Behavior Research Methods (2023) 55:963–980
https://doi.org/10.3758/s13428-021-01779-z

A new perspective on detecting performance decline: A change-point analysis based on Jensen-Shannon divergence

Dongbo Tu[1] · Yaling Li[1] · Yan Cai[1]

Accepted: 21 December 2021 / Published online: 6 May 2022
© The Psychonomic Society, Inc. 2022

Abstract
A common observation in ability assessment is that the probability of an examinee giving a correct response drops for end-of-test items due to low motivation, time limits or other factors. On the test-takers' side, this change can be considered performance decline (PD), which can strongly affect test validity and bias respondents' ability estimators. Currently, there is an increasing interest in the detection of PD among researchers and practitioners. Researchers and practitioners found that PD detection fails to achieve acceptable power, which is typically below 0.55. Change-point analysis (CPA), a well-developed statistical method, can be applied to item response sequences to identify whether an abrupt change exists. Existing CPA methods cannot be directly used to detect PD because they are appropriate for two-sided alternative hypotheses. To address these issues, this research firstly develops a CPA method based on Jensen-Shannon divergence to detect PD. Additionally, existing CPA statistics were converted into one-sided statistics to accommodate PD detection. Then, a simulation study was conducted to investigate the performance of the proposed method and compare it with modified CPA statistics. Results show that the proposed CPA method can detect PD with higher power while generating a well-controlled Type-I error rate. Compared against modified CPA statistics, the proposed method exhibits an augmentation in power from 1.0% to 8.2%, with average of 5.7% and higher accuracy in locating the change point. Finally, the proposed method was applied to two real datasets to demonstrate its utility.

Keywords Performance decline · Change-point analysis · Jensen-Shannon divergence · Person-fit statistics

Introduction

In the field of psychological and educational measurement, different tests have been developed to measure test-takers' latent traits. In addition to the intended latent trait being measured, many confounding factors, such as personal factors (e.g., motivation and physical condition of test-takers) and environmental factors (e.g., time limit and testing conditions), may also influence test performance. If these "nuisance" factors seriously affect test-takers' performance, failing to consider their effect would result in biased ability estimations and thus threaten test validity. Biased ability estimators may lead to incorrect interpretations of test scores and subsequent inappropriate decisions (e.g., academic admission) (Shao et al., 2015; Jin & Wang, 2014).

A primary purpose of large-scale educational assessments (e.g., the Program for International Student Assessment, or PISA) is to supply information on examinees' proficiency to policymakers. With no personal consequences, they have low stakes for examinees (Baumert & Demmrich, 2001; DeMars, 2000; Penk et al., 2014; Wise & DeMars, 2005); thus, for certain test-takers, the effort they make to answer items is likely lower than when they are exposed to high-stakes tests. (DeMars, 2000; Wise & DeMars, 2005; Wolf & Smith, 1995; Wolf, Smith & Birnbaum, 1995). Aberrant response behaviors, such as random guessing, rapid responding and omitting a mass of items, are often observed and are most salient at the end of a test (van Barneveld, 2007; Wise, 1996). In the field of psychometrical intelligence tests, similar problems also arise. In such situations, test results carry little or no meaning for the respondents themselves. Consequently, certain respondents may lose motivation or

✉ Yaling Li
lyl199681@aliyun.com

✉ Yan Cai
cy1979123@aliyun.com

[1] School of Psychology, Jiangxi Normal University, 99 Ziyang Ave, Nanchang 330022, Jiangxi, China

effort gradually as the test progresses, responding with more guesses and blanks on end-of-test items.

Test time limits also strongly affect examinees' performance on end-of-test items (Bolt, Cohen & Wollack, 2002; Glas & Pimentel, 2008; Goegebeur, De Boeck, Molenberghs & del Pino, 2006; Goegebeur, De Boeck, Wollack & Cohen, 2008). Unlike speeded tests, test time in power tests, which purports to measure and only measure cognitive ability of certain domains, should ideally be adequate to allow all respondents to try all items with maximum effort. In practice, most power tests are administered with time limits. Responses given in haste are thus frequently observed, particularly in high-stakes tests (Jin & Wang, 2014). Examinees under time pressure are inclined to respond to questions more rapidly or guess randomly on multiple-choice items and leave blanks on items that they could not reach before the end of the test (Lu & Sireci, 2007).

Such testing behaviors can lead to a decline in the probability that a test-taker answers a question correctly towards the end of a test. From the test-taker's perspective, this can be considered a performance decline (PD), which is viewed as a type of aberrant response behavior (Cao & Stokes, 2008; Schnipke & Scrams, 1997; Suh et al., 2012). PD is more likely to attribute to test speededness during high-stakes tests (Bolt et al., 2002), while PD in low-stakes tests is often associated with a decrease in motivation or effort (Wise & Kong, 2005). List et al. (2017) noted that if PD is present but not identified, measurement error increases and inference accuracy may suffer.

There are three approaches that are currently used to identify PD (Schüttpelz-Brauns et al., 2018). The first is to measure response time to items based on the assumption that respondents with less test-taking effort would take less time to complete items. Measuring response time is convenient in computer-based assessment but fails to distinguish between low test-taking effort and test-takers with high expertise, who can identify key words in items and decide in seconds whether they can answer them or not (Schüttpelz-Brauns et al., 2018). The second widely used method is the administration of self-report questionnaires after an assessment. This method does not require sophisticated statistical skills but may not yield adequate accuracy or validity (Wise & DeMars, 2005); less motivated respondents may respond more carelessly and untruthfully (Debeer et al., 2014). A third method is an appropriateness measurement, which evaluates the fit of a test-taker's response pattern to a chosen item response theory (IRT) model. Inferences are limited by model fit; thus, before conducting the person-fit test, the optimal IRT model should be identified based on the test-level model fit (Tendeiro & Meijer, 2012); if the model is misspecified, the inference may be invalid. More importantly, de la Torre and Deng (2008) used IRT person-fit statistics (l_z) to detect speeding and lack of motivation and found that they achieved limited power; the largest power was 0.125 and 0.524, respectively.

As a statistical process control (SPC) method, change-point analysis (CPA) can detect abrupt changes in a sequence of data. Recently, CPA has been used by psychometricians to detect aberrant response behaviors (Shao, 2016; Shao, Li & Cheng, 2015; Sinharay, 2016, 2017a, 2017b, 2017c; Yu & Cheng, 2019). An advantage of CPA is that it can detect aberrant response behavior and locate the change point (i.e., the item after which a respondent shows an aberrant response), which makes deletion of responses after the change point possible for data cleaning (Embretson & Reise, 2000; Shao et al., 2015). Another advantage of the CPA method is its flexibility: it does not need to know the distribution parameter before and after the change point, or fit a specific model that explicitly considers aberrant response behavior. Sinharay (2016) developed three CPA procedures to detect performance changes for computerized adaptive testing systems (CATs), which are more appropriate for the two-sided alternative hypothesis. PD can lead to a decline in the ability of the subtest after the change point; thus, those who perform worse on items after the change point are exactly what the proposed PD aims to detect. Consequently, these CPA methods are inappropriate for detecting PD. Sinharay (2016) also found that each CPA method had higher power in detecting performance change that have considerable differences in ability before and after the change point (−2 or 2). However, the success of detecting performance changes with fewer differences in ability (−1 or 1) was limited, achieving power of approximately 0.53 or even lower.

Compared to traditional outlier detection methods, CPA can estimate the change point, and this inference does not depend on the optimal IRT model. Thus, we use CPA to detect PD. In CPA, the problem in detecting aberrancy is recognizing whether performance on subtests before and after the change point changes significantly. Given the responses of a respondent, each subtest is characterized by the corresponding posterior distributions or point estimates of ability. Existing CPA methods, such as the method based on the Wald test, for detecting PD by identifying differences in the mean can be considered. In general, statistics based on the estimated moments fail to capture the difference between posterior entirely. Additionally, a difference between point estimates can be particularly unstable with an insufficient number of items in one of the subtests. For these issues, a common solution is to use Bayesian statistics and consider the respondent's ability as a distribution. Measuring the difference between posterior ability distributions of two subtests directly may be more stable and accurate. Consequently, a CPA method for detecting PD based on the Jensen-Shannon divergence is proposed in this study.

The remainder of the article is presented as follows. First, the CPA methods for PD are briefly introduced. Second, the proposed CPA method based on Jensen-Shannon divergence is introduced. Third, the performance of the proposed CPA method in detecting PD is evaluated and compared against modified CPA methods through a simulation study. Then, the proposed model is applied to two real-data examples. Finally, the strengths, limitations, and future directions of this research are discussed.

Method

Change-point analysis

For a process or variable, when a certain type of statistical property (e.g., model parameter) changes at a specific point under the influence of systematic factors, two subsequences before and after that point present different patterns. That point is considered to be the change point. As the name implies, CPA detects whether the statistical properties of a sequence change and estimates where a change occurs. CPA has been used in many fields, such as economics, statistics and medicine (e.g., Andrews, 1993; Barry & Hartigan, 1993; Robinson, Wager & Lindquist, 2010). Although it has a wide range of application in many fields, only a handful studies have applied CPA to detect aberrancy in the testing process. For example, a real-time continuous item monitoring program based on CPA was proposed to detect whether and when an item becomes compromised (Zhang, 2014). When test-takers' responses to item strings are considered to be of interest, CPA can be used to detect aberrant response behaviors within a test. Shao et al. (2015) were the first to apply CPA to individuals' item response data to detect whether and when each test-taker had speeded responses within the test process. Yu and Cheng (2019) proposed a CPA procedure based on weighted residuals to detect random responses in the context of low-stakes psychological assessment.

The CPA methods for PD

We denote the latent trait of a test to be measured (e.g., reading literacy or depression) as θ. Without a change point, it is assumed that response data that examinees provided in the order of presentation of items fit the 2-parameter logistic IRT model (2PL), one of the widely used IRT models for 0–1-scored data. The formula of 2PL is presented as follows:

$$P_i(\theta) = \frac{\exp\left[Da_i\left(\theta - b_i\right)\right]}{1 + \exp\left[Da_i\left(\theta - b_i\right)\right]}, \tag{1}$$

where $P_i(\theta)$ is the probability that the examinee with the latent trait θ correctly answered the i-th item; D is a scaling constant of 1.7; a_i and b_i are the discrimination parameter and difficulty /location parameter of item i, respectively.

With dichotomous items, Shao et al. (2015) and Sinharay (2016, 2017a, 2017b, 2017c) proposed three statistics for CPA: $L_{\max}$ based on the likelihood ratio test, $W_{\max}$ based on the Wald test, and $S_{\max}$ based on the score test. For all three statistics, their rationale was that the test can be divided into two subtests if a change in the latent trait occurs immediately after item j.

Before introducing these three statistics, we define the following notations. It is assumed that item j is the change point with J test items. And let S_1 containing item 1 to item j and S_2 including item $j+1$ to item J represent the subtest before the change point and the subtest after the change point, respectively. Let define the latent trait estimator from the scores on the entire test as $\hat{\theta}_0$, that for the scores on S_1 as θ_{1j}, and that for the scores on S_2 as $\hat{\theta}_{2j}$.

CPA statistic based on likelihood ratio test The LRT statistics (Rao, 1973) for testing the null hypothesis of equality of the respondent latent trait over S_1 and S_2 is given by:

$$L_j = -2\left[\begin{array}{c} L\left(\hat{\theta}_0; Y_1, Y_2, \ldots, Y_J\right) - L\left(\hat{\theta}_{1j}; Y_1, Y_2, \ldots, Y_j\right) \\ - L\left(\hat{\theta}_{2j}; Y_{j+1}, Y_{j+2}, \ldots, Y_J\right) \end{array}\right], \tag{2}$$

where, for example

$$L\left(\hat{\theta}_{1j}; Y_1, Y_2, \ldots, Y_j\right) = \sum_{i=1}^{j}\left[Y_i \log P_i\left(\hat{\theta}_{1j}\right) + \left(1 - Y_i\right)\log\left\{1 - P_i\left(\hat{\theta}_{1j}\right)\right\}\right], \tag{3}$$

where Y_1, Y_2, …, Y_J is a sequence of item responses, $L\left(\hat{\theta}_{1j}; Y_1, Y_2, \ldots, Y_j\right)$ is denoted as an examinee's log likelihood of Y_1, Y_2, …, Y_j at $\hat{\theta}_{1j}$.

The statistics L_j are appropriate for two-sized alternative hypotheses (i.e., L_j could test the equality of the respondent latent trait over S_1 and S_2). For PD, we intend to identify those who perform worse on S_2 and not those who perform worse on S_1 (i.e., $\hat{\theta}_{1j} \geq \hat{\theta}_{2j}$). Consequently, the alternative hypothesis in PD cases is one-sized. For one-sized alternatives, studies (Cox, 2006; Cox & Hinkley, 1974; Biehler, Holling & Doebler, 2014) have suggested the use of the signed likelihood ratio statistic, which, for PD, is given by:

$$L_{sj} = \begin{cases} \sqrt{L_j}, & if\, \hat{\theta}_{1j} \geq \hat{\theta}_{2j} \\ -\sqrt{L_j}, & if\, \hat{\theta}_{1j} < \hat{\theta}_{2j} \end{cases}. \tag{4}$$

Therefore, L_{sj} is positive if the respondent's estimated latent trait based on S_1 is greater than that based on S_2 and otherwise. (Sinharay, 2017a).

CPA statistic based on Wald test The Wald statistics (Rao, 1973) for testing the null hypothesis of equality of the respondent latent trait over S_1 and S_2 is given by:

$$W_j = \frac{\left(\hat{\theta}_{1j} - \hat{\theta}_{2j}\right)^2}{\frac{1}{I_1\left(\hat{\theta}_0\right)} + \frac{1}{I_2\left(\hat{\theta}_0\right)}}, \tag{5}$$

where $I_1\left(\hat{\theta}_0\right)$ and $I_2\left(\hat{\theta}_0\right)$ are the estimated test information based on S_1 and S_2, respectively, at $\hat{\theta}_0$. Because the alternative hypothesis in the proposed case is one-sided, it is more suitable to use the signed Wald statistics, which, for PD, is given by:

$$W_{sj} = \frac{\left(\hat{\theta}_{1j} - \hat{\theta}_{2j}\right)}{\sqrt{\frac{1}{I_1(\hat{\theta}_0)} + \frac{1}{I_2(\hat{\theta}_0)}}}. \tag{6}$$

When a respondent is affected by PD, his or her $\hat{\theta}_{1j}$ is greater than $\hat{\theta}_{2j}$, and then W_{sj} is positive. For a respondent with a non-PD aberrant response pattern (e.g., warm-up effect, the short-term effect of poor performance at the early stage of a test due to anxiety, tension), his or her $\hat{\theta}_{1j}$ is below $\hat{\theta}_{2j}$; thus, W_{sj} is negative.

CPA statistic based on the Score test The Score statistic (Rao, 1973) for testing the null hypothesis of equality of the respondent latent trait over S_1 and S_2 is given by:

$$S_j = \frac{\left[\nabla\left(\hat{\theta}_0; Y_1, Y_2, \ldots, Y_j\right)\right]^2}{I_1\left(\hat{\theta}_0\right)} + \frac{\left[\nabla\left(\hat{\theta}_0; Y_{j+1}, Y_{j+2}, \ldots, Y_J\right)\right]^2}{I_2\left(\hat{\theta}_0\right)} \tag{7}$$

where $\nabla\left(\hat{\theta}_0; Y_1, Y_2, \ldots, Y_j\right)$ and $\nabla\left(\hat{\theta}_0; Y_{j+1}, Y_{j+2}, \ldots, Y_J\right)$ are the first-order derivatives of the log likelihood of S_1 and S_2, respectively, at $\theta = \hat{\theta}_0$. For the same reason as mentioned earlier, it is modified to the signed score statistic (Cox, 2006), which for PD is given by:

$$S_{sj} = \begin{cases} \sqrt{S_j}, & if\, \hat{\theta}_{1j} \geq \hat{\theta}_{2j} \\ -\sqrt{S_j}, & if\, \hat{\theta}_{1j} < \hat{\theta}_{2j} \end{cases}. \tag{8}$$

In general, larger L_{sj}, W_{sj} and S_{sj} lead to a higher probability that the null hypothesis is incorrect, providing stronger evidence that there is a change point j in the response sequence. Sinharay (2017a) noted that L_{sj} and S_{sj} both asymptotically follow standard normal distribution; thus, L_{sj} and S_{sj} of examinee n can be compared to critical values obtained from standard normal distribution. If they are above the critical value, we can deduce that the change occurs immediately after item j in the response sequence of examinee n. However, in real practice, $J-1$ possible change points exist between item 1 and item $J-1$. Thus, all possible change points are investigated, and the maximum of all possible change points is considered the ultimate test statistic:

$$L_{\max} = \max_{1 \leq j \leq J-1} L_{sj}. \tag{9}$$

$$W_{\max} = \max_{1 \leq j \leq J-1} W_{sj}. \tag{10}$$

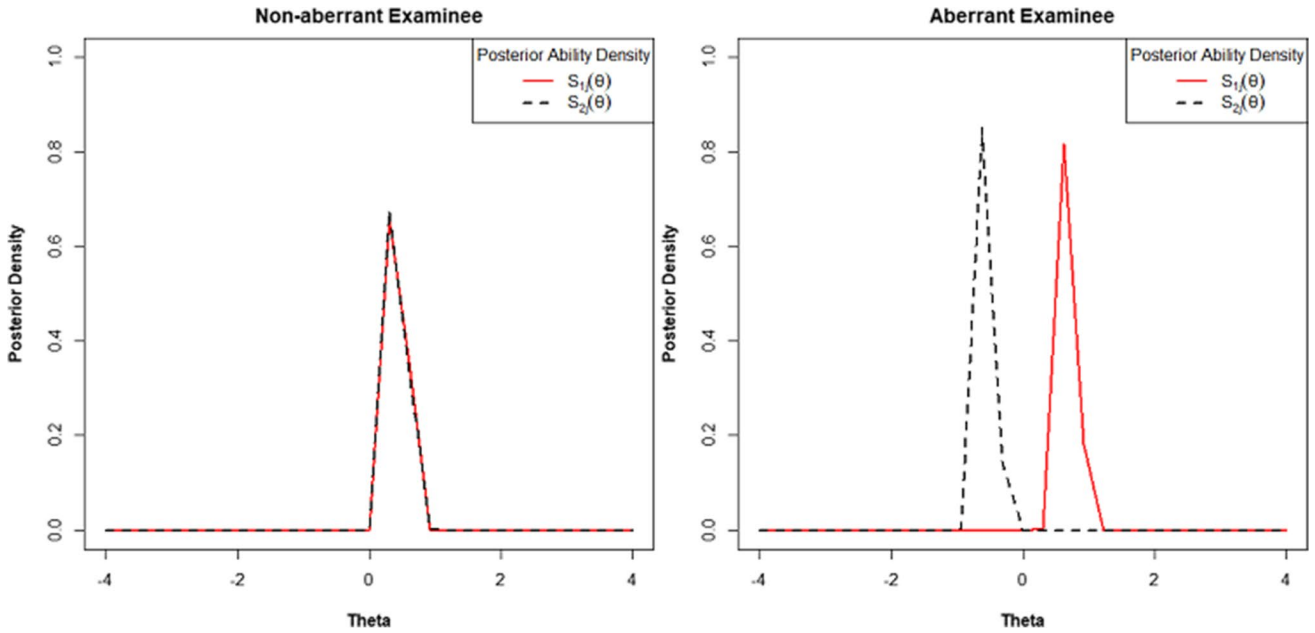
$$S_{\max} = \max_{1 \leq j \leq J-1} S_{sj}. \tag{11}$$

Despite the analytical distributions of $L_{\max}$, $W_{\max}$ and $S_{\max}$ is obtainable, considering the difficulty of calculation, or the poor approximation of the asymptotic distribution in short test lengths, we adopted the Monte Carlo simulation approach as done in Shao and Cheng (2017) and Yu and Cheng (2019) to establish the null distribution of the aforementioned three CPA statistics. One could infer whether a response sequence exists as a change point by comparing individuals' $L_{\max}$, $W_{\max}$ and $S_{\max}$ with the corresponding null distribution. If true, the point that has maximum values of $L_{\max}$, $W_{\max}$ and $S_{\max}$ is the change point estimated through CPA procedures.

Proposed CPA based on Jensen-Shannon divergence A CPA procedure based on Jensen-Shannon divergence (JS; Lin, 1991), which is called *JS*, is proposed to detect PD in this study. The *JS* is a symmetric measure of the difference between two probability distributions P and Q. In this study, *JS* is used to measure the difference between two posterior ability distributions estimated by S_1 and S_2.

To describe the rationality of *JS*, we simulated two respondents and plotted their posterior ability distributions. The responses to a 20-item test of respondent 1 without PD were simulated by the 2PL. For respondent 2 with PD, his or her responses to the first 10 items were generated similarly; however, the responses to the last 10 items were generated following the mixture performance decline model (MPDM; Jin & Wang, 2014) that is introduced in Eq. 24. The item parameters were simulated as described by Shao et al. (2015). For item i ($i = 1, 2 \ldots .20$), the difficulty parameter b_i was randomly generated from standardnormal distribution. The discrimination parameter a_i was generated randomly from *logN* (*0, 0.5*). Figure 1a shows two posterior ability distributions computed from S_1 and S_2 for respondent 1, a non-aberrant respondent. Figure 1b shows the same for respondent 2 with a change point in the middle of the response sequence. For respondent 1, the two estimated posterior ability distributions overlap considerably. However, for respondent 2, the posterior distribution based on S_1 is located far on the right side of that based on S_2.

Figure 1 shows that normal responses can generate two similar posterior distributions, while aberrant responses generate two posterior distributions that exhibit marked differences. Following this logic, the CPA based on the



(a) Responder 1 without PD

(b) Responder 2 with PD

Fig. 1 The estimated posterior ability distributions based on S_1 and S_2 for the respondent with PD and without PD. Note. $S_{1j}(\theta)$ is the posterior ability distribution based on S_1, and $S_{2j}(\theta)$ is the posterior ability distribution based on S_2.

Jensen-Shannon divergence statistic was proposed here to measure the difference between posterior ability distributions. The JS_j between two posterior ability distributions is computed by the following formula:

$$JS_j\left[S_{1j}(\theta)\middle\|S_{2j}(\theta)\right] = \frac{1}{2}\int_{-\infty}^{+\infty} S_{1j}(\theta)\log\left\{\frac{S_{1j}(\theta)}{\frac{S_{1j}(\theta)+S_{2j}(\theta)}{2}}\right\}d(\theta) + \frac{1}{2}\int_{-\infty}^{+\infty} S_{2j}(\theta)\log\left\{\frac{S_{2j}(\theta)}{\frac{S_{1j}(\theta)+S_{2j}(\theta)}{2}}\right\}d(\theta), \tag{12}$$

where $S_{1j}(\theta)$ and $S_{2j}(\theta)$ refer to the estimated posterior ability distribution based on S_1 and S_2, respectively. The values of JS are between 0 and 1, and when JS_j is equal to 0, $S_{1j}(\theta)$ and $S_{2j}(\theta)$ are identical. The larger the values of JS_j are, the greater the difference between $S_{1j}(\theta)$ and $S_{2j}(\theta)$; thus, a relatively larger JS_j indicates that the change occurs in the given response sequence.

Bayes' theorem expressed in terms of a probability density function is stated as:

$$f(\theta|X) = \frac{f(X|\theta)f(\theta)}{f(X)} = \frac{f(X|\theta)f(\theta)}{\int f(X|\theta)f(\theta)d\theta}, \tag{13}$$

where $f(\theta|X)$ is the posterior distribution for parameter θ, $f(X|\theta)$ is the sampling density for the data X, and $f(\theta)$ is the prior probability of θ. $f(X)$ refers to the marginal probability of the data X. When fitting Eq. 13 to the IRT, the $f(X|\theta)$ is expressed as the relative likelihood of the item response data given all of the model parameters. To simplify the calculation, a finite set $m = \{\theta_1,\theta_2,\ldots,\theta_l\}$ of ability values equally spaced in the interval $[-4, 4]$ was used to approximate the numerical value of Eq. 13, where $l = 27$. In the standard Bayesian method, prior information is fixed before response data are collected. The prior probability is obtained from the data within the empirical Bayesian method, obtaining more information for a parameter (Robbins, 1985). To accurately estimate the posterior ability distributions, the current study adopt this method. An initial standard normal prior is used. The prior for θ_{1j} and θ_{2j} is given in the following form:

$$X(\theta_m) = \frac{\sum_{i=1}^{J}\left[P(Y_i|\theta_m)W(\theta_m)\right]}{\sum_{k=1}^{l}\sum_{i=1}^{J}\left[P(Y_i|\theta_k)W(\theta_k)\right]}, \tag{14}$$

where $W(\theta_k)$ refers to the weight of θ_k, obtained from $N(0,1)$, and m is the finite set for ability quadrature points. Once the prior is obtained, Bayesian posteriors are computed based on the response data. The formula for the posterior probabilities for S_1 is given as follows:

$$S_{1j}(\theta_m) = \frac{\sum_{i=1}^{j}\left[P(Y_i|\theta_m)X(\theta_m)\right]}{\sum_{k=1}^{l}\sum_{i=1}^{j}\left[P(Y_i|\theta_k)X(\theta_k)\right]}, m = 1,\ldots,l, \tag{15}$$

where $S_{1j}(\theta_m)$ refers to the posterior probability of the quadrature points θ_m, and $X(\theta_m)$ is the prior calculated from Eq. 14. Similarly, the posterior distribution based on S_2 is:

$$S_{2j}(\theta_m) = \frac{\sum_{i=j+1}^{J}\left[P(Y_i|\theta_m)X(\theta_m)\right]}{\sum_{k=1}^{l}\sum_{i=j+1}^{J}\left[P(Y_i|\theta_k)X(\theta_k)\right]}, m = 1, \ldots, l. \tag{16}$$

There is one issue, however, when JS_j is used directly to detect PD. Individuals whose posterior ability distribution based on S_2 is located far to the right side of that based on S_1 might be flagged by JS_j. However, they do not experience PD and are not the objects that we aim to detect. Fortunately, this issue can be solved by fixing the JS_j of someone who outperforms S_2 on S_1 to zero (i.e., only those whose performance on S_1 is better than that on S_2 might be flagged by JS_j). Finally, JS_j between $S_{1j}(\theta)$ and $S_{2j}(\theta)$ is calculated by the following equation:

$$JS_j\left[S_{1j}(\theta)\middle\|S_{2j}(\theta)\right] = \begin{cases} \frac{1}{2}\sum_{m=1}^{l} S_{1j}(\theta_m)\log\left\{\frac{S_{1j}(\theta_m)}{\frac{S_{1j}(\theta_m)+S_{2j}(\theta_m)}{2}}\right\} + \frac{1}{2}\sum_{m=1}^{l} S_{2j}(\theta_m)\log\left\{\frac{S_{2j}(\theta_m)}{\frac{S_{1j}(\theta_m)+S_{2j}(\theta_m)}{2}}\right\} & if\hat{\theta}_{1j} \geq \hat{\theta}_{2j} \\ 0, & if\hat{\theta}_{1j} < \hat{\theta}_{2j} \end{cases} \tag{17}$$

Thus, $JS_j[S_{1j}(\theta)\|S_{2j}(\theta)]$ is equal to 0 for those who outperform S_2 on S_1. Because both posterior distributions in Eq. 17 are estimable, the values of JS_j could provide an index of similarity or difference for $S_{1j}(\theta)$ and $S_{2j}(\theta)$. Thus, test-takers with large values of JS_j might experience PD.

In fact, the actual change point is unknown, so all possible change points would be tested. The point with the maximum JS_j value is the change point estimated by the CPA procedure as:

$$JS_{\max} = \max_{1\leq j\leq J-1} JS_j. \tag{18}$$

This step is similar to the other aforementioned CPA statistics. As with the aforementioned three CPA statistics, the null distribution for $JS_{\max}$ is also obtained using the Monte Carlo method. The details of the simulation are shown in the following section. Once the null distributions are obtained, sample statistics can be compared to the critical values for all four CPA statistics to detect whether PD occurs, given a significance level.

Simulation study

A simulation study was conducted to investigate the performance of the proposed CPA procedure and three other modified CPA procedures. Normal response patterns were simulated using the 2PL model. Response patterns with PD were simulated using MPDM (Jin & Wang, 2014). The performance of the proposed method and three modified CPA methods was evaluated in two aspects. First, the power (the proportion of respondents with PD who are successfully detected) and the Type-I error rate (the proportion of normal respondents who are falsely specified as PD) were calculated. Second, the accuracy of four CPA statistics in locating the change point was evaluated. The difference between the estimated change point and true change point is denoted as lag, which is calculated in two ways. For respondents affected by PD who are successfully detected, the lag is the difference between the estimated change point and true change point. For respondents affected by PD who are incorrectly labeled as without PD, the lag is the difference between the length of the test and the true change point, in which the CPA statistic considers that there is no change point in their response sequence. Since the lag can be positive or negative, and can be offset if the average is taken, the absolute value of the lag was used and then mean was calculated.

Simulating response data with PD

Jin and Wang (2014) proposed a mixture IRT model for PD. The assumption of the MPDM is that examinees exert their utmost effort to attempt items until a certain item and then start to attempt items with less effort, which is consistent with the premise of CPA. Thus, the MPMD was adopted to simulate response with PD in this study.

The MPDM takes the following form:

$$P(Y_i = 1) = c_i + \frac{(1-c_i)\exp\left[a_i(\theta - b_i - \omega_i)\right]}{\exp\left[a_i(\theta - b_i - \omega_i)\right] + 1}, \tag{19}$$

where c_i is the guessing parameter of item i; $\omega_i(\omega_i \geq 0)$ is the attenuation parameter at item position i and is used to adjust the PD due to a decline in test-taking effort, speededness, or any factor.

$$\omega_i \begin{cases} 0, & if\ i \leq \delta \\ \gamma_\delta, & if\ i > \delta \end{cases}, \tag{20}$$

where δ is the change point, an integer with a value ranging from $[1, J]$. When $\delta = i$, PD will start after item i. If $\delta = J$, PD will not occur throughout the test; $\gamma_\delta(\gamma_\delta \geq 0)$ is the decrement when the change point is δ:

$$\gamma_\delta = k(J-\delta), \tag{21}$$

$$\gamma_\delta = k_1(J-\delta) + k_2(J-\delta)^2 \tag{22}$$

where k ($k>0$) is the slope of the line formed by connecting the decrement of change points. Jin and Wang (2014) also proposed a quadratic function for γ_δ. However, they found that the linear function for γ_δ (Eq. 21) fits empirical data well and the value of k_2 approaches 0, which indicates that the quadratic term is not essential.

Thus, the new MPDM for 2PL takes the following formula:

$$P_i(\theta) = \begin{cases} \frac{\exp\left[Da_i(\theta-b_i)\right]}{1+\exp\left[Da_i(\theta-b_i)\right]}, & i \le \delta \\ \frac{\exp\left[Da_i\{\theta-b_i-k(J-\delta)\}\right]}{1+\exp\left[Da_i\{\theta-b_i-k(J-\delta)\}\right]}, & i > \delta \end{cases}, \tag{23}$$

The following example helps interpret the new MPDP for 2PL. Let there be a four-item test ($J=4$) and $k=0.1$. Therefore, respondents consist of four groups, namely $\delta=1$, $\delta=2$, $\delta=3$ and $\delta=4$. According to Eq. 23, the decrements are $\gamma_1=k(J-\delta)=0.3$, $\gamma_2=k(J-\delta)=0.2$, $\gamma_3=k(J-\delta)=0.1$ and $\gamma_4=k(J-\delta)=0$ for the four groups, respectively. If a respondent is categorized into group 1, θ is engaged in item 1, while $\theta-0.3$ is engaged in items 2 to 4; if categorized into group 2, θ is engaged in items 1 to 2, while $\theta-0.2$ is engaged in items 3 to 4; if categorized into group 3, θ is engaged in items 1 to 3, while $\theta-0.1$ is engaged in item 4; if categorized into group 4, θ is engaged in items 1 to 4. In summary, the closer that the location of the change point is to the end of the test, the greater the ability decrement.

In Jin et al.'s simulation study, k was set to 0.1 and 0.2 when PD occurred. In a long test (e.g., 40-item), suppose k is set to 0.2 and a respondent experiences PD after the fifth item. Based on Eq. 23, γ_δ is equal to 7, while the difference between the upper (3) and lower (−3) bounds of θ is usually 6. Therefore, we set k to 0.1. Jin and Wang (2014) used a complex method to simulate the respondent change points. For simplicity, a method that is similar to those used by Wollack and Cohen (2004), Shao et al. (2015) and Yu and Cheng (2019) was used to simulate the change point in this study. The change point of examinee n is thus assumed to be at $100\eta_n\%(0<\eta_n<1)$ of a test, which indicates that for item i, if $\frac{i}{J} \le \eta_n$, $\omega_{ni}=0$; otherwise $\omega_{ni}=k(J-\delta_n)$. We can express this fact with the following formula:

$$P_{ni}(\theta) = \begin{cases} \frac{\exp\left[Da_i(\theta_n-b_i)\right]}{\exp\left[Da_i(\theta_n-b_i)\right]+1}, & if\ \frac{i}{J} \le \eta_n \\ \frac{\exp\left[Da_i\{\theta_n-b_i-k(J-\delta_n)\}\right]}{\exp\left[Da_i\{\theta_n-b_i-k(J-\delta_n)\}\right]+1}, & if\ \frac{i}{J} > \eta_n \end{cases}. \tag{24}$$

The values of η_n for examinees are different. Finally, Eq. 24 was used to generate response patterns with PD.

Simulation design

Two tests of different lengths (40 and 60 items) were included in the simulation study. Item parameters were simulated in the same way as in Shao et al. (2015). Threshold parameters were generated randomly from N (0, 1). Discrimination parameters were generated from $logN$ (0, 0.5). Then, 1,000 responders whose true abilities were generated from N (0, 1) were simulated. To retain more information and keep consistent with previous studies (Shao et al., 2015; Sinharay, 2016; Yu & Cheng, 2019), different levels of prevalence of PD were considered. However, it should be noted that the PD prevalence should not affect person-level detection, because the true item parameters were used here. List et al. (2017) found that the percentage of respondents affected by PD for three mixture PD models was 9%, 18% and 32% in empirical research; thus, three levels of prevalence were simulated in this study by $m=10\%$, 20% and 30%.

Examinees with PD may be affected to varying degrees; some may have more responses affected by PD than others. To simulate different PD severity levels, the method used by Shao et al. (2015) was used to generate η, which refers to the change point and follows a beta distribution. We generated four beta distributions with different medians and variances to depict different severities of PD. Figure 2 shows the density curve of η and indicates that as the median increases, the change point moves closer to the end of the test. Also, as the variance increases, the change point between respondents exhibits more variability. Two levels of the median (0.5 and 0.6) and the variation (0.001 and 0.01) of η were coupled, resulting in four conditions.

An overview of the data generation scheme is shown in Table 1. There are 2 (test lengths) × 3 (PD prevalence)× 4 (PD severities) = 24 simulation conditions, and each condition was replicated 50 times.

To determine whether PD occurs in a given examination, sample statistics were computed and compared to respective critical values. The same methods used by Yu and Cheng (2019) and Worsley (1979) were used to obtain the critical values after simulating response data of 10,000 normal examinees (no PD) with test lengths of 40 and 60 items. The true abilities of the examinees were generated from N (0, 1). A total of 10,000 sample values of $JS_{\max}$, $L_{\max}$, $W_{\max}$ and $S_{\max}$ constitute the null distributions of each test statistic at each test length. The critical value was obtained at the cutoff of the 95th quantile. Sinharay (2017a) and Shao et al. (2015) used maximum likelihood estimation (MLE) to estimate the latent traits. When estimating a change point that is late or early, which indicates that one subtest may have an insufficient number of items, the MLE can be unstable for all 1 or 0 responses. Similar to Yu and Cheng (2019), the expected a posteriori (EAP) with a prior of N (0,1) was used to estimate the latent trait. Figure 3 shows a null distribution of four

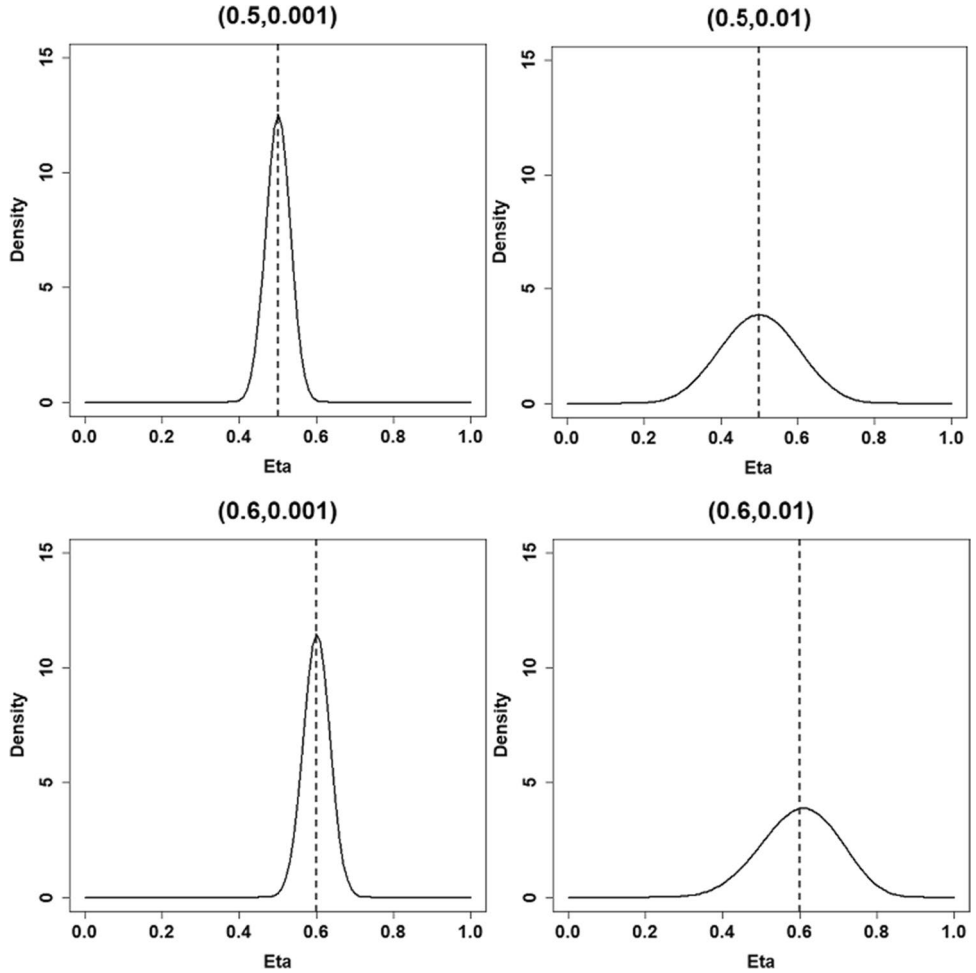


Fig. 2 Density curve of the 4 η distributions. *Note*. 0.5 is median of density curve, 0.001 is variance of density curve.

Table 1 Summary of simulation conditions

Type	Parameter	Condition	Median	Variance
MPDP parameters	η(Beta distribution)	C1	0.5	0.001
		C2	0.5	0.010
		C3	0.6	0.001
		C4	0.6	0.010
Item parameters	a	Lognormal (0, 0.5)		
	b	Normal (0, 1)		
Latent trait	θ	Normal (0, 1)		

CPA statistics, and the null distributions of JS_{max}, L_{max}, W_{max} and S_{max} are positively skewed.

The critical value of the four CPA statistics is obtained from the top 500 values, given the 10,000 sample values for JS_{max}, L_{max}, W_{max} and S_{max}, and a significance level of 0.05. After 50 replications at each test length condition, the average of 50 times was considered to be the final critical value for each CPA statistic. Table 2 reports the critical value for the four CPA statistics, along with the standard deviations (SD), across the 50 replications. Table 2 shows that as the length of the test increases, the critical value of each test statistic increases. Additionally, the standard deviation of the critical value for JS_{max}, L_{max}, W_{max} and S_{max} is small, indicating that the critical value is rather stable. For each sample value of JS_{max}, L_{max}, W_{max} and S_{max}, if it is above the respective critical value of the corresponding test length,

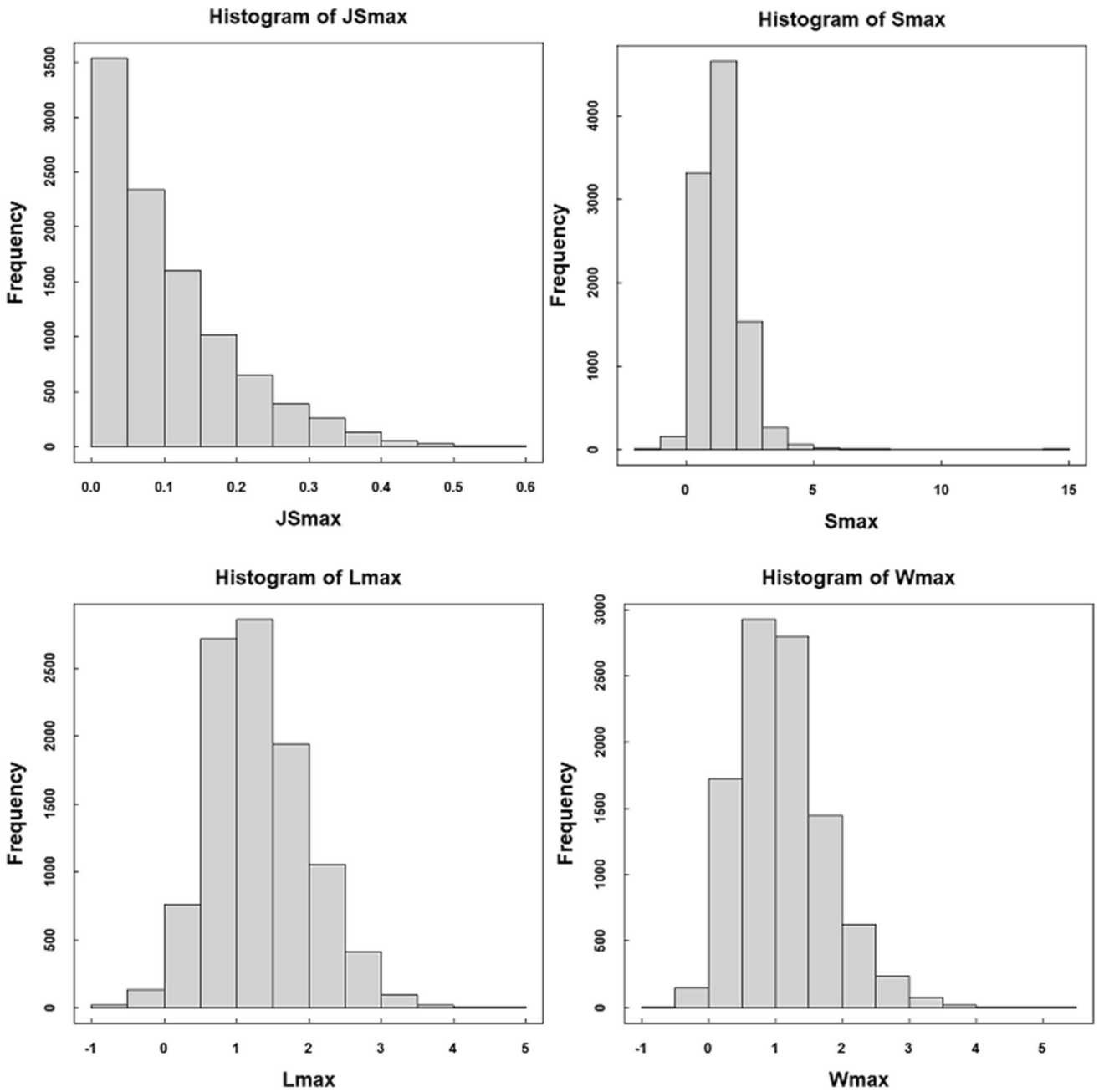


Fig. 3 Histogram of four CPA statistics.

Table 2 The average (and SD) of critical values for the four CPA statistics

Statistic	Test length	
	40	60
L_{max}	2.496 (0.016)	2.600 (0.014)
W_{max}	2.293 (0.018)	2.669 (0.018)
S_{max}	2.753 (0.026)	2.922 (0.035)
JS_{max}	0.289 (0.003)	0.311 (0.002)

then the null hypothesis is rejected, a personal misfit inference is concluded, and the estimated change point by the CPA procedures is the value of j, where $L_{\max} = L_{sj}$, $W_{\max} = W_{sj}$, $S_{\max} = S_{sj}$ $JS_{\max} = JS_j$.

Results

Table 3 shows the Type-I error rate and power averaged over 50 replications of the proposed statistic under each condition. The Type-I error rate under all experimental conditions is approximately 0.05, which implies that the Jensen-Shannon divergence-based CPA method can generate a well-controlled Type-I error rate when detecting PD under various conditions. For a 40-item test, the power ranges between 0.730 and 0.849, regardless of the severity of PD. For a 60-item test, the power is between the low .90s and the middle .90s. Thus, longer tests typically result in higher power compared to shorter tests.

With a decline in the median of η (i.e., more PD responses are present), power increases, as expected. Thus, conditions C1–C2 could generate higher power than conditions C3–C4. To understand this trend, we provide an analogy between a greater number of PD responses and a larger effect size. Larger effect sizes are more likely to be discovered and more likely to be detected statistically. When the variance of η declines—that is, when the starting point of PD has small variabilities—the power also increases. This increase occurs because it has more difficulty in correctly detecting PD for respondents with few responses (e.g., 10 or 5% of the items) affected by PD, or with many items affected by PD (e.g., 80% of the items, in which case respondents are more likely to be miscategorized as low-ability examinees). As the variance of η declines, change points become more concentrated at 50% or 60% of the test, which implies that fewer respondents have change points at the beginning or end of the test. Thus, a greater number of respondents with PD are diagnosed correctly.

By comparing the three PD prevalence results, we found that conditions with 10%, 20% and 30% PD respondents generated similar power, which implies that the PD prevalence has little effect on power.

Results based on L_{max} are shown in Table 4, which reveal patterns similar to those shown in Table 3. By comparing Tables 3 and 4, JS_{max} is shown to perform better than L_{max} in detecting PD with the same test length. When the test length is 40 items, the power of the former is approximately 2.2–3.9% higher. When the test length increases to 60, test-takers' responses affected by PD increase to 24 or 30 items, so that less sensitive CPA methods could also successfully detect them. Consequently, the gap in power between the methods is narrowing. However, JS_{max} remains higher than the latter by 1.0–2.1%. Both JS_{max} and L_{max} resulted in Type-I error rates near 0.05. Therefore, JS_{max} is generally preferable for detecting PD.

The results based on the Wald test statistics W_{max} and Score test statistic S_{max} are shown in Tables 5 and 6, respectively. Both statistics generated a well-controlled Type-I error rate, but lower powers compared to the proposed method JS_{max}; S_{max} generated the lowest power.

To facilitate comparison of the performance of the four statistics, Table 7 summarizes the average power and

Table 3 Power and Type-I error rates for PD detection based on the proposed JS_{max}

Prevalence	Severity	40		60	
		Power	TIE	Power	TIE
10%	C1	0.845	0.052	0.957	0.052
	C2	0.821	0.049	0.952	0.053
	C3	0.795	0.052	0.952	0.052
	C4	0.730	0.049	0.912	0.052
	Average	0.798	0.051	0.943	0.052
20%	C1	0.849	0.049	0.960	0.052
	C2	0.812	0.050	0.952	0.050
	C3	0.798	0.050	0.948	0.051
	C4	0.741	0.050	0.911	0.053
	Average	0.800	0.050	0.943	0.052
30%	C1	0.845	0.051	0.962	0.050
	C2	0.815	0.050	0.951	0.050
	C3	0.790	0.050	0.947	0.051
	C4	0.731	0.048	0.912	0.052
	Average	0.795	0.050	0.943	0.051

Note. TIE refers to Type-I error; C1–C4 are the four conditions of PD severity, where C1 refers to PD starting position with median = 0.5, variance = 0.001.

Table 4 Power and Type-I error rates for PD detection based on L_{max}

Prevalence	Severity	40		60	
		Power	TIE	Power	TIE
10%	C1	0.822	0.051	0.939	0.051
	C2	0.796	0.049	0.931	0.052
	C3	0.756	0.051	0.936	0.052
	C4	0.704	0.050	0.900	0.052
	Average	0.770	0.050	0.927	0.052
20%	C1	0.824	0.048	0.945	0.050
	C2	0.790	0.049	0.936	0.051
	C3	0.764	0.050	0.930	0.050
	C4	0.712	0.050	0.901	0.052
	Average	0.773	0.049	0.928	0.051
30%	C1	0.818	0.051	0.948	0.051
	C2	0.789	0.051	0.935	0.051
	C3	0.757	0.051	0.931	0.050
	C4	0.706	0.048	0.902	0.051
	Average	0.768	0.050	0.929	0.051

Note. TIE refers to Type-I error; C1–C4 are the four conditions of PD severity, where C1 refers to PD starting position with median = 0.5, variance = 0.001.

Table 5 Power and Type-I error rates for PD detection based on $W_{\max}$

Prevalence	Severity	40		60	
		Power	TIE	Power	TIE
10%	C1	0.811	0.051	0.922	0.049
	C2	0.780	0.052	0.913	0.052
	C3	0.743	0.051	0.919	0.050
	C4	0.676	0.050	0.882	0.051
	Average	0.753	0.051	0.909	0.051
20%	C1	0.813	0.049	0.928	0.049
	C2	0.769	0.050	0.916	0.051
	C3	0.750	0.050	0.910	0.050
	C4	0.683	0.049	0.882	0.052
	Average	0.754	0.050	0.909	0.051
30%	C1	0.807	0.051	0.932	0.049
	C2	0.774	0.052	0.917	0.052
	C3	0.741	0.050	0.913	0.050
	C4	0.679	0.048	0.885	0.052
	Average	0.750	0.050	0.912	0.051

Note. TIE refers to Type-I error; C1–C4 are the four conditions of PD severity, where C1 refers to PD starting position with median = 0.5, variance = 0.001.

Table 6 Power and Type-I error rates for PD detection based on $S_{\max}$

Prevalence	Severity	40		60	
		Power	TIE	Power	TIE
10%	C1	0.789	0.051	0.921	0.051
	C2	0.773	0.052	0.905	0.051
	C3	0.713	0.051	0.910	0.050
	C4	0.670	0.051	0.861	0.050
	Average	0.736	0.051	0.899	0.051
20%	C1	0.794	0.049	0.927	0.051
	C2	0.759	0.049	0.911	0.049
	C3	0.723	0.050	0.902	0.050
	C4	0.676	0.051	0.859	0.051
	Average	0.738	0.050	0.899	0.050
30%	C1	0.790	0.050	0.930	0.050
	C2	0.761	0.049	0.914	0.053
	C3	0.717	0.050	0.905	0.050
	C4	0.672	0.050	0.865	0.051
	Average	0.735	0.050	0.904	0.051

Note. TIE refers to Type-I error; C1–C4 are the four conditions of PD severity, where C1 refers to PD starting position with median = 0.5, variance = 0.001.

average Type-I error rate for the four CPA procedures. We can conclude that (1) as the test length increases, the power of all four statistics increases; (2) all four statistics generate a well-controlled Type-I error rate in detecting PD under various conditions; and (3) based on power, $JS_{\max}$ performs best in detecting PD, followed by $L_{\max}$, $W_{\max}$ and $S_{\max}$. It results in power typically ranging from the low .70s to the middle .90s. Compared to $L_{\max}$, $W_{\max}$ and $S_{\max}$, $JS_{\max}$ resulted in comparable Type-I error and an increase in power of between 1.0% and 8.2%.

Table 7 Average power and Type-I error for PD detection based on four CPA statistics

Index	Prevalence	Statistic	Test length	
			40	60
Average power	10%	JS_{max}	0.798	0.943
		L_{max}	0.770	0.927
		W_{max}	0.753	0.909
		S_{max}	0.736	0.899
	20%	JS_{max}	0.800	0.943
		L_{max}	0.773	0.928
		W_{max}	0.754	0.909
		S_{max}	0.738	0.899
	30%	JS_{max}	0.795	0.943
		L_{max}	0.768	0.929
		W_{max}	0.750	0.912
		S_{max}	0.735	0.904
Average Type-I error	10%	JS_{max}	0.051	0.052
		L_{max}	0.050	0.052
		W_{max}	0.051	0.051
		S_{max}	0.051	0.051
	20%	JS_{max}	0.050	0.052
		L_{max}	0.049	0.051
		W_{max}	0.050	0.051
		S_{max}	0.050	0.050
	30%	JS_{max}	0.050	0.051
		L_{max}	0.050	0.051
		W_{max}	0.050	0.051
		S_{max}	0.050	0.051

Table 8 presents the information for the mean absolute lag. Comparing the 40-item and 60-item conditions, one can find that the latter has a relatively small mean lag due to the decreased number of respondents affected by PD who are incorrectly labeled as without PD. In addition, as the variance of η declines, the mean lag also declines. In the 40-item test, there is a clear pattern that the mean lag increases with the median η. In contrast, this pattern is not observed in the 60-item test. Comparing the conditions with 10%, 20% and 30% respondents with PD, there is little difference in the mean of the absolute lag. Similar to the results for the power, $JS_{\max}$ has the best accuracy in locating the change point, followed by $L_{\max}$, then $W_{\max}$ and $S_{\max}$ has the worst performance.

Table 8 Absolute lag of change-point detection for four CPA statistics

Prevalence	Severity	40				60			
		JS_{max}	L_{max}	W_{max}	S_{max}	JS_{max}	L_{max}	W_{max}	S_{max}
10%	C1	5.523	6.262	6.395	8.247	3.352	3.920	4.517	5.217
	C2	6.153	6.920	7.086	8.433	3.805	4.522	5.286	6.104
	C3	6.782	7.760	7.825	9.487	3.715	3.979	4.614	5.549
	C4	7.592	8.388	8.641	9.960	4.912	5.055	5.866	6.659
	Average	6.513	7.333	7.487	9.032	3.946	4.369	5.071	5.882
20%	C1	5.436	6.205	6.356	7.998	3.333	3.938	4.805	5.306
	C2	6.199	7.044	7.254	8.749	3.762	4.340	5.175	5.927
	C3	6.561	7.605	7.545	9.443	3.884	4.200	4.918	5.770
	C4	7.264	8.116	8.440	10.014	5.016	5.248	5.947	6.688
	Average	6.365	7.243	7.399	9.051	3.999	4.432	5.211	5.923
30%	C1	5.508	6.374	6.484	8.085	3.309	3.842	4.484	5.144
	C2	6.170	6.908	7.139	8.557	3.796	4.345	5.086	5.801
	C3	6.745	7.775	7.803	9.545	3.907	4.228	4.851	5.718
	C4	7.482	8.291	8.543	12.911	4.949	4.988	5.712	6.559
	Average	6.476	7.337	7.492	9.775	3.990	4.351	5.033	5.806

Note. C1–C4 are the four conditions of PD severity, where C1 refers to PD starting position with median = 0.5, variance = 0.001.

Real data application

To evaluate the utility of the proposed CPA procedures, we applied CPA to two empirical datasets, one from the PISA data and the other from the Raven Advanced Progressive Matrices test (see "Real data application 1" and "Real data application 2," respectively). This section is used to demonstrate the proposed model; thus, we suggest caution in over-interpreting the results.

Real data application 1: Detection of PD in PISA

The PISA, which is an age-based survey designed to assess the performance of 15-year-old students in three primary fields of mathematics, reading and science, is typically a low-stakes testing program (List et al., 2017); thus, certain examinees may not apply their full effort throughout the test. Additionally, examinees are required to complete the test within a certain time, which might lead to speeded responses. As a result, PD was expected in these test results. We used data from the sixth booklet of PISA 2009, which exclusively covers reading items. Fifty-eight items were dichotomous, and one polytomous item was excluded from the analysis.

We used the listwise deletion method (Adams & Wu, 2002) to remove 6,202 examinees with missing data, leaving 17,101 examinees, from which we randomly sampled 5,000 respondents for further analysis. R and Mplus were used to perform all analyses in this section. First, to explore the structure of the data, we randomly divided it into two subsamples: one for exploratory factor analysis (EFA), and the other for confirmatory factor analysis (CFA). In EFA, the fitting indices of the one-factor model were as follows: CFI = 0.974, TLI = 0.973 and RMSEA = 0.021; those of the two-factor model were CFI = 0.984, TLI = 0.982 and RMSEA = 0.017. Given that the fitting indices of the one-factor model are above the critical value (0.9), and for simplicity, we conducted one-factor CFA based on the second subsample, where CFI = 0.973, TLI = 0.973 and RMSEA = 0.022. These results show that the data fit well with the one-factor model. Second, CPA procedures were used to detect PD with estimated item and ability parameters. Specifically, the item parameters of 58 items were estimated using marginal maximum likelihood estimation (MMLE) in the MIRT package (Chalmers, 2012). Based on these item parameter estimators, the ability estimation of each test-taker was obtained through EAP estimation. Given the item parameter and ability estimations, CPA statistics were computed for each test-taker.

A total of 314 examinees were flagged by JS_{max} as exhibiting PD. Of the 314 JS_{max} flagged cases, 251 were also flagged by L_{max}, 260 were also flagged by W_{max}, and 181 were also flagged by S_{max}. Figure 4 plots the sample statistic values for JS_{max} and L_{max}. The red line perpendicular to the x-axis is the critical value of JS_{max}, while the gray line perpendicular to the y-axis is the critical value of L_{max}. Hence, the dots on the lower left indicate respondents who are labeled as without PD by both statistics, and those in the upper right refer to respondents who are calibrated as with PD by both procedures. Fleiss kappa coefficient is suitable

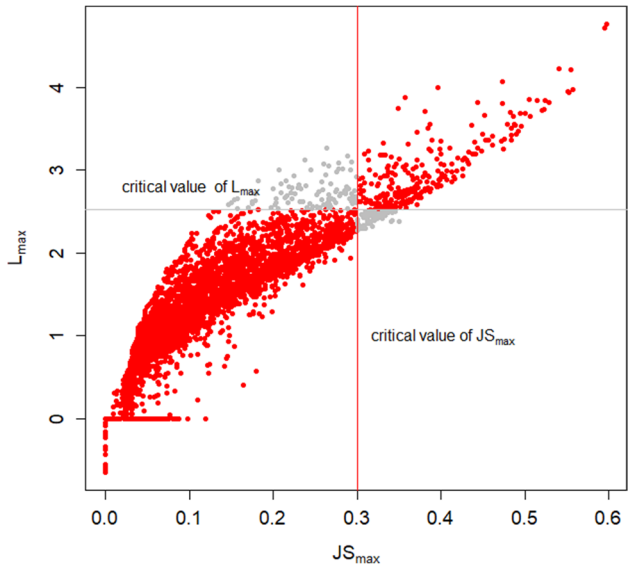


Fig. 4 The sample statistic values for JS_{max} and L_{max}. *Note*. The red dots in the figure indicate the same determination of the respondent for JS_{max} and L_{max}, while the gray dots indicate different determinations for the respondent.

for the consistency test of the analysis when repeated three or more times. Because there are four CPA methods, Fleiss kappa coefficient was calculated to evaluate the consistency of the PD detection results by the four methods. The Fleiss kappa coefficient was 0.671 ($P < .001$). A Fleiss kappa coefficient between 0.61 and 0.80 indicates that the detection results of multiple analyses are highly consistent (An et al., 2020); thus, the detection results regarding PD by the four CPA methods have high consistency in the PISA dataset.

Figures 5 and 6 compare the posterior ability distributions for two flagged respondents and two normal respondents as identified by the JS_{max} method, respectively. With regard to flagged respondents, we found that the posterior distributions based on S_1 are located far to the right of those based on S_2. For normal respondents, the posterior distributions based on S_1 and S_2 overlap considerably. These results imply that the proposed method JS_{max} can identify respondents affected by PD in a real dataset.

Real data application 2: Detection of PD in Raven's Advanced Progressive Matrices test

Raven's Advanced Progressive Matrices test (APM) is a psychological assessment that measures inductive reasoning and analogical ability. The results carry little or no meaning for the respondents themselves. Therefore, certain respondents might gradually lose motivation as the test progresses. Additionally, this test is administered with time constraints; thus, certain test-takers might respond rapidly to end-of-test items. PD was therefore expected.

We recruited a total of 1,008 students from 10 Chinese colleges. After removing 111 respondents with missing data, 897 respondents (61.3% female) were included in the analysis. First, many studies have confirmed that APM has a

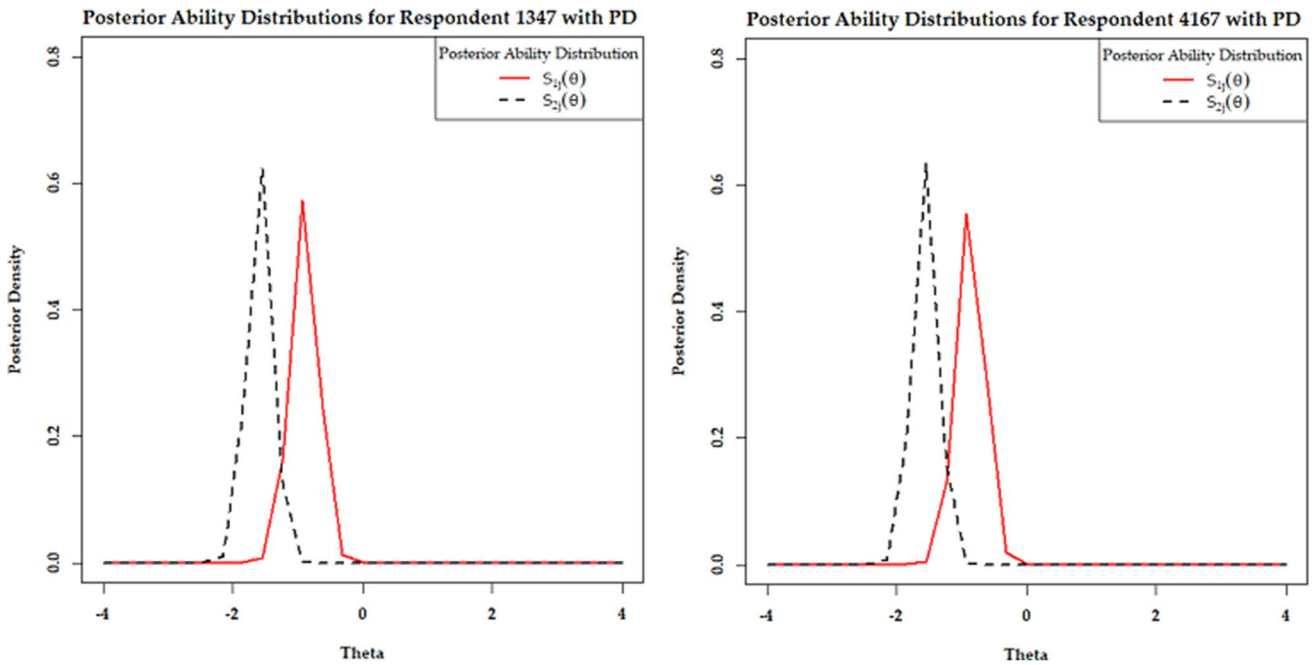


Fig. 5 Posterior distributions for two test-takers detected as having PD by JS_{max}.

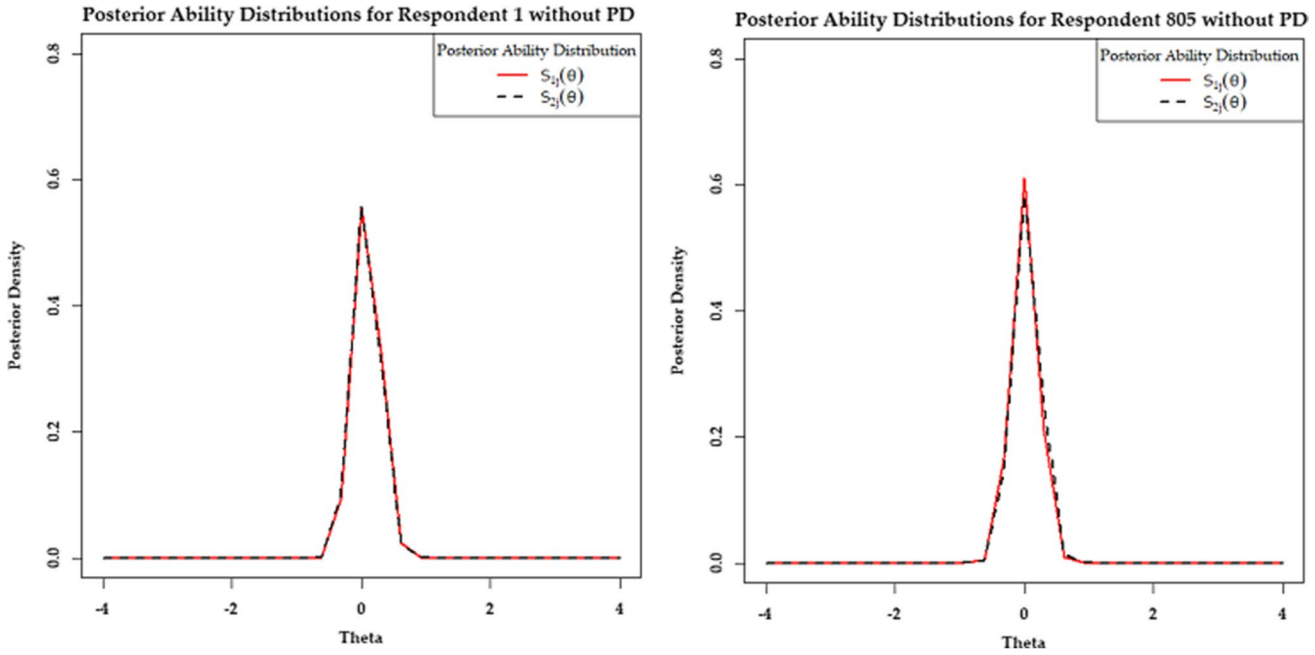


Fig. 6 Posterior distributions for two test-takers detected as not having PD by $JS_{\max}$.

unidimensional structure; therefore, we conducted a one-factor CFA model on the data, where CFI = 0.895, TLI = 0.888 and RMSEA = 0.034. The results clearly showed that the data share one common factor. Second, we computed a sample statistic with estimated item and ability parameters. Specifically, we used the same method as in "Real data application 1" to obtain item and ability parameters.

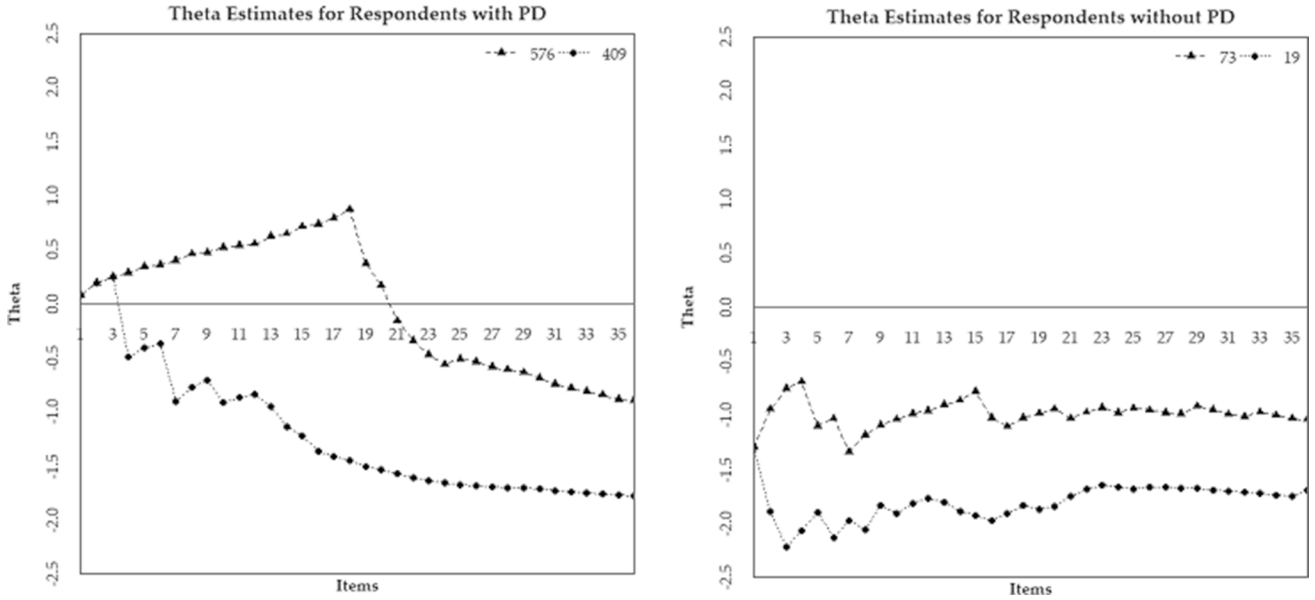
A total of 98 examinees were flagged as PD by $JS_{\max}$. Of the 98 $JS_{\max}$ flagged cases, 85 were also flagged by $L_{\max}$, 77 were also flagged by $W_{\max}$, and 55 were also flagged by $S_{\max}$. Similarly, the Fleiss kappa coefficient was calculated. The Fleiss kappa coefficient was 0.721 ($P < .001$), which again confirms that the PD detection results by the four CPA methods in real data are highly consistent.

Figure 7a and b compare the instantaneous ability estimators of the two flagged subjects and two normal subjects by the $JS_{\max}$ method, respectively. By observing the instant ability estimators for flagged subjects, we discovered that most instant ability estimators fluctuated considerably and that there was a clear downward trend as the test progressed. For normal respondents, their instant ability estimators fluctuated marginally at the beginning of the assessment, which may have occurred because few items were responded to at the beginning of the test, resulting in unstable instant ability estimates. As the number of items answered increased, their instant ability estimator tended to stabilize, which again verified that the proposed method $JS_{\max}$ can identify subjects affected by PD in a real dataset.

Discussion

Given that the traditional approaches of identifying PD have their respective flaws and that existing CPA methods are inappropriate for PD detection, this study first modified three existing CPA statistics to accommodate PD detection. Then, we proposed the Jensen-Shannon divergence-based CPA method, investigated its performance and compared it with modified CPA methods through a simulation study, and finally elaborated its effectiveness in two real datasets. Results show that the power and accuracy in locating the change point of the Jensen- Shannon divergence-based CPA statistic was superior to that of the three modified CPA statistics, while retaining a Type-I error rate near the nominal level. Two empirical studies also show that the statistic is capable of identifying respondents whose response pattern is affected by PD, and the four CPA methods for detecting PD have high consistency.

The primary advantages of the proposed method and the contributions of this article include the following: (1) The proposed method and three modified CPA methods are specialized for PD detection with higher power. Many IRT-based person-fit statistics, such as l_z, are applicable for PD detection. However, the maximum power of such a broad-spectrum method is below 0.55 for speededness and lack of motivation detection (de la Torre & Deng, 2008). Given the prevalence of PD, a targeted detection method (e.g., the method proposed in this study) must be developed. (2) Existing CPA methods determine whether a change point exists in a given response sequence by examining whether there is



(a) Subjects detected of PD by $JS_{\max}$ (b) Subjects detected of without PD by $JS_{\max}$

Fig. 7 Instant ability estimators for subjects. **a** Subjects detected as being affected PD by $JS_{\max}$, **b** subjects detected as not being affected by PD by $JS_{\max}$

a significant difference between two ability point estimates before and after the change point. The proposed method detects aberrancy by quantifying the difference between the posterior ability distributions before and after the change point. A difference between point estimates might be particularly unstable with an insufficient number of items in one of the subtests. Measuring the difference between the posterior distributions directly can provide greater stability and accuracy, resulting in higher power for $JS_{\max}$. The proposed method obtains the prior distributions from the entire response sequence and thus uses more information to obtain more accurate estimated posterior ability distributions. Thus, the proposed method captures the performance change before and after the change point more accurately and sensitively than other CPA methods. (3) Compared to the other three modified CPA methods, the proposed method yields a comparable Type-I error and a gain in power of between 1.0% to 8.2%, which implies that the proposed method is more accurate in detecting PD.

Despite many advantages, there are certain limitations of this study. First, this article generates ability from $N(0, 1)$ to construct the null distribution of four CPA statistics as done in Worsley (1979) and Shao and Cheng (2017). However, the critical value may not be the same among different ability levels; using $N(0,1)$ to simulate the null responses is equivalent to calculating the average critical value across different ability levels. In empirical studies, the ability distribution of a group may be completely different from the standard normal distribution. In that scenario, using $N(0,1)$ to construct the null distribution may not be the optimal choice. Hence, when constructing the null distribution in an empirical study, we can consider using the existing prior information about the respondents' ability or the distribution of the composition of ability estimator based on the response data to construct the null distribution. Second, in practice, respondents may answer items randomly, and not in sequence; thus, the effect of PD may be shown on all items, as all items could possibly appear at the beginning or end of a test. How this possibility may affect the detection of respondents with PD is interesting and should be investigated in future research. Third, the proposed method does not consider situations where respondent responses have multiple change points. For example, a response pattern would be affected by the warm-up effect at the beginning of a test and PD at the end of a test, which may be a more common phenomenon in practice. Our procedure can be easily extended to adapt to multiple change points. The first change point should be searched first, and then the second change point is determined given the first change point. Each search attempts to maximize the $JS_{\max}$. To examine whether these change points are statistically significant, the Monte Carlo simulation is still used to obtain the critical value of $JS_{\max}$ for the first and second change points. Fourth, in practice, more difficult items are typically placed at the end of a test to avoid frustrating respondents at the beginning of a test. In such cases, it is completely reasonable to obtain lower scores on end-of-test items. However, it is difficult to differentiate

between this reasonable response pattern and the response pattern with PD for the proposed method.

Considering the relative newness of CPA in psychometrics, future research should be informed by this study. First, the proposed method was used to detect PD based on the 2PL model. However, an educational and psychological test typically examines multiple underlying factors. A multidimensional extension is important and necessary (e.g., replacing 2PL with multidimensional 2PL). Once the model is confirmed, the posterior distribution can be computed and $JS_{\max}$ can be used to detect PD directly. Second, the proposed method requires known item parameters or parameters estimated from the response data. The presence of substantial proportions of aberrant responses (e.g., PD) can result in biased estimators of item parameters. Thus, outliers may fail to be successfully flagged, as the proposed method depends on item parameter estimators; this phenomenon is proverbially called the "masking effect" in the field of model-based outlier detection (Fung, 1993). Thus, certain outliers might be "masked" in that when structural parameters have been distorted by those outliers, they no longer appear to be outlying observations. Any model-based outlier detection method would be affected by the masking effect; thus, the problem of masking is not exclusive to the proposed method, or even CPA methods in general. Consequently, the degree of prevalence and severity at which the approaches fail to perform effectively should be studied. In addition, we re-preformed the simulation study to investigate the power of four CPA statistics when using the estimated item parameter. The results showed that although the powers of four CPA statistics all declined somewhat when using the estimated item parameter, the ranking of the power of the four CPA statistics was same as that using the true item parameters. For example, when using the estimated item parameter, the power of $JS_{\max}$, $L_{\max}$, $W_{\max}$ and $S_{\max}$ are 0.762, 0.733, 0.726 and 0.702, respectively, with the 40-item condition and the median and variance setting 0.6 and 0.001, respectively. Future studies can use estimated item parameters to investigate the performance of each CPA method, but should justify why the estimated parameters are chosen over the true parameters. However, the estimated item parameters may be contaminated with aberrant responses, thus affecting the performance of the CPA statistics. In this context, we suggest that future studies might consider selecting a subset of seemingly normal responses from the dataset to estimate item parameters, and then detect aberrant responses using the estimated item parameters.

Third, two similar median levels of 0.5 and 0.6 were set in the simulation of PD severity η. Considering that PD often happens at the end of the test, we added a simulation with a median and variance of 0.9 and 0.001 for η and 10% prevalence of PD within the 40-item test. The results showed that the power of $JS_{\max}$, $L_{\max}$, $W_{\max}$, and $S_{\max}$ was 0.386, 0.373, 0.351 and 0.341, the Type-I error rate was 0.054, 0.055, 0.053 and 0.051, and the mean absolute lag of change-point detection was 13.095, 13.406, 13.160 and 14.143, which indicates that future research should seek to further improve the power of CPA methods in detecting those who respond randomly at the end of a test. Based on these supplementary results, we can conclude that the superiority of $JS_{\max}$ diminishes when the change point approaches the end of a test, and the power of the four CPA statistics examined is extremely close. This phenomenon may be because fewer items are involved in the calculation of parameters (i.e., $\hat{\theta}_{2j}$, $S_{2j}(\theta)$) for CPA statistics before the change point, and thus those parameters will be less informative.

Funding The work was supported by the National Natural Science Foundation of China (32160203, 62167004 and 31960186).

Authors' Note

Ethics approval The questionnaire for this study was approved by mental health at the Research Center of Mental Health of Jingxi Normal University.

Consent to participate Informed consent was obtained from all participants in accordance with the Declaration of Helsinki

References

Adams, R., & Wu, M. (2002) *PISA 2000 technical report*. https://doi.org/10.1787/9789264167872-en.

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

An, W. P., Cheng, X. B., & Liu, Y. (2020). Application of Flessis' Kappa coefficient in Bayesian decision tree algorithm Computer Engineering and Applications, *Journal of Computer Engineering and Applications 56*(7), 137-140.

Barry, D., & Hartigan, J. A. (1993). A Bayesian Analysis for Change Point Problems. *Journal of the American Statistical Association, 88*(421), 309–319. https://doi.org/10.1080/01621459.1993.10594323.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*, 441–462. https://doi.org/10.1007/BF03173192

Biehler, M., Holling, H., & Doebler, P. (2014). Saddlepoint Approximations of the Distribution of the Person Parameter in the Two Parameter Logistic Model. *Psychometrika, 80*(3), 665–688. https://doi.org/10.1007/s11336-014-9405-1.

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*, 331–348. https://doi.org/10.1111/j.1745-3984.2002.tb01146.x

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. Psychometrika, 73, 209–230. https://doi.org/10.1007/s11336-007-9045-9

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1–29. https://doi.org/10.18637/jss.v048.i06

Cox, D. R. (2006). *Principles of statistical inference*. : Cambridge University Press

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. : Chapman and Hall.

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, School, and Country Differences in Sustained Test-Taking Effort in the 2009 PISA Reading Assessment. *Journal of Educational and Behavioral Statistics, 39*(6), 502–523. doi:https://doi.org/10.3102/1076998614558485

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*, 55–77. https://doi.org/10.1207/s15324818ame1301_3

de la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45*(2), 159–177. https://doi.org/10.1111/j.1745-3984.2008.00058.x

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. : Erlbaum, Inc.

Estrella, A., & Rodrihues, A. (2005). *One-sided test for an unknown breakpoint: Theory, computation, and application to monetary theory (staff Report No. 232)*. Federal Reserve Bank of New York.

Fung, W. K. (1993). Unmasking outliers and leverage points: A confirmation. *Journal of the American Statistical Association*, *88*, 515–519. https://doi.org/10.1080/01621459.1993.10476302

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, *68*, 907–922. https://doi.org/10.1177/0013164408315262

Goegebeur, Y., De Boeck, P., Molenberghs, G., & del Pino, G. (2006). A local-influencebased diagnostic approach to a speeded item response theory model. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *55*, 647–676. https://doi.org/10.1111/j.1467-9876.2006..00558.x

Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, *73*, 65–87. https://doi.org/10.1007/s11336-007-9031-2

Jin, K.-Y., & Wang, W.-C. (2014). Item Response Theory Models for Performance Decline During Testing. *Journal of Educational Measurement, 51*(2), 178–200. https://doi.org/10.1111/jedm.12041

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory, 37*(1), 145–151. https://doi.org/10.1109/18.61115

List, M. K., Robitzsch, A., Lüdtke, O., Köller, O., & Nagy, G. (2017). Performance decline in low-stakes educational assessments: Different mixture modeling approaches. *Large-scale Assessments in Education, 5*,1–25. https://doi.org/10.1186/s40536-017-0049-3

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, *26*(4), 29–37. https://doi.org/10.1111/j.1745-3992.2007.00106.x

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-scale Assessments in Education*. https://doi.org/10.1186/s40536-014-0005-4

Rao, C. R. (1973). *Linear statistical inference and its applications (2nd ed)*. : John Wiley.

Robbins, H. (1985). The Empirical Bayes Approach to Statistical Decision Problems. *Herbert Robbins Selected Papers*, 49–68. https://doi.org/10.1007/978-1-4612-5110-1_4

Robinson, L. F., Wager, T. D., & Lindquist, M. A. (2010). Change point estimation in multi-subject fMRI studies. *NeuroImage, 49*, 1581–1592. https://doi.org/10.1016/j.neuroimage.2009.08.061

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. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

Schüttpelz-Brauns, K., Kadmon, M., Kiessling, C., Karay, Y., Gestmann, M., & Kämmer, J. E. (2018). Identifying low test-taking effort during low-stakes tests with the new Test-taking Effort Short Scale (TESS) – development and psychometrics. *BMC Medical Education, 18*(1). https://doi.org/10.1186/s12909-018-1196-0

Shao, C. (2016). *Aberrant response detection using change-point analysis*. (Doctoral dissertation). University of Notre Dame, Notre Dame, IN.

Shao, C., & Cheng, Y. (2017, April). Detection of test speededness using change-point analysis with response time data. Paper presented at the annual Meeting of National Council for Measurement in Education, San Antonio, TX.

Shao, C., Li, J., & Cheng, Y. (2015). Detection of Test Speededness Using Change-Point Analysis. *Psychometrika, 81*(4), 1118–1141. https://doi.org/10.1007/s11336-015-9476-7

Sinharay, S. (2016). Person fit analysis in computerized adaptive testing using tests for a change point. *Journal of Educational and Behavioral Statistics, 41*, 521–549. https://doi.org/10.3102/1076998616658331

Sinharay, S. (2017a). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics, 42*, 46–68. https://doi.org/10.3102/1076998616673872

Sinharay, S. (2017b). Some remarks on applications of tests for detecting a change point to psychometric problems. *Psychometrika, 82*, 1149–1161. https://doi.org/10.1007/s11336-016-9531-z

Sinharay, S. (2017c). Which statistic should be used to detect item preknowledge when the set of compromised items is known? *Applied Psychological Measurement, 41*, 403–421. https://doi.org/10.1177/0146621617698453

Suh, Y., Cho, S.-J., & Wollack, J. A. (2012). A comparison of item calibration procedure in the presence of test speededness. *Journal of Educational Measurement, 49*, 285–311. https://doi.org/10.1111/j.1745-3984.2012.00176.x

Tendeiro, J. N., & Meijer, R. R. (2012). A CUSUM to Detect Person Misfit. *Applied Psychological Measurement, 36*(5), 420–442. https://doi.org/10.1177/0146621612446305

van Barneveld, C. (2007). The effect of test-taker motivation on test construction within an IRT framework. *Applied Psychological Measurement, 31,* 31–46. https://doi.org/10.1177/0146621606286206.

Wise, S. L. (1996, April). A persistence model of motivation and test performance. Paper presented at the annual meeting of the American Educational Research Association, New York, NY

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10,* 1–17. https://doi.org/10.1207/s15326977ea1001_1

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. https://doi.org/10.1207/s15324818ame1802_2

Wollack, J. A., & Cohen, A. S. (2004, April). A model for simulating speeded test data. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*, 227–242. https://doi.org/10.1207/s15324818ame0803_3

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test, motivation, and mentally taxing items. *Applied Measurement in Education*, *8*, 341–351. https://doi.org/10.1207/s15324818ame0804_4

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. https://doi.org/10.2307/2286336

Yu, X. F., & Cheng, Y. (2019). A Change-Point Analysis Procedure Based on Weighted Residuals to Detect Back Random Responding. *Psychological Methods, 24*(5). https://doi.org/10.1037/met0000212

Zhang, J. (2014). A sequential procedure for detecting compromised items in the item pool of a CAT system. *Applied Psychological Measurement, 38*, 87–104. https://doi.org/10.1177/0146621613510062

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.