AL_SD are relatively small, indicating precise change point estimation. As η variance increases, the power decreases, but the accuracy of change point estimation remains high. Furthermore, as the test length increases, the performance of recovering change points for both methods deteriorate, which is consistent with findings of simulation study I. Under the condition of 30% speeded examinee proportion with $\eta_{\text{median}} = 0.9$ and $\eta_{\text{var}} = 0.04$, SIC-CPA outperforms Shao-CPA in change point estimation accuracy, showing lower values of AL_mean and AL_SD.

Table 3 shows the average running time for SIC-CPA method, Wald test, and likelihood ratio test under the test lengths of 20, 50, and 80 when sample size is 500. It is observed that both the SIC-CPA method and Wald test exhibit similar running times. In addition, SIC-CPA is more efficient than the likelihood ratio test.

3.5. Simulation study III

This simulation study aims to evaluate the performance of the proposed SIC-CPA approach using responses and RTs. In addition, we compared the results with those from the other two SIC-CPA approaches that used either responses or RTs solely. The speeded responses and RTs were simulated using the modified gradual change joint model for responses and RTs (please see Eqs. (30)–(33)). Van der Linden's (2006) hierarchical response and RT model (i.e., Eqs. (30)–(33) with $\eta_i = 0$ and $\lambda_i = 0$) was used to generate the non-speeded responses and RTs. For speeded examinee i, to ensure that the speeded response and RT data are aligned at the same change point, the same values of η_i and λ_i in Eqs. (30) and (31) were used for the modified gradual change joint model for responses and RTs. The person parameters were sample from $\begin{pmatrix} \theta_i \\ \tau_i \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_p = \begin{pmatrix} 1 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & 0.25 \end{pmatrix}\right)$. Here, we manipulated two levels of correlation coefficients between ability and speed parameters, that is, 0.5 and 0.8, which corresponds to $\sigma_{\theta\tau} = 0.25$ and 0.4 respectively. The item parameters were simulated as follows: $\begin{pmatrix} b_j \\ \beta_j \end{pmatrix} \sim \text{MVN}\left(\mu_I = \begin{pmatrix} 0 \\ 3.6 \end{pmatrix}, \Sigma_I = \begin{pmatrix} 1 & 0.15 \\ 0.15 & 0.2 \end{pmatrix}\right)$ (Gorney et al., 2024), $a_j \sim \log N(0, 0.5)$, and $\alpha_j \sim U(1.75, 3.25)$ (Lu & Wang, 2020; Lu et al., 2020, 2023). Again, the time limits for tests with 20, 50, and 80 items were set at 30, 75, and 120 min, respectively. In addition, we evaluated the sensitivity of our proposed SIC-CPA approaches under four distinct scenarios. Scenarios 1–4 explored two common underlying

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

Table 4. Generated models and parameter settings for responses and RTs under four scenarios in simulation study III.

Scenario	Generated model	Responses	RTs
1	GCM: Eqs. (30)–(33)	$\lambda_i \sim \log N(3.912,1)$	$\lambda_i \sim \log N(3.912,1)$
2	HM+ GCM: Eqs. (31)–(34)	$g = 0$	$\lambda_i \sim$ Uniform (10,20)
3	HM+ GCM: Eqs. (31)–(34)	$g = 0.25$	$\lambda_i \sim \log N(3.912,1)$
4	HM: Eqs. (32)–(35)	$g = 0$	$C_{ij} \sim U(1\text{ sec}, 20\text{ sec})$

mechanisms of test speededness (Yu & Cheng, 2020): gradual change model (GCM) and hybrid model (HM) with abrupt changes, respectively. Scenario 1 serves as a baseline scenario, similar to the setups used in simulations 1 and 2. Equations (34) and (35) describe the HM models for responses and RTs, respectively. Thus, Eqs. (32)–(35) constitute the joint HM model for responses and RTs. Scenarios 2 and 3 manipulated the abrupt changes in correct response probability together with gradual changes in RTs after the change points. This occurs when examinees shift to providing incorrect or random answers, leading to reduced accuracy and progressively faster responses (i.e., gradual changes in RTs) as they adapt to the new answering strategy or experience cognitive disengagement. We set $g = 0$ to represent an extreme but practically relevant scenario in which examinees stop solving items due to severe time pressure or disengagement, yielding nearly uniformly incorrect answers toward the end of the test; a similar pattern was reported in Shao et al. (2016, see Figure 4 in their real data analysis). This setting is also reasonable for high-stakes tests with difficult end-of-test items, where accurate responses cannot be made without spending reasonable amount of time. In contrast, $g = 0.25$ represents random guessing behavior (Wang & Xu, 2015), typical in multiple-choice questions with four options. In scenario 2, we use the uniform distribution to simulate λ_i, representing mild speededness where most examinees show gradual performance changes without extreme fluctuations. In contrast, the log-normal distribution of λ simulates heterogeneous speededness, including both gradual and stable declines in performance, as well as extreme declines, reflecting diverse real-world scenarios. Table 4 depicts data generation manners for responses and RTs under four scenarios. In total, there are 720 (i.e., 90 baseline conditions [please see Table 1] × 2 (correlation coefficients) × 4 (scenarios)) simulation conditions. Figure 1 shows the histogram of log RTs for all person-by-item combinations for four scenarios:

$$P_{ij}^{*} = \begin{cases} p(y_{ij} = 1 \mid \theta_i, a_j, b_j), & \text{before the change point } \eta_i \\ g, & \text{after the change point } \eta_i \end{cases}. \tag{34}$$

$$T_{ij}^{*} = \begin{cases} \log(t_{ij}) \sim N\left(\beta_j - \tau_i, \alpha_j^{-2}\right), & \text{before the change point } \eta_i \\ C_{ij}, & \text{after the change point } \eta_i \end{cases}. \tag{35}$$

Tables 5 and 6, along with Supplementary Tables S8–S11, show the detection results of scenario 1 in simulation study III. Overall, compared to the SIC-CPA using only RT data, the SIC-CPA using response and RT data shows slightly higher CCRs, nearly identical power except when the median of η is 0.9, and lower Type I errors. The power of the SIC-CPA using combined data is notably higher than that using only response data. Regarding the accuracy of change point location recovery, the SIC-CPA using only response data performs the worst; in most cases, the SIC-CPA using RT data performs slightly better than when using combined data. This could be attributed to the variability in discrete response data caused by test speededness, which impacts the accuracy of recovering change point locations. In particular, when the change points occur at the end of the test, as seen when the median of η is 0.8 and 0.9, the AL_mean and AL_SD for SIC-CPA using RTs are smaller than those for SIC-CPA using combined data.

When the number of items increases, the power of the SIC-CPA using only response data significantly improves. The Type I errors of both the SIC-CPA using combined data and the SIC-CPA using response data slightly increase. The detection results of the SIC-CPA using only RT data are almost unaffected by

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

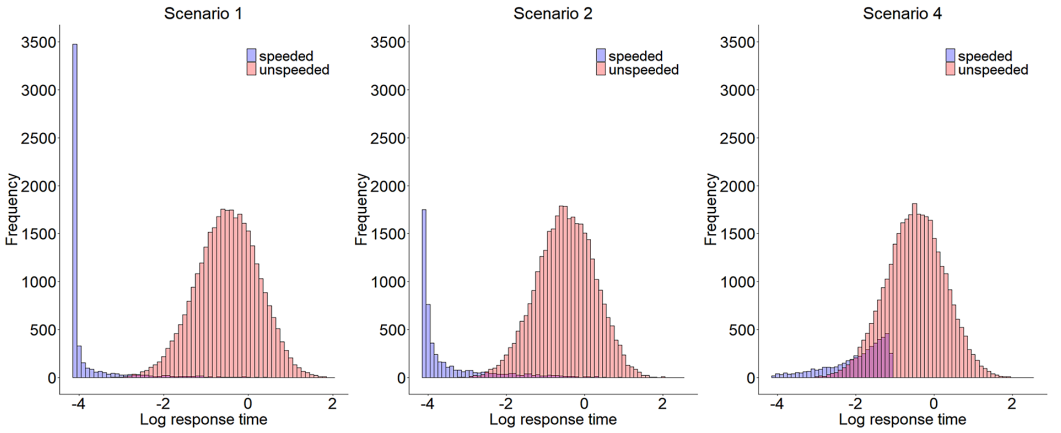


Figure 1. Histogram of log RTs for all person-by-item combinations for four scenarios. Note that the histogram of scenario 3 is identical to that of scenario 1. The "speeded" refers to all speeded RTs at person-by-item level, "unspeeded" indicates all normal RTs at person-by-item level.

the increase in test length. In terms of detecting speeded examinees, the performance of the SIC-CPA using combined data and SIC-CPA using RT data is nearly identical. As the number of items increases, the accuracy of change point location recovery declines for the SIC-CPA using all three types of data, as evidenced by larger AL_mean and AL_SD.

In summary, compared to the SIC-CPA using only response data, the SIC-CPA using combined data and the SIC-CPA using RT data present higher CCRs and power, as well as lower AL_mean and AL_SD. The SIC-CPA using combined data and the SIC-CPA using RT data outperform the SIC-CPA using only response data in detecting speeded examinees and in recovering change point locations.

When the correlation coefficient between θ and τ is 0.8, the trends of results in terms of these evaluation criteria are almost identical to those when the correlation coefficient is 0.5. As the correlation coefficient between θ and τ increases, SIC-CPA using combined data shows particularly notable improvements in detecting speeded examinees, showing higher power compared with SIC-CPA using RTs.

The results of scenarios 2, 3, and 4 under different correlation coefficients between θ and τ, as well as different test lengths, are essentially consistent with the results of scenario 1. Due to the page limit, we only present the results for the correlation coefficient between θ and τ of 0.8, for 500 examinees, across 15 conditions. The specific conditions are shown in Table 7. The detailed results can be found in Supplementary Tables S12–S20.

Figure 2 displays the power, Type I error, and AL_mean of 15 simulation conditions for four scenarios for the condition of 10% speeded examinee proportion with test length 50. In scenario 2, unlike the gradual decline correct response probability observed in scenario 1, the correct response probability of speeded responses suddenly drops to zero. This change significantly improves the performance of the SIC-CPA using response data in detecting speeded examinees and pinpointing change point locations. Furthermore, scenario 2 demonstrates that while the results using RT data alone are less satisfactory than in scenario 1, the results from using response data are superior. Moreover, combining both types of data yields better detection results than using RT data alone, with a higher power than that of using response data alone.

In scenario 3, the correct response probability of speeded responses suddenly drops to 0.25. Compared to scenario 1, the performance of the SIC-CPA using response data for detecting speeded examinees and pinpointing change points diminishes. The power of the SIC-CPA using combined data is identical to those using RT data alone except for the median of η is 0.9. In most conditions, the use of combined data results in lower Type I errors compared to using RT data alone. Therefore,

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

Table 5. Results of the proposed three SIC-CPA approaches when the correlation coefficient of person parameters is 0.5 and test length is 50 for scenario 1 in simulation study III.

			CCR			Power			Type_I			AL_mean			AL_SD			NF%
%	η_{median}	η_{var}	Y&T	RT	Y	Y&T	RT	Y	Y&T	RT	Y	Y&T	RT	Y	Y&T	RT	Y	
10	0.5	0.001	0.994	0.991	0.963	1	1	0.713	**0.007**	**0.010**	0.006	0.664	0.638	1.096	0.679	0.664	1.500	9.132
	0.6	0.001	0.993	0.992	0.956	1	1	0.651	0.008	0.009	0.007	0.589	0.587	2.018	0.645	0.640	2.429	9.248
	0.7	0.001	0.994	0.992	0.946	1	1	0.561	0.007	0.009	0.007	0.593	0.589	2.214	0.620	0.621	3.224	9.352
	0.8	0.001	0.993	0.991	0.934	1	1	0.452	0.008	0.010	0.008	0.532	0.533	2.771	0.591	0.588	4.274	9.272
	0.9	0.001	0.992	0.990	0.917	1	1	0.222	0.008	0.009	0.008	0.435	0.463	2.562	0.526	0.496	4.871	9.304
	0.5	0.01	0.993	0.991	0.962	1	1	0.712	0.007	0.010	0.007	0.671	0.650	1.539	0.722	0.686	2.256	9.100
	0.6	0.01	0.992	0.991	0.955	1	1	0.643	0.009	0.010	0.007	0.622	0.611	1.818	0.663	0.650	2.619	9.212
	0.7	0.01	0.993	0.992	0.946	1	1	0.556	0.008	0.009	0.007	0.565	0.565	2.232	0.624	0.625	3.334	9.284
	0.8	0.01	0.993	0.992	0.931	0.999	0.999	0.418	0.007	0.009	0.008	0.527	0.529	2.539	0.590	0.603	4.178	9.272
	0.9	0.01	0.974	0.975	0.954	**0.895**	**0.795**	**0.357**	0.008	0.009	0.008	0.493	0.492	2.824	0.934	0.636	4.977	9.264
	0.5	0.04	0.993	0.991	0.955	1	1	0.652	0.008	0.010	0.008	0.657	0.644	1.641	0.695	0.673	2.299	9.204
	0.6	0.04	0.993	0.992	0.948	1.000	1.000	0.579	0.008	0.009	0.007	0.597	0.591	1.694	0.655	0.630	2.513	9.180
	0.7	0.04	0.992	0.990	0.939	0.985	0.987	0.492	0.007	0.009	0.007	0.559	0.555	1.989	0.631	0.624	3.005	9.304
	0.8	0.04	0.981	0.981	0.923	0.892	0.899	0.335	0.008	0.009	0.007	0.521	0.520	2.446	0.767	0.755	3.904	9.420
	0.9	0.04	0.924	0.922	0.914	**0.439**	**0.428**	**0.165**	0.008	0.009	0.007	0.620	0.660	2.621	1.202	0.804	3.679	9.148

(Continued)

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

Table 5. Continued.

			CCR			Power			Type_I			AL_mean			AL_SD			NF%
%	η_{median}	η_{var}	Y&T	RT	Y	Y&T	RT	Y	Y&T	RT	Y	Y&T	RT	Y	Y&T	RT	Y	
30	0.5	0.001	0.994	0.993	0.895	1	1	0.689	0.008	0.011	0.007	1.047	1.030	1.382	1.086	1.079	1.947	7.072
	0.6	0.001	0.995	0.993	0.879	1	1	0.637	0.008	0.010	0.008	0.974	0.973	2.161	0.995	0.990	2.554	7.416
	0.7	0.001	0.994	0.993	0.861	1	1	0.581	0.008	0.011	0.008	0.899	0.899	2.567	0.868	0.865	3.667	7.840
	0.8	0.001	0.994	0.993	0.824	1	1	0.459	0.008	0.010	0.009	0.796	0.806	3.028	0.784	0.786	5.013	8.268
	0.9	0.001	0.990	0.986	0.746	1	1	0.204	0.008	0.010	0.009	0.645	0.542	3.401				
	0.5	0.01	0.994	0.993	0.895	1	1	0.692	0.008	0.010	0.009	1.086	1.061	1.734	1.119	1.090	2.382	7.180
	0.6	0.01	0.995	0.994	0.877	1	1	0.632	0.008	0.010	0.008	0.990	0.982	2.051	0.991	0.982	2.829	7.432
	0.7	0.01	0.995	0.994	0.856	1	1	0.561	0.008	0.009	0.007	0.891	0.892	2.551	0.896	0.892	3.593	7.860
	0.8	0.01	0.993	0.993	0.810	0.995	0.997	0.413	0.008	0.009	0.008	0.773	0.775	2.472	0.771	0.770	4.025	8.264
	0.9	0.01	0.940	0.921	0.748	**0.821**	**0.742**	**0.201**	0.008	0.009	0.008	0.631	0.560	2.804	1.125	0.670	4.953	8.657
	0.5	0.04	0.995	0.993	0.879	1.000	1.000	0.638	0.008	0.011	0.007	1.033	1.016	1.969	1.082	1.065	2.680	7.272
	0.6	0.04	0.993	0.992	0.864	0.997	0.998	0.590	0.008	0.010	0.009	0.965	0.960	2.058	1.014	0.999	2.956	7.588
	0.7	0.04	0.986	0.984	0.832	0.975	0.976	0.485	0.009	0.012	0.007	0.878	0.867	2.241	1.010	0.962	3.630	8.044
	0.8	0.04	0.951	0.952	0.788	0.861	0.870	0.337	0.008	0.010	0.007	0.764	0.802	2.426	1.065	1.373	4.019	8.540
	0.9	0.04	0.819	0.782	0.740	**0.403**	**0.247**	**0.112**	0.009	0.0.10	0.008	0.895	0.720	2.840	1.784	1.768	4.892	8.820

Note: "Y&T" indicates SIC-CPA approach based on responses and RTs, "RT" indicates SIC-CPA approach based on response times, and "Y" indicates SIC-CPA approach based on responses. Boldface values indicate substantially lower detection performance than other conditions within the same scenario.

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

Table 6. Results of the proposed three SIC-CPA approaches when the correlation coefficient of person parameters is 0.8 and test length is 50 for scenario 1 in simulation study III.

			CCR			Power			Type_I			AL_mean			AL_SD			NF%
%	η_{median}	η_{var}	Y&T	RT	Y	Y&T	RT	Y	Y&T	RT	Y	Y&T	RT	Y	Y&T	RT	Y	
10	0.5	0.001	0.993	0.991	0.960	1	1	0.701	0.008	0.010	0.008	0.645	0.626	1.060	0.673	0.667	1.396	8.256
	0.6	0.001	0.993	0.991	0.957	1	1	0.658	0.008	0.010	0.007	0.601	0.601	1.932	0.635	0.634	2.354	8.204
	0.7	0.001	0.992	0.990	0.948	1	1	0.577	0.009	0.011	0.007	0.595	0.593	2.343	0.617	0.617	3.466	8.230
	0.8	0.001	0.993	0.992	0.937	1	1	0.480	0.008	0.009	0.007	0.526	0.526	2.683	0.588	0.587	4.180	8.415
	0.9	0.001	0.993	0.991	0.916	1	1	0.235	0.008	0.009	0.007	0.426	0.477	2.401	0.519	0.505	4.640	8.804
	0.5	0.01	0.993	0.991	0.961	1	1	0.699	0.008	0.010	0.007	0.672	0.652	1.493	0.711	0.689	2.009	8.332
	0.6	0.01	0.992	0.991	0.954	1	1	0.647	0.008	0.010	0.008	0.622	0.616	1.832	0.657	0.649	2.571	8.125
	0.7	0.01	0.993	0.992	0.947	1	1	0.567	0.008	0.009	0.007	0.572	0.574	2.208	0.632	0.631	3.431	8.404
	0.8	0.01	0.992	0.991	0.933	0.998	0.998	0.436	0.009	0.010	0.007	0.526	0.521	2.373	0.588	0.585	3.897	8.415
	0.9	0.01	0.979	0.978	0.917	**0.872**	**0.778**	**0.234**	0.009	0.009	0.008	0.486	0.457	2.828	0.927	0.611	4.800	8.342
	0.5	0.04	0.993	0.991	0.957	1.000	1.000	0.665	0.008	0.010	0.007	0.639	0.622	1.605	0.677	0.662	2.298	8.320
	0.6	0.04	0.992	0.991	0.950	0.999	0.999	0.610	0.009	0.010	0.008	0.603	0.605	1.890	0.643	0.640	3.064	8.326
	0.7	0.04	0.992	0.990	0.941	0.989	0.990	0.508	0.007	0.010	0.007	0.556	0.556	1.961	0.626	0.619	3.271	8.521
	0.8	0.04	0.982	0.982	0.925	0.903	0.913	0.366	0.009	0.010	0.008	0.506	0.517	2.230	0.726	0.838	3.962	8.614
	0.9	0.04	0938	0.935	0.912	**0.448**	**0.409**	**0.172**	0.009	0.010	0.007	0.617	0.667	2.570	1.101	0.794	3.849	8.201

(Continued)

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

Table 6. Continued.

			CCR			Power			Type_I			AL_mean			AL_SD			NF%
%	η_{median}	η_{var}	Y&T	RT	Y	Y&T	RT	Y	Y&T	RT	Y	Y&T	RT	Y	Y&T	RT	Y	
30	0.5	0.001	0.994	0.992	0.903	1	1	0.716	0.009	0.012	0.008	1.050	1.036	1.447	1.074	1.072	2.115	6.602
	0.6	0.001	0.995	0.993	0.884	1	1	0.653	0.008	0.010	0.008	0.990	0.993	2.219	1.011	1.010	2.668	6.504
	0.7	0.001	0.994	0.993	0.865	1	1	0.592	0.008	0.010	0.007	0.896	0.897	2.504	0.872	0.872	3.568	6.904
	0.8	0.001	0.995	0.993	0.834	1	1	0.486	0.008	0.010	0.007	0.795	0.797	2.880	0.784	0.776	4.459	7.502
	0.9	0.001	0.992	0.987	0.755	1	1	0.216	0.009	0.010	0.007	0.600	0.533	2.374	0.670	0.621	5.348	8.464
	0.5	0.01	0.994	0.992	0.902	1	1	0.711	0.009	0.011	0.008	1.083	1.069	1.715	1.106	1.099	2.282	6.612
	0.6	0.01	0.994	0.993	0.882	1	1	0.647	0.009	0.010	0.008	0.981	0.975	2.062	1.002	0.994	2.870	6.606
	0.7	0.01	0.994	0.994	0.861	1	1	0.579	0.008	0.009	0.008	0.890	0.888	2.495	0.888	0.885	3.685	7.024
	0.8	0.01	0.992	0.992	0.817	0.994	0.995	0.432	0.008	0.009	0.008	0.769	0.771	2.516	0.776	0.773	4.192	7.606
	0.9	0.01	0.942	0.924	0.756	**0.834**	**0.752**	**0.220**	0.009	0.009	0.008	0.623	0.585	2.700	1.004	0.688	5.040	8.680
	0.5	0.04	0.995	0.994	0.886	0.999	0.999	0.661	0.007	0.009	0.008	1.042	1.027	1.877	1.079	1.066	2.480	6.724
	0.6	0.04	0.993	0.992	0.868	0.996	0.997	0.601	0.008	0.010	0.007	0.967	0.955	2.114	1.045	1.024	3.211	7.002
	0.7	0.04	0.987	0.987	0.839	0.978	0.980	0.505	0.008	0.010	0.007	0.850	0.844	2.344	0.966	0.930	3.887	7.416
	0.8	0.04	0.950	0.952	0.793	0.857	0.869	0.355	0.008	0.010	0.008	0.753	0.804	2.424	1.083	1.448	4.249	7.724
	0.9	0.04	0.817	0.777	0.737	**0.417**	**0.255**	**0.150**	0.008	0.010	0.008	0.900	0.714	2.844	1.800	1.750	4.666	8.768

Note: "Y&T" indicates SIC-CPA approach based on responses and RTs, "RT" indicates SIC-CPA approach based on response times, and "Y" indicates SIC-CPA approach based on responses. Boldface values indicate substantially lower detection performance than other conditions within the same scenario.

Table 7. Simulation conditions with different values for the median and variance of η in four scenarios of simulation study III.

Simulation condition	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Median of η	0.5	0.6	0.7	0.8	0.9	0.5	0.6	0.7	0.8	0.9	0.5	0.6	0.7	0.8	0.9
Variance of η	0.001	0.001	0.001	0.001	0.001	0.01	0.01	0.01	0.01	0.01	0.04	0.04	0.04	0.04	0.04

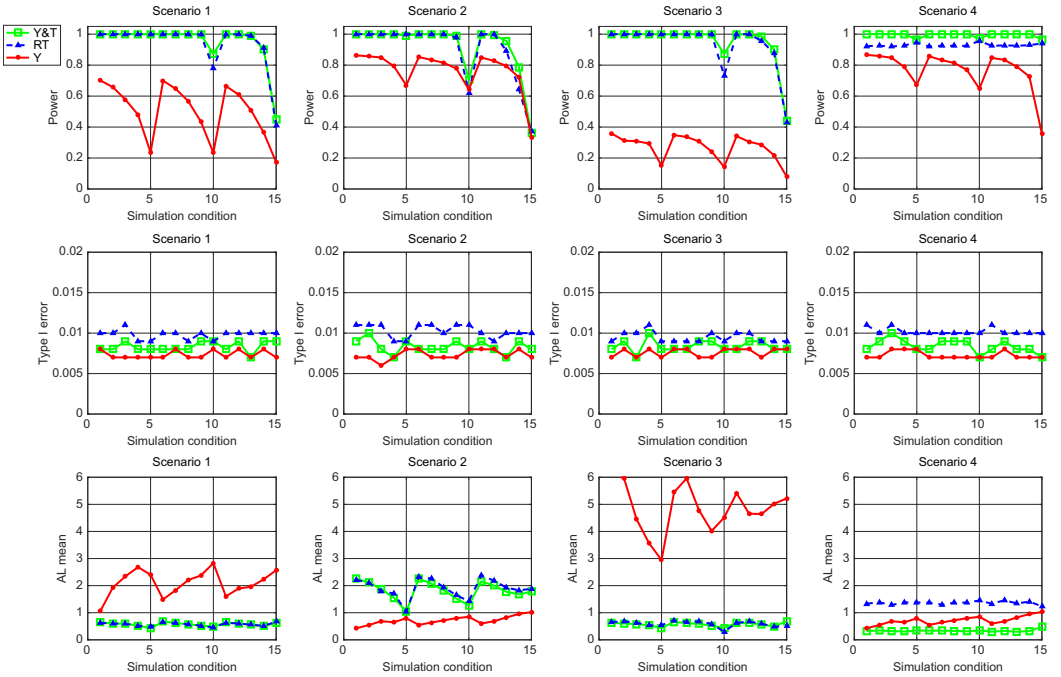


Figure 2. Power, Type I errors, and AL_mean of 15 simulation conditions for 4 scenarios in simulation study III.

combining response and RT data proves to be more effective than using either response or RT data alone. Furthermore, even when response data alone shows reduced performance, the inclusion of RT data allows the SIC-CPA to maintain high performance, demonstrating the robustness of the combined data approach.

Figure 1 shows that in scenario 4, the speeded RTs largely overlap with normal RTs, resulting in suboptimal detection results when using RTs alone, as expected. However, even with a high degree of overlap between speeded and normal RTs, the power of the SIC-CPA using RT data consistently exceeds 90%, with Type I errors staying below 0.012. Figure 2 illustrates that using combined data provides greater advantages in identifying speeded examinees and pinpointing change point locations. For instance, all AL_mean values detected using combined data are consistently lower than those detected using response data alone. Therefore, even when there is little distinction between normal and speeded RTs, with a clear distinction between normal and speeded response data, the SIC-CPA using combined data still effectively detects the speeded examinees and provides better accuracy in change point recovery.

Table 8 displays the average running times for the SIC-CPA using three types of data, under the test lengths of 20, 50, and 80. It is observed that the SIC-CPA using only RT data operates faster, while the running time for using combined data and response data are closely comparable.

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

Table 8. Average running time of SIC-CPA approach when sample size is 500 in simulation study III.

Test length	Y&T (seconds)	RT (seconds)	Y (seconds)
20	1.1	0.3	0.9
50	2.9	0.5	2.7
80	6.0	0.6	5.3

Note: "Y&T" indicates SIC-CPA approach based on responses and RTs, "RT" indicates SIC-CPA approach based on response times, and "Y" indicates SIC-CPA approach based on responses.

3.6. Simulation study IV

This simulation study was conducted to evaluate the performance of the SIC-CPA approach based on RTs to identify multiple change points. The study does not focus on aberrant behaviors occurring at isolated items, as these could lead to a large number of change points, making accurate detection more challenging. Instead, we manipulated the common aberrant test-taking behavior scenarios, that is, the warm-up effect at the very beginning of the test and the test speededness at the end of the test due to the time limit, resulting in two change points. Essentially, these two common aberrant behaviors can be seen as forms of speeding-up behavior, reflected in an increasing trend of speed parameters throughout different phases of the test. Specifically, examinees often struggle to settle at the beginning of the test, and the warm-up effect leads to longer RTs (Shao, 2017), resulting in lower speed parameters. Afterward, they respond at a normal pace, exhibiting typical speeds. Finally, due to time pressure, they speed up toward the end of the test, which is reflected in higher speeds.

Following the simulation study designed by Zhu et al. (2023) to manipulate RTs with the warm-up effect and test speededness, we generated the RTs using the following model proposed by van der Linden and van Krimpen-Stoop (2003):

$$\log\left(t_{ij}\right) = \beta_j - \tau_i + L + \varepsilon_{ij}, \varepsilon_{ij} \sim N\left(0, \alpha_j^{-2}\right),$$

where L captures the speed shift caused by aberrant behavior, reflecting the extent of an abrupt change in an examinee's speed. When $L > 0$, longer RTs indicate warm-up effect, whereas when $L < 0$, shorter RTs reflect test speededness. When $L = 0$, the model simplifies to Eq. (13), representing the regular behavior. We set the first change point occurring between items 5 and 8, which means the warm-up effect is observed across items 1–4, 1–5, 1–6, and 1–7, respectively. The second change point is determined by the median and variance of η, with the specific values provided in Table 1. For the RTs associated with the warm-up effect, we set $L = 1$, and for the RTs related to test speededness, $L = -1$ or -2. The other model parameters are consistent with those in simulation study II. Thus, 2 (speed shifts for test speededness, we denote two scenarios as $L_{\text{speeded}} = -1$ or -2) $\times$ 90 (baseline condition in Table 1) produces 180 simulation conditions.

Supplementary Tables S21–S26 show the detection results of the SIC-CPA approach based on RTs for identifying two change points in simulation study IV. Figure 3 depicts the results of simulation study IV under 10% speeded examinee proportion, including power, type I error rate, and AL_mean for each of the two change points (i.e., $\text{Al_mean}_{\text{warm-up}}$, $\text{Al_mean}_{\text{speeded}}$) as well as their AL_mean summation (i.e., $\text{Al_mean}_{\text{sum}}$). Across all test lengths, SIC-CPA approach under $L_{\text{speeded}} = -2$ consistently shows higher power than those under $L_{\text{speeded}} = -1$. When test length is 20, the power under $L_{\text{speeded}} = -2$ is below 0.9, while power under $L_{\text{speeded}} = -1$ falls under 0.6. When η_{median} is set at 0.5 or 0.9, the power is notably lower than when η_{median}, which takes values of 0.6, 0.7, or 0.8. For a test length of 20, the reduced power for η_{median} =0.5 and 0.9 can be attributed to a shortage of items between the two change points. Specifically, for example, when η_{median} is 0.5, the second change point shows only minor variability around item 10 under $\eta_{\text{var}} = 0.001$, while the first change point is located between items 5 and 8. This creates an interval of only one to four items between the two change points, severely limiting the

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

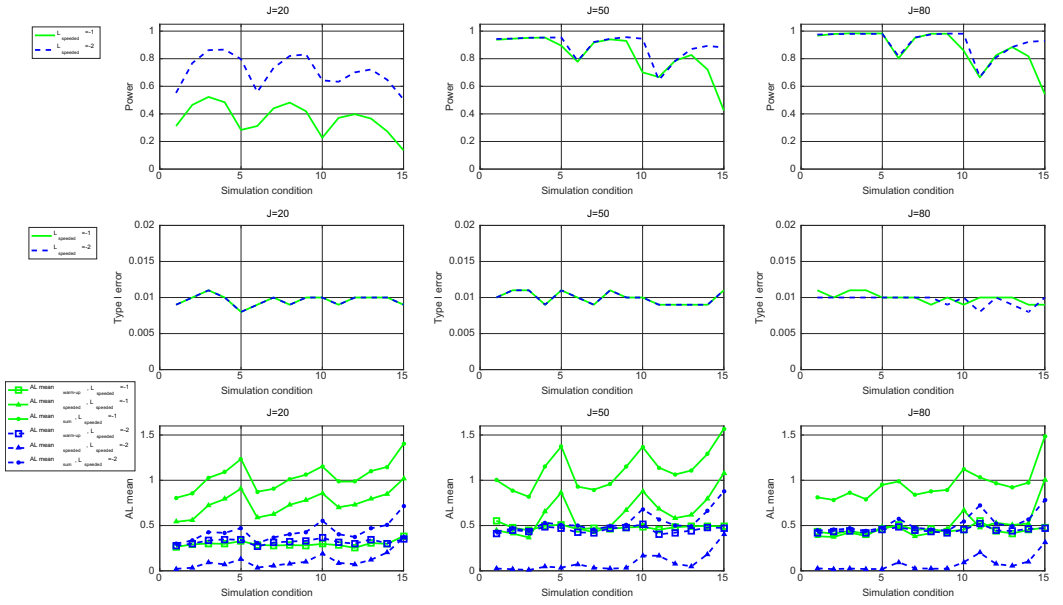


Figure 3. Power, Type I errors, and AL_mean of 15 simulation conditions for $L_{\text{speeded}} = -1$ and -2 in simulation study IV.

SIC-CPA detection accuracy for the first change points. A similar issue arises when η_{median} is 0.9, as the few items following the second change point hinder effective change point recovery for the second change points. Thus, when the test length is 20, for the fixed η_{var} conditions, power tends to increase with η_{median} initially but then decreases finally. This occurs because when η_{median} is 0.5 or 0.9, that is, the average number of items between two change points is too short to effectively detect the two change points. For longer tests (50 and 80 items), the power pattern differs with the speed shift condition. Under $L_{\text{speeded}} = -2$, for the fixed η_{var}, increasing η_{median} results in a larger number of items with regular RTs between two change points, which in turn enhances power. In contrast, under $L_{\text{speeded}} = -1$, power initially increases with η_{median} but eventually declines. This decline is due to the second change point moving closer to the end of the test, where the difference between regular and speeded RTs becomes less distinct, making it more difficult to identify both change points.

Across all conditions, the type I error rates are effectively controlled around 0.01. Regarding the AL_mean, the total AL_mean for both change points is higher than that of each change point individually; however, the total AL_mean remains below 1.5, indicating the satisfactory performance in recovering the true change point locations. Note that under $L_{\text{speeded}} = -1$, the recovery of the second change point is typically less accurate than that of the first change point, likely because the difference between regular and speeded RTs is relatively modest. In contrast, under $L_{\text{speeded}} = -2$, the larger RT difference enhances the recovery of the second change point. For test length of 20, given that the absolute RT deviation for a speed shift of -1 (i.e., L_{speeded}) is equivalent to that for a shift of 1 (i.e., $L_{\text{warm-up}}$) relative to regular RTs, the improved recovery under $L_{\text{speeded}} = -2$ indicates that the speeded change points (i.e., the second change points) are more distinctly identified than warm-up change points (i.e., the first change points) associated with $L_{\text{warm-up}} = 1$.

3.7. Simulation study V

In the main simulation studies I–IV, item parameters are treated as known to focus on the change point detection performance of SIC-CPA. This assumption is common in many testing programs where item parameters are assumed estimated without error (Cheng & Shao, 2022; Shao et al., 2016; Tu et al., 2023; Yu & Cheng, 2020). However, item parameters are often estimated using the same operational

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

data that may already include speeded responding. Such contamination can bias item calibration, thereby affecting both (1) the estimation of person parameters used in the test statistics and (2) the null distribution used to determine critical values. Thus, we conduct an additional simulation in which item parameters are unknown and need to be estimated from contaminated response and RT data, which closely aligned with the real data example. We compare the performance of SIC-CPA, the likelihood ratio test, and the Wald test under two two-stage calibration-and-detection procedures.

Case 1: Non-iterative calibration-detection

We adopt a two-stage procedure that mirrors common practice. First, item parameters are estimated from the full dataset using a joint model of responses and RTs. These estimates are then treated as fixed and used in three change point detection approaches.

Case 2: Iterative detect-clean-recalibrate

We further examine whether an iterative approach can improve detection. The detection methods are applied to all examinees using the initial item parameter estimates. After detection, the responses and RTs occurring after each examinee's detected change point are deleted. The item parameters are then re-estimated using the remaining pre-change data only so that post-change (potentially speeded) observations do not influence calibration. This "Iterative detect-clean-recalibrate" procedure is repeated until the maximum absolute change between two consecutive sets of item parameter estimates is smaller than 0.01, or until five iterations are reached to prevent non-terminating updates.

We employed the *LNIRT* package in R (Fox et al., 2023; R Core Team, 2022) to estimate the model parameters. Critical values were generated under the null hypothesis conditional on the item parameter estimates. This design ensures that both the detection statistics and their critical values are evaluated within a coherent model framework, which is essential when item parameters are unknown in practice. This approach reflects practical testing conditions and maintains internal consistency between model estimation and subsequent detection. Specifically, within each condition and case, we generate null datasets from the fitted *no-change-point* joint model, recompute the corresponding test statistic using the same estimation and detection procedure, and take the $(1-\alpha)$-quantile as the critical value ($\alpha = 0.01$). Thus, the nominal Type I error is defined relative to the joint response and RT model implied by the estimated parameters. Note that item parameter estimation is affected by sample size. Accordingly, in addition to the 90 baseline simulation conditions summarized in Table 1, we further included conditions with a larger sample of 1,000 examinees, resulting in a total of 180 simulation settings. The results exhibited similar overall patterns across conditions, and due to the page limit, we report only the results corresponding to $\eta_{\text{median}} = 0.7$ in the main text.

The likelihood ratio test yields slightly lower detection accuracy than SIC-CPA across the contaminated calibration conditions. Thus, Table 9 summarizes the performance of SIC and Wald based on the joint response and RT data in these two cases. Compared with Supplementary Table S5 where item parameters are known, Case 1 shows a clear decline in detection performance. Under Case 1, both SIC and Wald achieved high CCR and power across most conditions, indicating that the combination of responses and RTs provides substantial information for detecting test speededness. However, the key limitation of Case 1 emerged under heavy contamination (i.e., 30% speeded examinee proportion): the Type I errors were noticeably inflated in several settings, and this inflation was typically larger for Wald than for SIC. This likely occurs because non-iterative calibration–detection procedure on a heavily contaminated sample can systematically bias item parameters, especially for later items. The inflation is stronger when $\eta_{\text{var}} = 0.001$ or 0.01 because change points cluster near $0.7 \times J$, so many speeded examinees affect the same set of later items, creating a concentrated calibration bias that increases false alarms for normal examinees. When $\eta_{\text{var}} = 0.04$, change points are more dispersed, the speeded impact is spread across different item positions, and the calibration distortion is less concentrated, helping keep Type I error closer to the nominal level. For the same reason, localization can be less distorted and thus more stable, leading to smaller AL_mean and AL_SD, even though change point locations are

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

Table 9. Results of Wald test and the proposed SIC-CPA approach when test length is 20 and η =0.7 in simulation study V.

				CCR		Power		Type I error		AL_mean		AL_SD		Number of iterations	
N	%	η_{var}	Case	SIC	Wald	SIC	Wald	SIC	Wald	SIC	Wald	SIC	Wald	SIC	Wald
500	10	0.001	1	0.989	0.986	0.984	0.980	0.005	0.007	0.483	0.461	0.600	0.601	1	1
		0.001	2	0.993	0.990	0.987	0.983	0.006	0.006	0.418	0.431	0.528	0.596	3.95	3.95
		0.01	1	0.985	0.986	0.980	0.976	0.004	0.003	0.417	0.406	0.579	0.584	1	1
		0.01	2	0.992	0.987	0.982	0.978	0.005	0.006	0.365	0.379	0.534	0.57	3.95	3.55
		0.04	1	0.987	0.981	0.879	0.873	0.006	0.012	0.332	0.356	0.513	0.539	1	1
		0.04	2	0.990	0.982	0.882	0.880	0.006	0.007	0.321	0.384	0.507	0.67	4.45	4.2
	30	0.001	1	0.933	0.932	0.957	0.974	0.078	0.089	0.508	0.543	0.689	0.686	1	1
		0.001	2	0.995	0.992	0.993	0.994	0.005	0.008	0.414	0.436	0.551	0.587	4.45	4.4
		0.01	1	0.954	0.953	0.961	0.975	0.050	0.058	0.378	0.382	0.555	0.565	1	1
		0.01	2	0.989	0.984	0.981	0.978	0.006	0.003	0.328	0.347	0.509	0.545	4.75	4.5
		0.04	1	0.954	0.944	0.871	0.863	0.018	0.015	0.328	0.342	0.518	0.57	1	1
		0.04	2	0.958	0.955	0.885	0.886	0.007	0.002	0.321	0.375	0.526	0.645	4.9	4.95
1000	10	0.001	1	0.986	0.983	0.982	0.985	0.007	0.002	0.482	0.445	0.595	0.608	1	1
		0.001	2	0.992	0.990	0.983	0.980	0.007	0.004	0.390	0.402	0.537	0.569	4.05	3.95
		0.01	1	0.991	0.989	0.991	0.987	0.007	0.003	0.390	0.405	0.568	0.587	1	1
		0.01	2	0.993	0.990	0.969	0.966	0.007	0.005	0.349	0.387	0.539	0.589	4.45	4.25
		0.04	1	0.985	0.981	0.904	0.903	0.006	0.005	0.338	0.378	0.532	0.592	1	1
		0.04	2	0.986	0.983	0.903	0.900	0.007	0.002	0.338	0.408	0.546	0.653	4.8	4.4
	30	0.001	1	0.933	0.928	0.958	0.975	0.078	0.094	0.498	0.530	0.658	0.671	1	1
		0.001	2	0.995	0.982	0.994	0.990	0.005	0.002	0.418	0.427	0.559	0.576	4.7	4.75
		0.01	1	0.950	0.955	0.962	0.973	0.056	0.053	0.385	0.388	0.557	0.573	1	1
		0.01	2	0.985	0.984	0.971	0.973	0.008	0.012	0.336	0.384	0.512	0.585	4.9	4.8
		0.04	1	0.957	0.954	0.876	0.871	0.015	0.016	0.327	0.35	0.527	0.569	1	1
		0.04	2	0.962	0.961	0.906	0.900	0.008	0.011	0.322	0.38	0.548	0.684	5	5

Note: % shows the proportion of speeded examinees, SIC indicates the proposed SIC-CPA approach and Wald indicates Wald test.

more heterogeneous. The Wald test is particularly sensitive because it relies directly on the estimated covariance (inverse information) of the person parameter estimates, making it more susceptible to calibration-induced instability.

In contrast, Case 2 largely maintained the high power of Case 1 while substantially improving Type I control, especially for the conditions with 30% speeded examinee proportion, and it typically yielded equal or smaller AL_mean and AL_SD, indicating more accurate and stable localization. This improvement is consistent with the purpose of Case 2: iterative refinement reduces the impact of speeded patterns on calibration, yielding more stable inference. The last two columns of Table 9 show that the iterative procedure converges with an average of about four iterations across conditions, suggesting that only a small number of updates are needed to stabilize the calibration and the subsequent detection results.

4. Real data analysis

The real data were collected from a computer-based stage mathematics power test taken by ninth-grade students at a secondary school. Six hundred students answered 75 dichotomously scored multiple-choice items, with a time limit of 3 h. Four students were removed due to RT of 0 at the end of the test, clearly indicating test speededness. The remaining 596 students were analyzed using the proposed SIC-CPA methods, which are based on three types of data: responses, RTs, and a combination of both.

For this dataset with unknown item parameters, we adopted the Iterative detect-clean-recalibrate strategy (i.e., Case 2 in simulation study V) to detect change points and flag examinees with aberrant behavior. Finally, 26, 25, and 28 examinees identified as speeded using SIC-CPA based on responses, RTs, and a combination of both, respectively. Fleiss' Kappa, a metric that assesses consistency in detection results, was used to evaluate the SIC-CPA methods using three types of data. The value of 0.278 indicates fair consistency among these three methods. Meanwhile, we also conducted the real data analysis using likelihood ratio test and Wald test based on responses, RTs, and a combination of both. The Fleiss' Kappa value for the likelihood ratio test based on three types of data is 0.273, and the Fleiss' Kappa value for the Wald test based on three types of data is 0.275. In addition, we found high Fleiss' Kappa of 0.972 for the test-taking indicators across these three different methods for each data type, where speeded examinees were flagged as 1 and non-speeded examinees as 0. This reveals different patterns of test speededness within the dataset, that is, some speeded examinees showed reduced accuracy, others exhibited increased speed, and some displayed both decreased accuracy and increased speed. A similar finding of heterogeneity in the speeded patterns of examinees was also reported in Gorney et al. (2024). Further analysis will focus on the data from individual examinees to more thoroughly figure out these patterns.

Table 10 shows examinees' correct response rates, average RTs, ability parameter estimates, and speed parameter estimates before and after the change points. Examinee 245 was flagged as speeded using SIC-CPA based on RT data, with the change point location of 56. Both before and after the change point, this examinee's correct response rates and ability parameters were consistently low. Additionally, there was a significant decrease in RT and an increase in speed parameters after the change point, which exactly occurs test speededness for this examinee. Examinee 254 was identified as speeded using SIC-CPA based on response data, with a change point location of 59. After this change point, there was a slight decrease in the examinee's RTs and an increase in speed parameters, which is consistent with characteristics of test speededness. Furthermore, there was a significant drop in the correct response rates and ability parameters following the change point. As a result, the test speededness could be detected using the SIC-CPA method that relies solely on response data.

Taking examinee 403 as an example, he/she was identified as speeded using SIC-CPA method, which combines response and RT data. The RTs for this examinee, as shown in Figure 4, suggest that it is challenging to determine test speededness based solely on RT data. The SIC-CPA using combined data indicated that the examinee began to speed up from item 41 onwards. The average RT for the first 40 items was 2.4 min per item, with a correct response rate of 83%. From items 41–75, the average RT decreased to 1.2 min per item, and the correct response rate dropped to 50%. Additionally, ability parameter estimates calculated from the joint model before and after the change point showed the ability

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

Table 10. Examinees' correct response rates, average RTs, ability parameter estimates, and speed parameter estimates before and after the change points.

Examinee		Correct response rates	Average RTs	Ability parameter estimate	Speed parameter estimate
245	Before the detected change point	0.290	0.900	−1.106	0.630
	After the detected change point	0.260	0.200	−1.196	2.521
254	Before the detected change point	0.800	1.700	0.414	−0.203
	After the detected change point	0.500	1.600	−0.295	−0.114

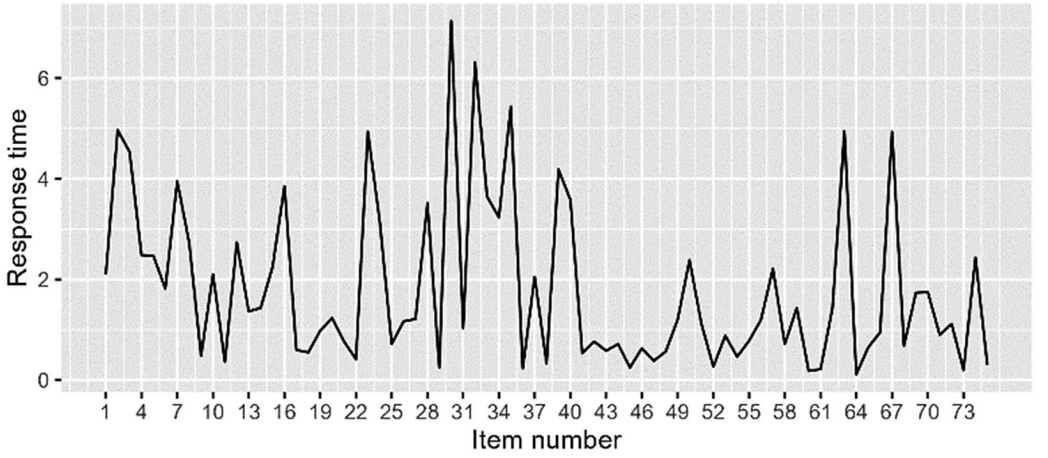


Figure 4. Response times of 75 items for examinee 403. Note that the unit of response time is minutes.

estimate of 1.3 before and a notable decrease to 0.09 afterwards, which aligns with the characteristics of test speededness.

5. Discussion

This article provides a clear and robust decision framework for hypothesis testing, that is, SIC-CPA method, to detect test speededness in educational testing. To identify the change point, one simply needs to compare the SIC values of two models (with and without the change point) by iterating through all possible change point locations and calculating the corresponding SICs. The model with the lowest SIC is selected as the most likely change point location. This new method is particularly suitable for post-hoc analysis of collected data after the test is completed. It can be applied to various testing modes, including both traditional paper-and-pencil tests with only response data and computer-based tests that record RTs. Importantly, this study represents the first comprehensive analysis of test speededness using model selection principle within a hypothesis testing framework, considering both response and RT data simultaneously.

Simulation studies confirm the superior performance of the SIC-CPA approach that integrates both response and RT data, which significantly enhances power and reduces Type I errors. This comprehensive use of data types is especially advantageous in cases involving examinees with lower abilities (or high speeds), where RT (or response) data provide critical additional insights that are not apparent from response accuracy (or response times) alone. This advantage is particularly evident when a high proportion of examinees exhibit speeded behaviors, where the SIC-CPA method demonstrates

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

superior capability in identifying the exact locations of change points, especially as examinees manifest speededness closer to the end of the test. Even in scenarios where there is some overlap between speeded and normal RTs, the approach focusing on RT data maintains a high detection power, while keeping low Type I error rates. This robustness in detection exemplifies the method's efficacy in accurately distinguishing between regular and speeded behaviors under various testing conditions considering the diverse data variability. However, a limitation of our method is that accurate estimation of person parameters requires a sufficient number of test items. Our simulations indicate that for the SIC-CPA approach based on response data, at least eight items after the speeded change point are necessary to achieve a power above 0.7. For the SIC-CPA approach based on RTs or combined data, at least four items are needed to achieve a power above 0.8.

For the detection of change point due to test speededness, the SIC-CPA approach offers several advantages over likelihood ratio and Wald tests: First, it incorporates a penalty for model complexity, preventing the selection of overly complex models and thus reducing the occurrence of false alarms (Type I error). However, likelihood ratio and Wald tests do not account for model complexity, which can lead to overfitting and potential misidentification of change points. The SIC-CPA approach tends to prefer models without a change point unless the including a change point significantly enhances model fit. Second, the SIC-CPA approach is robust in handling data variability, focusing on different levels of changes in RTs caused by test speededness and the variations in the time dispersion parameter. Besides, Gupta and Chen (1996) introduced the significance level α and its associated critical value c_α, which effectively addresses the issue of fluctuations in responses and/or RTs causing subtle differences in SICs. However, likelihood ratio and Wald tests are relatively more sensitive to random fluctuations in the data, which increases the risk of false alarms. Third, SIC-CPA method is more efficient than these competing approaches, that is, likelihood ratio and Wald tests. Fourth, although the likelihood ratio and Wald tests theoretically follow asymptotic distributions (e.g., chi-square and normal distributions) with known critical values, practical factors—such as test length and the precision of item parameter estimates—can affect their performance. Besides, previous studies (Andrews, 1993; Cheng & Shao, 2022; Sinharay, 2016) suggest that asymptotic critical values can only be directly applied when the change point occurs neither too early nor too late in the test. This means that theoretical asymptotic critical values may not match the empirical critical values observed in simulations or real data, particularly when dealing with small samples or complex models, even if the change points occur at 50% or 90% of the test length. Therefore, Monte Carlo simulations are used to obtain empirical critical values that are more accurate and better suited to the specific testing conditions in this study. The SIC-CPA method can accurately detect the two change points resulted from common warm-up effect and test speededness under the abrupt speed change conditions in the current simulation design. However, its performance is relatively poor when the median of η is 0.5 and 0.9. Future research could explore additional simulation scenarios with multiple change points to further evaluate the SIC-CPA method. Note that detecting multiple change points in practice is often complex, as later change points may be influenced by those detected earlier. This limitation arises from the fact that earlier detected change points can affect next subsequences, potentially leading to inaccuracies in identifying new change points as well as computational efficiency. Simulation Study V mimics the real scenarios when item parameters are estimated from operational data that contain speeded responses. Under heavy contamination, the non-iterative calibrate–detect procedure (Case 1) can show inflated Type I error, with Wald typically more affected than SIC-CPA. The iterative detect–clean–recalibrate procedure (Case 2) improves error control by limiting the impact of post-change observations on calibration, while largely preserving detection power.

The current approach has several limitations, and potential directions for future research are outlined below. First, after detecting speededness and the change point, the final estimation approach of person parameters will depend on the study's goals and the potential consequences of misclassification. To minimize bias while retaining all data, robust estimation (e.g., Hong & Cheng, 2019; Schuster & Yuan, 2011) can be used to downweight aberrant segments. If aberrant data invalidates subsequent responses, it may be better to remove it after the change point before estimating parameters. Second, the SIC-CPA method requires recalculating the SIC value for each item position. This process can be computationally

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

intensive, especially when handling large datasets or numerous potential change points, as it involves repeatedly dividing subsequences. A similar challenge is observed in the approaches by Shao et al. (2016) and Cheng and Shao (2022), where the likelihood ratio test and Wald test also require recalculating parameters for each subsequence. Future research could explore more efficient approaches, such as Bayesian change point detection approaches (Lu et al., 2024), which allow for the detection of multiple change points simultaneously. Third, the constant speed assumption of the SIC-CPA approach overlooks real-world test-taking behaviors such as changes in pace due to fatigue or time management. While this assumption remains prevalent, future studies should investigate models that accommodate variable speeds, such as those explored by Fox and Marianti (2016), to better capture the dynamics of test-taking behavior. Fourth, further research should investigate the reliability of SIC-CPA methods across diverse real-world scenarios, particularly in cases involving skewed or multimodal distributions with aberrant behaviors like cheating and rapid guessing. Note that our proposed model is based on traditional IRT models and lognormal RT models. Therefore, when the data deviates from these models, the SIC-CPA approach may lead to model misfit, which in turn affects the accuracy of change point detection. Therefore, we recommend first testing the model fit of the response and/or RT data. If the IRT or lognormal model is not a good fit, we suggest selecting an alternative response or RT model and applying the SIC-CPA approach for change point detection. In such cases, the IRT or lognormal RT model can be replaced with more appropriate models to estimate the ability or speed parameters. Fifth, RT data may exhibit considerable variability due to factors such as item difficulty, examinee proficiency, and attention levels. In our study, we addressed item difficulty and proficiency using van der Linden's (2007) hierarchical joint model, which estimates speed parameters while accounting for the correlation among person parameters and the correlation among item characteristics. However, the lognormal RT model we employed may not fully capture the complexities introduced by attention variations. Advanced methods, such as diffusion models (Kang et al., 2022), can better elucidate the impact of attention on decision accuracy and response times. Furthermore, monitoring individual fluctuations in attention may help identify aberrant behaviors like carelessness (Curran, 2016), thereby refining the understanding of RT data. From a practical perspective, the SIC-CPA method is not intended to replace joint or process models, nor to provide a full behavioral classification. Instead, it serves as a screening and localization tool that detects when response behavior changes (i.e., change points). In practice, identifying where aberrant responding begins is informative for test design (e.g., test length and item ordering), time-limit evaluation (e.g., assessing whether time pressure emerges too early) to minimize construct-irrelevant variance, as well as test monitoring (e.g., identifying test forms or administrations that show unusually strong performance drops near the end of the test). Accordingly, we recommend using SIC-CPA as a first-step diagnostic procedure to flag potential speededness and localize change points, after which more detailed mixture or process models may be applied to a smaller subset of examinees or test segments for in-depth investigation. In this sense, SIC-CPA complements existing modeling approaches by guiding subsequent analyses and reducing the scope of more computationally intensive modeling in large-scale studies, particularly in settings with pre-calibrated item banks where item parameters are routinely treated as fixed.

Supplementary material. To view supplementary material for this article, please visit http://doi.org/10.1017/psy.2026.10094.

Data availability statement. The real data are not made available due to a confidentiality agreement.

Funding statement. This research was supported by the general project of National Social Science Fund of China on Statistics (Grant No. 23BTJ067).

Competing interests. The authors declared no potential competing interests with respect to the research, authorship, and/or publication of this article.

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.

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

References

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, *61*, 821–856.

Armstrong, R. D., & Shi, M. (2009). Model-free CUSUM methods for person fit. *Journal of Educational Measurement*, *46*(4), 408–428.

Bejar, I. I. (1985). *Test speededness under number-right scoring: An analysis of the test of English as a foreign language* (Report No. ETS-RR-85-11). Educational Testing Services.

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

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331–348.

Boughton, K. A., & Yamamoto, K. (2007). A HYBRID model for test speededness. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 147–156). Springer.

Chen, J., & Gupta, A. K. (1997). Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association*, *92*(438), 739–747.

Chen, J., Gupta, A. K., & Gupta, A. K. (2000). *Parametric statistical change point analysis* (Vol. 192). Birkhäuser.

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.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19.

Demirkaya, O. (2022). *Detecting examinees with item pre-knowledge in computer based large-scale assessment* [Doctoral dissertation]. University of Illinois at Urbana-Champaign.

Demirkaya, O., Bezirhan, U., & Zhang, J. (2023). Detecting item preknowledge using revisits with speed and accuracy. *Journal of Educational and Behavioral Statistics*, *48*(4), 521–542.

Egberink, I. J. L., Meijer, R. R., Veldkamp, B. P., Schakel, L., & Smid, N. G. (2010). Detection of aberrant item score patterns in computerized adaptive testing: A empirical example using the CUSUM. *Personality and Individual Differences*, *48*(8), 921–925.

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement*, *9*(2), 123–131.

Fox, J. P., Klotzke, K., & Simsek, A. S. (2023). R-package LNIRT for joint modeling of response accuracy and times. *PeerJ Computer Science*, *9*, e1232.

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

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.

Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, *73*(1), 65–87.

Gorney, K., Sinharay, S., & Liu, X. (2024). Using item scores and response times in person-fit assessment. *British Journal of Mathematical and Statistical Psychology*, *77*(1), 151–168.

Gorney, K., & Wollack, J. A. (2022). Two new models for item preknowledge. *Applied Psychological Measurement*, *46*(6), 447–461.

Gupta, A. K., & Chen, J. (1996). Detecting changes of mean in multidimensional normal sequences with applications to literature and geology. *Computational Statistics*, *11*(3), 211–221.

Hong, M., & Cheng, Y. (2019). Robust maximum marginal likelihood estimation for item response theory models. *Behavior Research Methods*, *51*, 573–588.

Kang, I., De Boeck, P., & Ratcliff, R. (2022). Modeling conditional dependence of response accuracy and response time with the diffusion item response theory model. *Psychometrika*, *87*(2), 725–748.

Lu, J., & Wang, C. (2020). A response time process model for not-reached and omitted items. *Journal of Educational Measurement*, *57*(4), 584–620.

Lu, J., Wang, C., & Shi, N. (2023). A mixture response time process model for aberrant behaviors and item nonresponses. *Multivariate Behavioral Research*, *58*(1), 71–89.

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. *The 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.

Lu, Y., & Sireci, S. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, *26*, 29–37.

Luecht, R., & Ackerman, T. A. (2018). A technical note on IRT simulation studies: Dealing with truth, estimates, observed data, and residuals. *Educational Measurement Issues and Practice*, *37*(3), 65–76.

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press

Marianti, S., Fox, J.-P., Marianna, A., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*, 426–451.
Meijer, R. R. (2002). Outlier detection in high-stakes certification testing. *Journal of Educational Measurement*, *39*(3), 219–233.
Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107–135.
Meijer, R. R., & Sotaridona, L. S. (2006). *Detection of advance item knowledge using response times in computer adaptive testing* (Technical Report 03-03). Law School Admission Council.
Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*, 200–219.
R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.2.0) [Computer software]. https://www.Rproject.org/
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*, 213–232.
Schuster, C., & Yuan, K.-H. (2011). Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics*, *36*(6), 720–735.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464.
Shao, C. (2017). *Aberrant response detection using change-point analysis*. University of Notre Dame.
Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika*, *81*, 1118–1141.
Sinharay, S. (2016). Person ft analysis in computerized adaptive testing using tests for a change point. *Journal of Educational and Behavioral Statistics*, *41*, 521–549.
Sinharay, S. (2017). Some remarks on applications of tests for detecting a change point to psychometric problems. *Psychometrika*, *82*, 1149–1161.
Sinharay, S., & Johnson, M. S. (2020). The use of item scores and response times to detect examinees who may have benefited from item preknowledge. *British Journal of Mathematical and Statistical Psychology*, *73*, 397–419.
Suh, Y., Cho, S. J., & Wollack, J. A. (2012). A comparison of item calibration procedures in the presence of test speededness. *Journal of Educational Measurement*, *49*(3), 285–311.
Tendeiro, J. N., & Meijer, R. R. (2012). A CUSUM to detect person misfit: A discussion and some alternatives for existing procedures. *Applied Psychological Measurement*, *36*(5), 420–442.
Tu, D., Li, Y., & Cai, Y. (2023). A new perspective on detecting performance decline: A change-point analysis based on Jensen-Shannon divergence. *Behavior Research Methods*, *55*, 963–980.
van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204.
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.
van der Linden, W. J. (2011). Setting time limits on tests. *Applied Psychological Measurement*, *35*(3), 183–199.
van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*(2), 251–265.
van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, *26*(2), 199–217.
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.
Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163–183.
Wollack, J. A., & Cohen, A. S. (2004). *A model for simulating speeded test data* [Paper presentation]. Paper Presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.
Worsley, K. J. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association*, *74*, 365–367.
Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model (TOEFL Technical Report No. TR-10)*. Educational Testing Service.
Yao, Y. C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, *6*(3), 181–189.
Yu, X., & Cheng, Y. (2020). A comprehensive review and comparison of CUSUM and change-point-analysis methods to detect test speededness. *Multivariate Behavioral Research*, *57*(1), 112–133.
Zhu, H., Jiao, H., Gao, W., & Meng, X. (2023). Bayesian change-point analysis approach to detecting aberrant test-taking behavior using response times. *Journal of Educational and Behavioral Statistics*, *48*(4), 490–520.

Cite this article: Lu, J., Wang, C., Zhang, J. and Liu, Z. (2026). Detecting Test Speededness Using Responses and/or Response Times: Change Point Analysis Approaches Based on Schwarz Information Criterion. *Psychometrika*, 965–994. https://doi.org/10.1017/psy.2026.10094

https://doi.org/10.1017/psy.2026.10094 Published online by Cambridge University Press