Exploring Change Point Analysis of Response Time for Detecting Speeded Responding

Authors: Zhong Xiaoyuan, Yu Xiaofeng, Miao Ying, Qin Chunying, Peng Yafeng, Tong Hao, Xiaofeng Yu

Date: 2022-05-14T00:00:00+00:00

Abstract

Compared to traditional discrete response data, response time as continuous data can provide additional information. Change point analysis represents a relatively novel technique in psychological and educational research. This paper provides both a comprehensive synthesis and analysis of change point analysis applications in psychometrics, and extends two change point statistics from response data to response time data, applying change point methodology to detect anomalous test-taking behavior: speededness. Two testing methods—the likelihood ratio test and Wald test—are implemented to detect anomalous response patterns under conditions of known and unknown item parameters. Results demonstrate that the proposed methods exhibit high power for detecting speeded behavior while maintaining adequate control of Type I error rates. Empirical data analysis further substantiates the practical utility of the methods presented herein.

Full Text

Preamble

Exploration and Research on Change-Point Analysis Based on Response Time for Detecting Test Speededness

Zhong Xiaoyuan[1], Yu Xiaofeng[1], Miao Ying[1], Qin Chunying[2], Peng Yafeng[1], Tong Hao[1]
([1] School of Psychology, Jiangxi Normal University, Nanchang, 330022, China)
([2] School of Mathematics and Information Science, Nanchang Normal University, Nanchang, 330032, China)

Abstract

Compared to traditional discrete response data, response time as continuous data can provide substantially more information. Change-point analysis (CPA) represents a relatively new technique in psychological and educational measurement. This paper provides a comprehensive summary and analysis of CPA applications in psychometrics while extending two CPA statistics from response data to response time data, applying CPA to detect an aberrant response pattern: test speededness. Two testing methods—the likelihood ratio test and Wald test—are employed to detect aberrant response patterns under conditions of both known and unknown item parameters. Results demonstrate that the proposed methods exhibit high power for detecting speededness while effectively controlling Type I error rates. Empirical data analysis further confirms the practical utility of the methods presented herein.

Keywords: change-point analysis, aberrant response behavior, response time, test speededness, statistical process control

1. Introduction

High-quality measurement data is essential for accurate assessment of examinee ability. However, numerous factors can introduce systematic errors that compromise data quality, with aberrant response behaviors being among the most common. Typical aberrant behaviors during testing include warm-up effects and test speededness (Luo et al., 2020; Zhang et al., 2020). When examinees engage in aberrant behavior, their response data differ significantly from normal response patterns. Data generated under aberrant conditions are termed aberrant response data or aberrant response patterns. The presence of such data degrades test quality and adversely affects subsequent analyses, causing model-data misfit, biased estimation of examinee and item parameters (Stefan et al., 2016), and compromising test reliability and validity (Guo et al., 2009). Consequently, detecting aberrant response data is critically important, and researchers have continuously sought effective solutions (e.g., Bejar, 1985; Evans & Reilly, 1972; Shao et al., 2016; Bradlow et al., 1998; McLeod et al., 2003; Wise & Kong, 2005; Yu & Cheng, 2019, 2020).

Change-point analysis (CPA; Page, 1955; Shao et al., 2016; Sinharay, 2016) is a widely used method for detecting anomalies in process data, primarily designed to identify distributional changes in sequential data. Its fundamental principle states that when samples are ordered chronologically, if the underlying distribution or sample characteristics (such as mean or variance) change significantly from a certain point onward, a change point has occurred (Hawkins et al., 2003), indicating a qualitative shift.

Recent research has introduced CPA into psychological and educational measurement to detect aberrant response behaviors or patterns (Zhang, 2014; Shao, 2016; Shao et al., 2016; Sinharay, 2016, 2017a, 2017b, 2017c; Yu et al., 2019, 2020). Studies by Shao et al. (2016), Sinharay (2016), and Yu et al. (2019, 2020)

have demonstrated CPA' s advantages in detecting aberrant behaviors. During testing, examinee responses to each item form unique sequential data that typically follow a specific distribution; for instance, response time data often follow a lognormal distribution. When aberrant behavior occurs, the response data undergo a qualitative change, with the location of aberrant behavior (item number) representing the change point.

CPA can detect aberrant behavior using two data sources: response data (correct/incorrect) and response time data (duration per item). Response time data, being continuous, contain information about both examinee ability and item characteristics (Marianti et al., 2014), thereby improving ability estimation precision and optimizing test design. With technological advances, computer-based testing and online assessment have made response time data increasingly accessible and attracted scholarly attention. For example, van der Linden and van Krimpen-Stoop (2003) used response time data to detect item preknowledge and test speededness; van der Linden and Guo (2008) and Pan and Wollack (2021) employed response time data to detect item compromise. Researchers have also developed various response time models for different applications, demonstrating that incorporating response time data aids parameter estimation and expands its utility (Wang & Xu, 2015; Guo & Luo, 2019; Zhan, 2019; Zhan et al., 2020).

Previous CPA-based research on aberrant behavior detection has primarily focused on response data, yet the advantages of response time data are now evident. Integrating response time data into analyses represents an important trend. Moreover, test speededness is among the most common and prevalent aberrant behaviors (Goegebeur et al., 2008), significantly impacting data quality and attracting considerable research attention (e.g., Bolt et al., 2002; Oshima, 1994; Suh et al., 2012; Yu et al., 2020). This study therefore focuses on response time data, applying CPA methods to detect aberrant response patterns caused by speededness under both known and unknown item parameter conditions. Notably, while CPA is fundamentally a method for detecting anomalous data, it can also detect aberrant patterns caused by other behaviors such as item preknowledge or warm-up effects. We now introduce CPA technology in detail.

2. Change-Point Analysis (CPA) Technology

CPA is widely applied in biology, statistics, and economics. Although introduced to psychometrics, it remains underdeveloped. Key CPA-based studies for aberrant behavior detection include Zhang (2014), Shao et al. (2016), Shao (2016), and Sinharay (2016, 2017a, 2017b, 2017c), Yu et al. (2019, 2020).

Zhang (2014) addressed item compromise due to preknowledge, proposing a real-time sequential item monitoring method. Shao et al. (2016) employed CPA with likelihood ratio tests to detect speededness, enabling classification of examinees into speeded and non-speeded groups while accurately locating the onset

of aberrant behavior. This capability allows test administrators to improve ability estimation accuracy by removing suspected speeded responses and provides references for setting appropriate test lengths. Shao et al. (2016) used test statistics $\Delta l_i = 2(l_i^{H_a} - l_i^{H_0})$, where $l_i^{H_a}$ and $l_i^{H_0}$ represent log-likelihood values under speeded and normal response conditions, respectively. For a given examinee i' s score data, MLE (Baker & Kim, 2004) can estimate ability θ_i, yielding $l_i^{H_0}$. Under H_a, $l_i^{H_a} = l_i^{j-} + l_i^{j+}$, where l_i^{j-} and l_i^{j+} are likelihood functions for two subtests split at change point j (subtest 1: items 1 to j; subtest 2: items $j+1$ to n). The location where Δl_i reaches its maximum indicates where speededness begins. Its null distribution and critical values can be obtained through permutation distributions (Shao et al., 2016), empirical distributions (Yu et al., 2020), or theoretical approximations (Sinharay, 2016).

Sinharay (2016) used three CPA statistics to examine person-fit in CAT, computing Type I error rates and power. Approximate null distributions were used, revealing good performance in detecting examinees with aberrant behavior. The three CPA statistics are:

$$L_j = -2\{L(\hat{\theta}_0; Y_1, Y_2, \ldots, Y_n) - L(\hat{\theta}_{1j}; Y_1, Y_2, \ldots, Y_j) - L(\hat{\theta}_{2j}; Y_{j+1}, Y_2, \ldots, Y_n)\},$$

$$W_j = \frac{(\hat{\theta}_{1j} - \hat{\theta}_{2j})^2}{\frac{1}{I_{1j}(\hat{\theta}_0)} + \frac{1}{I_{2j}(\hat{\theta}_0)}},$$

$$S_j = \frac{(\nabla(\hat{\theta}_0; Y_1, Y_2, \ldots, Y_j))^2}{I_{1j}(\hat{\theta}_0)} + \frac{(\nabla(\hat{\theta}_0; Y_{j+1}, Y_{j+2}, \ldots, Y_n))^2}{I_{2j}(\hat{\theta}_0)}.$$

Since the change point is unknown, the three test statistics become:

$$L_{\max} = \max_{1 \le j \le n-1} L_j, \quad W_{\max} = \max_{1 \le j \le n-1} W_j, \quad S_{\max} = \max_{1 \le j \le n-1} S_j.$$

Here, $I_{1j}(\hat{\theta}_0)$ and $I_{2j}(\hat{\theta}_0)$ are Fisher information estimates based on items 1 to j and items $j+1$ to n when $\theta = \hat{\theta}_0$. $\nabla(\hat{\theta}_0; Y_1, Y_2, \ldots, Y_j)$ is the first derivative of the log-likelihood for $Y_1, Y_2, \ldots, Y_j$ at $\theta = \hat{\theta}_0$. Sinharay (2016) also used ROC curves to compare CUSUM and CPA methods, showing CPA's superiority under many conditions.

Sinharay (2017a) proposed likelihood ratio test (L) and Lagrange multiplier (score test) statistics (R) to detect item preknowledge. The test was divided into parts s and $\bar{s}$. For normal examinees (without preknowledge), posterior distributions and ability estimates based on s and $\bar{s}$ are similar, but differ significantly when preknowledge exists.

$$L = 2[L(\hat{\theta}_s; y_j, j \in s) + L(\hat{\theta}_{\bar{s}}; y_j, j \in \bar{s}) - L(\hat{\theta}_0; y_j, j = 1, 2, \dots, n)],$$

$$R = \frac{[\nabla(\hat{\theta}_0; y_j, j \in s)]^2}{I_s(\hat{\theta}_0)} + \frac{[\nabla(\hat{\theta}_0; y_j, j \in \bar{s}}{)}]^2 I_{\bar{s}}(\hat{\theta}_0).$$

Here, $\hat{\theta}_s$, $\hat{\theta}_{\bar{s}}$, and $\hat{\theta}_0$ are ability estimates based on s, $\bar{s}$, and all items ($j = 1, 2, \dots, n$). y_i represents scores, $L(\hat{\theta}_s; y_j, j \in s)$ is the log-likelihood for part s, $\nabla(\hat{\theta}_0; y_j, j \in s)$ is the first derivative of log-likelihood (Baker et al., 2004), and $I_s(\hat{\theta}_0)$ is the sum of item information for part s at ability $\hat{\theta}_0$.

These statistics apply to adaptive and non-adaptive tests with dichotomous and polytomous items, following asymptotic standard normal distributions—highly advantageous for practical application. Results showed controllable Type I error rates and relatively high power.

Sinharay (2017b) provided a general CPA detection framework, discussing statistic selection, critical value acquisition, and related issues, offering solutions. Using the Rasch model, three real-data examples demonstrated CPA' s application to psychometric problems.

Sinharay (2017c) compared two CPA statistics—likelihood ratio statistic L_s and posterior shift statistic PSS (Belov, 2016)—for detecting item preknowledge under the three-parameter logistic model (3PLM; Birnbaum, 1968). The detection principle mirrors Sinharay (2017a): when examinees benefit from preknowledge, ability estimates or posterior distributions differ substantially between compromised (c) and uncompromised (u) sections, with c section estimates being higher/right-shifted. $|L_s|$ equals the square root of L (Sinharay, 2017a). PSS quantifies the distance between posterior distributions. Results showed similar Type I error rates and detection rates.

Yu et al. (2019) proposed a CPA statistic based on weighted residuals, comparing it with three other CPA statistics (Sinharay, 2016) for detecting back random responding (BRR). The weighted residual statistic accurately detected BRR in tests with 20+ items, controlling Type I error while achieving 17%-42% higher power.

Yu et al. (2020) compared 12 CUSUM statistics with 3 CPA statistics for detecting speededness. To evaluate robustness and flexibility, they simulated two acceleration mechanisms: gradual speed change and abrupt speed change, using the graduate change model (GCM) and hybrid model (HM). Variables included test length, prevalence of speeded behavior (proportion of speeded examinees), and severity (proportion of items affected).

To provide a concrete understanding of CPA applications in psychometrics, Table 1 summarizes relevant studies from multiple perspectives. Notably, all studies used response data; regarding critical values, only Shao et al. (2016) used

permutation distributions and Sinharay (2016) used approximate values, with others employing empirical critical values. Empirical critical values are advantageous due to simple implementation and universal applicability across statistics, unlike approximate values (limited to certain statistics) or permutation methods (computationally intensive).

Although Table 1 contains no CPA studies based on response time data, related research exists. Choe et al. (2018) used sequential analysis to detect compromised items based on response data, response time data, and combined data. Their results showed that: (1) response-time-only methods had substantially higher power than response-only methods at the same Type I error rate; (2) combined-data methods showed mixed results—one approach had slightly higher power than response-time-only, while another was far inferior. Response-time-based detection also showed the smallest detection lag. Choe et al. (2018) demonstrated that response time data indeed provides more test information, yielding substantial power improvements.

With response time data increasingly accessible and offering inherent advantages for detecting aberrant behavior over response data, response-time-based detection shows excellent research prospects. For example, when an examinee' s response pattern is [1111101010], speededness is difficult to detect from scores alone, but combined with response time data [57, 48, 51, 36, 42, 23, 18, 13, 7, 6], acceleration becomes evident as speededness directly manifests in response times.

3. CPA-Based Statistics

Table 1 identifies four common CPA statistics: likelihood ratio test statistics (L_j), score test statistics ($\nabla l_i, L_{\max}, L_S$), Wald test statistics ($W_{\max}, W_i$), and residual test statistics ($R_S, S_{\max}, R_{\max}$). All test the null hypothesis of no significant change in latent traits or response data to detect aberrant behavior. This study focuses on response time data for aberrant behavior detection, selecting likelihood ratio and Wald test statistics.

We specifically examine speededness, using one-tailed tests. After speededness onset, examinee response speed increases. When change point k is known, the null hypothesis states that examinee i' s speed parameter equals before and after item k: $\hat{\tau}_{i,k-} = \hat{\tau}_{i,k+}$. The alternative hypothesis posits that speed increases: $\hat{\tau}_{i,k-} < \hat{\tau}_{i,k+}$. We now introduce the two statistics for response time data.

3.1 Likelihood Ratio Test

Van der Linden' s (2006) lognormal model is the most widely used response time model, showing good fit in many empirical studies. This study adopts this model, assuming examinee i' s response time t_{ij} on item j follows:

$$f(t_{ij};\tau_i,\alpha_j,\beta_j)=\frac{\alpha_j}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{[\alpha_j(\ln t_{ij}-(\beta_j-\tau_i))]^2}{2}\right\},$$

or equivalently:

$$\ln(t_{ij})=\beta_j-\tau_i+\varepsilon_{ij},\quad \varepsilon_{ij}\sim N(0,\alpha_j^{-2}).$$

Here, $\beta_j\in(-\infty,\infty)$ is the time intensity parameter (higher values indicate more time required); $\tau_i\in(-\infty,\infty)$ is examinee i's speed parameter, typically assumed normal; α_j is the time discrimination parameter, analogous to discrimination parameters in item response models.

From this model, examinee i' s response time data t_i has likelihood:

$$L(\tau_i;t_i)=\prod_{j=1}^{n}\frac{\alpha_j}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{[\alpha_j(\ln t_{ij}-(\beta_j-\tau_i))]^2}{2}\right\},$$

where $t_i=(t_1,t_2,\ldots,t_n)$ is the vector of response times. The log-likelihood is:

$$l(\tau_i;t_i)=\ln L(\tau_i;t_i)=-\sum_{j=1}^{n}\ln(t_{ij}\sqrt{2\pi})-\frac{1}{2}\sum_{j=1}^{n}\{[\alpha_j(\ln t_{ij}-(\beta_j-\tau_i))]^2\}.$$

The likelihood ratio test formula follows Shao et al. (2016):

$$\Delta l_i(k)=2(l_i^{H_a}(k)-l_i^{H_0}),$$

where $l_i^{H_0}=l(\hat{\tau}_{i,0};t_i)$ is the log-likelihood using speed parameter $\hat{\tau}_{i,0}$ estimated from all items. Assuming speed changes abruptly after item k, the speeded log-likelihood is:

$$l_i^{H_a}(k)=l(\hat{\tau}_{i,k-};t_i(k-))+l(\hat{\tau}_{i,k+};t_i(k+)),$$

where $\hat{\tau}_{i,k-}$ is estimated from the first k items ($t_i(k-)=(t_{i1},t_{i2},\ldots,t_{ik})$) and $\hat{\tau}_{i,k+}$ from items $k+1$ to J.

Since the change point location is unknown in practice, the test statistic is the maximum $\Delta l_i(k)$ across all possible change points:

$$\Delta l_{\max,i}=\max_{k=1,2,\ldots,(J-1)}\Delta l_i(k).$$

When $\Delta l_{\max,i}$ exceeds the acceptable range at a given confidence level, we reject the null hypothesis, indicating a change point in examinee i' s response time data.

3.2 Wald Test

The one-tailed Wald test statistic is:

$$W_i(k) = \frac{(\hat{\tau}_{i,k-} - \hat{\tau}_{i,k+})^2}{\frac{1}{I_{k-}(\hat{\tau}_{i,0})} + \frac{1}{I_{k+}(\hat{\tau}_{i,0})}},$$

where $I_{k-}(\hat{\tau}_{i,0})$ and $I_{k+}(\hat{\tau}_{i,0})$ are Fisher information estimates based on the first k and last $(J-k)$ items. When k is unknown, the test statistic becomes:

$$W_{\max,i} = \max_{k=1,2,\ldots,(J-1)} W_i(k).$$

When $W_{\max,i}$ exceeds the critical value, we reject the null hypothesis, indicating a change point. This study uses the item following the maximum of $l_i(k)$ and $W_i(k)$ as the estimated acceleration point where examinee i begins speeded behavior.

3.3 Obtaining Critical Values for CPA Statistics

Like likelihood ratio tests for response data, $\Delta l_{\max,i}$ based on response time data lacks a closed-form distribution. As shown in Table 1, critical values can be obtained via permutation distributions, empirical critical values, or approximate critical values. Permutation methods are computationally intensive, while approximate critical values are suitable when change points occur near the middle (e.g., the central 70% of the test; Sinharay, 2016). Since speededness typically occurs in later test stages when time pressure is greatest, this study employs empirical critical values.

When k is known, $W_i(k)$ follows a chi-square distribution with 1 degree of freedom. When k and $W_i(k)$ are unknown, $W_{\max,i}$ also lacks a closed-form distribution. Sinharay (2016) noted that $W_{\max,i}$' s asymptotic null distribution matches that of the likelihood ratio statistic. Thus, Wald test critical values also use empirical values. Following Worsley (1979), we generated 10,000 normal response time patterns for test lengths of 40, 60, and 80 using Equation 11. We calculated 10,000 values each for $\Delta l_{\max,i}$ and $W_{\max,i}$, sorted them, and extracted the 500th, 100th, and 10th largest values ($c_{0.05}$, $c_{0.01}$, $c_{0.001}$) as approximate critical values for $\alpha = 0.05, 0.01, 0.001$. Each condition was replicated 100 times, with averaged values serving as final empirical critical values.

4. Response Time Model for Speededness

Speededness typically occurs in time-limited tests. As time expires, unfinished examinees accelerate, reducing response times. Previous studies have simulated speeded response times using: (1) fixed time levels (e.g., 10s, 20s, 30s; van der Linden et al., 2008); or (2) adding a constant L to speed parameter τ_i in the lognormal model (van der Linden et al., 2003 used $L = 0.375$ and 0.750). Both treat speededness as a fixed effect, which is unrealistic.

Yu et al. (2019) reviewed two potential acceleration mechanisms: abrupt speed change and gradual speed change, modeled by the hybrid model (HM) and graduate change model (GCM). HM assumes abrupt speed changes at random change points, while GCM assumes gradually decreasing correct response probabilities after acceleration.

This study adopts the more plausible gradual speed change approach. Wollack and Cohen (2004) developed a speededness model for response data where correct response probabilities "gradually decrease," with each speeded examinee having a unique acceleration pattern. Goegebeur et al. (2008) examined this model' s parameter estimation. The gradual-change three-parameter model is:

$$P_{ij}^{*} = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]} \times \min\left(1, \left[1 - \left(\frac{j - \eta_i}{J - \eta_i}\right)\right]^{\lambda_i}\right),$$

where c_j is the guessing parameter, $\frac{\exp[a_j(\theta_i - b_j)]}{1+\exp[a_j(\theta_i - b_j)]}$ is the standard 2PLM, η_i $(0 \leq \eta_i \leq 1)$ indicates the acceleration location (e.g., $\eta_i = 0.8$ means speededness begins on the last 20% of items), and λ_i modulates the rate of probability decline.

Analogously, this study constructs a gradual-decline response time model based on the lognormal model:

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

Parameters η_i and λ_i have the same meaning as above. Before reaching the stage indicated by η_i, $j < \eta_i$, so $[1 - (\frac{j-\eta_i}{J-\eta_i})] > 1$ and $\min(1, \cdot) = 1$, meaning response times follow the standard lognormal model. After reaching η_i, $j > \eta_i$, the factor becomes < 1, reducing $\ln(t_{ij})$ below normal values, indicating speeded behavior.

To fully demonstrate CPA' s application and evaluate its performance on response time data, we conducted simulation studies and empirical data analysis. Since item parameters may be known (e.g., in adaptive testing systems) or

unknown, we examined both conditions. For known parameters, only speed parameters were estimated using EAP (Shao, 2016). For unknown parameters, MCMC (Fox et al., 2021) estimated them from all examinee data.

5. Simulation Study

5.1 Simulation Design

The simulation fixed the number of examinees at 1,000, with test lengths of 40, 60, and 80 items and total test times of 60, 90, and 120 minutes, respectively (Shao, 2016). The proportion of speeded examinees was 10%, 20%, and 30%, representing low, medium, and high prevalence. Change point locations η_i were generated from four distributions (detailed in data generation). Tests terminated when time expired, with unfinished items receiving response times of 0 and examinees marked as speeded. The full factorial design yielded $3 \times 3 \times 4 \times 2 = 72$ conditions, each replicated 50 times. Simulations were conducted in R.

Table 2 presents simulation conditions including test length, proportion of speeded examinees, change point parameters, and item parameter conditions (known/unknown).

5.2 Data Generation

Following Patton (2015), item discrimination a and difficulty b were set as $a \sim \ln N(0, 0.5)$, $b \sim N(0, 1)$. Response times for normal and speeded examinees were generated using Equations (11) and (20), with time discrimination $\alpha_j \sim U(1.75, 3.25)$. Time intensity β_j and speed τ_i followed Patton (2015): β_j had mean 4 and SD 1/3, with correlation 0.3-0.5 with a and b; $\tau_i \sim N(0, 0.25)$. For speeded examinees, speed modulation parameter $\lambda_i \sim \log N(3.912, 1)$ (Suh et al., 2012). Change point η_i followed beta distributions with medians 0.6 and 0.7 and variances $\sigma^2_{\eta_i} = 0.001$ and 0.04: beta(143.367, 95.689), beta(2.970, 2.091), beta(146.345, 62.910), and beta(3.033, 1.490). **Figure 1 [Figure 1: see original paper]** shows these distributions. Note that η_i represents acceleration location as a percentage. For a 40-item test, $\eta_i = 0.6$ indicates speededness begins at item 25. When $\sigma^2_{\eta_i} = 0.001$, change points cluster near the median; when $\sigma^2_{\eta_i} = 0.04$, they can appear anywhere, even near the test end. We focus on mid-to-late test stages because speededness typically occurs later, and we want to examine whether approximate critical values remain applicable when change points are not central (Sinharay, 2016).

Critical values for both statistics were first generated via Monte Carlo simulation. For known item parameters, calculations used true values; for unknown parameters, estimates were used. The process then evaluated speededness detection.

Based on the CPA statistics analysis, we applied likelihood ratio and Wald statistics sequentially to each examinee' s response time data: (1) compute likelihood ratios for each item, using the maximum as the statistic (Wald test

similar); (2) compare statistics to condition-specific critical values, flagging data as aberrant when exceeded; (3) record aberrant behavior locations; (4) evaluate detection effectiveness using predetermined metrics.

5.4 Evaluation Metrics

Performance was evaluated using Type I error rate and power (averaged across replications). We also computed the proportion of students not finishing within the time limit (%NF) and absolute detection lag (ADL) between detected and true change points:

$$\text{Type I error} = \frac{\text{Number of normal examinees falsely flagged}}{\text{Total normal examinees}},$$

$$\text{Power} = \frac{\text{Number of speeded examinees correctly flagged}}{\text{Total speeded examinees}},$$

$$\text{ADL} = \frac{\sum_{i=1}^{N} |\hat{p}_i - p_i|}{N},$$

where $\hat{p}_i$ and p_i are detected and true change point locations for examinee i, and N is the number of examinees. %NF assesses whether test time and length settings are reasonable, providing useful design information.

5.5 Simulation Results

Table 3 presents critical values for known and unknown item parameters, showing means and standard deviations across conditions. Since likelihood ratio and Wald critical values were nearly identical, only likelihood ratio values are shown. Unknown item parameters had minimal impact on critical values. Critical values varied substantially with α levels but only slightly with test length, increasing modestly as length increased. Variances were small for $\alpha = 0.05$ and 0.01, indicating stability. Larger variance at $\alpha = 0.001$ is expected given the 10,000-sample distribution' s tail behavior.

Sinharay (2016) reported approximate critical values ranging 8.45-9.84 for $\alpha = 0.05$ and 11.69-13.01 for $\alpha = 0.01$. Our empirical values differ somewhat, likely because approximate values suit long tests with central change points, whereas our simulations used shorter tests with change points anywhere. Stable empirical critical values across test lengths and replications support their appropriateness. **Table 3** values serve as critical values for $\alpha = 0.05, 0.01, 0.001$.

Tables 4 and **5** show power and Type I error rates for known and unknown parameters. Both statistics performed slightly better with known parameters, but trends were consistent. Power was high (often near 1) across most conditions except test length 80 at $\alpha = 0.001$. Power increased with test length

and proportion of affected items. Compared to response-data-based methods (Shao et al., 2016: power 0.60-0.90; Sinharay, 2016: generally lower; Yu et al., 2020: lower than our results), response-time-based detection shows substantial advantages. Type I error rates were well-controlled, slightly exceeding nominal levels.

ADL means and SDs were small when $\eta_{var} = 0.001$. When $\eta_{var} = 0.04$ with test length 80, ADL reached nearly 14. As **Figure 1** shows, larger η_{var} allows change points anywhere, including near the end, making accurate detection difficult. Research (Andrews, 1993; Hawkins et al., 2003) suggests CPA works best in medium-length tests with change points in the central 70% (Andrews, 1993).

%NF results showed 3.9%-5% of examinees failed to finish the 40-item test, 5.9%-7.3% for the 60-item test, and 5.9%-8% for the 80-item test when 30% were speeded. These settings appear reasonable, as most speeded examinees still finished. Unfinished examinees were flagged as aberrant. Power near 1 indicates CPA detects less severe speededness—examinees who accelerate but still finish on time.

Table 5 shows results for unknown item parameters. Power remained high (minimum 0.89) with well-controlled Type I error, though slightly lower than **Table 4**. Trends persisted: power increased with test length (40-item average: 0.94-0.93; 60-item: 0.97-0.96). Type I error rates approximated nominal levels at $\alpha = 0.05, 0.01$, but were smaller at $\alpha = 0.001$ due to extreme-condition empirical critical values. ADL increased with test length (from 3.19 to 5.99) and was affected by η_{median} and η_{var}, with larger η_{var} causing greater delays (mean 7.92, SD 8.44).

Our conditions resemble Shao et al. (2016) and Yu et al. (2020). Though not directly comparable, results are informative. For known parameters with 40 items, response-data-based likelihood ratio and Wald tests showed power of 0.50-0.94, lower than our 0.89-0.97.

6. Empirical Data Analysis

To demonstrate CPA' s application to real data, we analyzed mathematics test data from a regional basic education assessment. One test form contained 30 items with 45-minute time limit and 36,000 examinees' response times. All items were multiple-choice, computer-administered. We cleaned the data by removing examinees with total times <5 minutes and those with zero response times on end-of-test items (to examine detection of milder speededness). This retained 33,000 examinees; we randomly sampled 5,000 for analysis.

Figure 2 [**Figure 2: see original paper**] shows the speed parameter distribution histogram. Using the LNIRT package (Fox et al., 2007, 2021), we fitted the lognormal model to obtain item parameters α_j, β_j and examinee speed parameters $\tau_{i,0}, \tau_{i,j-}, \tau_{i,j+}$. The 5,000 examinees' speed parameters had mean 0

and SD 0.267, showing a negatively skewed distribution.

These parameters calculated likelihood ratio and Wald statistics. Results were very similar, so we present Wald test results only. Using thresholds 8.068, 11.214, and 15.702, we identified 675, 361, and 271 speeded examinees, respectively.

Figure 3 [Figure 3: see original paper] shows response times for examinee #1034 (flagged as aberrant), along with expected times, sample mean times for aberrant examinees, and overall sample means. The blue line shows #1034' s times, red shows overall item means, green shows expected times, and gray shows mean times for all "aberrant" examinees. Examinee #1034 responded faster than average on the first 18 items but slower thereafter. The last 12 items were completed in $ $30 seconds (some ~10 seconds), a sharp decrease from earlier items. Both #1034 and the overall sample showed decreasing times near test completion, but "aberrant" examinees' times decreased more dramatically.

Average response times were 30-65 seconds per item (except the final item at ~26 seconds). Examinee #1034' s times matched or exceeded averages initially but fell below after item 18, diverging increasingly from expected times. This supports flagging #1034 as aberrant.

7. Discussion

This study applied two CPA test statistics to response time data for detecting speededness. Simulation and empirical analyses showed both statistics achieved very high power with well-controlled Type I error, demonstrating CPA' s high effectiveness for aberrant behavior detection using response time data.

Response time and response data each have unique advantages. Combining them could further analyze aberrant behavior types (Wang et al., 2018). Response time' s continuous nature provides richer information and detection advantages, while response data helps identify behavior types. Future research should explore combining polytomous response data (Chen et al., 2010; Cheng et al., 2012) with response time data.

Our finding that fixed critical values work across test lengths simplifies application—no need to recompute when adding or removing items. The method should extend easily to CAT or multistage adaptive testing (Li & Ding, 2018; Xiong et al., 2018). While we focused on speededness, CPA can detect other aberrant behaviors (e.g., low motivation in survey data).

CPA also shows promise for multidimensional tests, which are increasingly common (e.g., math tests in English; Zhang et al., 2020). Multidimensional response time models are emerging (Zhan et al., 2020), showing multidimensional latent processing speed structures that match latent abilities, enabling speed parameter estimation in multidimensional contexts.

This research has practical value. Developing aberrant behavior detection methods is crucial for test quality control. Problems persist due to inaccurate pa-

rameter estimation, equating bias, and misinterpretation of behavior. In lengthy summative tests, determining appropriate length to allow most examinees adequate time is essential (van der Linden, 2011; Patton, 2015). Recording item-level response times enables future CPA-based detection.

Limitations and Future Directions

First, while most tests record response data and some studies detect speededness using it, our study used response time data. When response-data and response-time-based detection conflict, additional information (test content analysis, other statistics, video records, historical data) is needed for careful evaluation (Wang et al., 2018). Second, we assumed known probability structures before/after change points, but these may be unknown in practice. Model-free change-point detection methods need exploration. Third, CPA cannot determine why data changed—low motivation and speededness both reduce response times. Expert domain knowledge is needed to identify causes. Future research should combine response and time data to develop CPA methods that leverage response time' s high detection power. For high-stakes tests, joint inference from both data sources is more appropriate. Finally, most CPA research uses large samples; future work should examine CPA' s effectiveness with small samples.

References

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4), 821–856. https://doi.org/10.2307/2951764.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Taylor & Francis Group. http://ebookcentral.proquest.com/lib/pqopenlayer/detail.action

Bejar, I. I. (1985). Test speededness under number-right scoring: An analysis of the test of English as a foreign language. *ETS Research Report*, 1985(1), i–57. https://doi.org/10.1002/j.2330-8516.1985.tb00096.x.

Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, 40(2), 83–97. https://doi.org/10.1177/0146621615603327.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee' s ability. In *Statistical theories of mental test scores* (pp. 395–479).

Bolt, D. M., Cohen, A. S., & Willock, 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. https://doi.org/10.1111/j.1745-3984.2002.tb01146.x.

Bradlow, E. T., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical*

Association, 93(443), 910–919. https://doi.org/10.1080/01621459.1998.1047374.

Choe, E. M., Zhang, J., & Chang, H.-H. (2018). Sequential detection of compromised items using response times in computerized adaptive testing. *Psychometrika*, 83(3), 650–673. https://doi.org/10.1007/s11336-017-9596-3.

Chen, Q., Ding, S. L., Zhu, L. Y., & Xu, Z. Y. (2010). Three-parameter graded response model and its parameter estimation. *Journal of Jiangxi Normal University (Natural Science)*, 34(2), 117–122.

Cheng, X. Y., Ding, S. L., Zhu, L. Y., & Wu, H. F. (2012). The stratified item selection strategy with maximal information under graded response model. *Journal of Jiangxi Normal University (Natural Science)*, 36(5), 446–451.

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. https://doi.org/10.1111/j.1745-3984.1972.tb00767.x.

Fox, J.-P., Entink, R. K., & Linden, W. J. van der. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, 20(7), 1–14. https://doi.org/10.18637/jss.v020.i07.

Fox, J.-P., Klotzke, K., & Simsek, A. S. (2021). LNIRT: An R package for joint modeling of response accuracy and times. *arXiv:2106.10144* [stat]. http://arxiv.org/abs/2106.10144.

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. https://doi.org/10.1007/s11336-007-9031-2.

Guo, J., Tay, L., & Drasgow, F. (2009). Conspiracies and test compromise: An evaluation of the resistance of test systems to small-scale cheating. *International Journal of Testing*, 9(4), 283–308. https://doi.org/10.1080/15305050903351901.

Guo, X. J., & Luo, Z. S. (2019). A psychometric model for speed-accuracy tradeoff and application. *Psychological Exploration*, 39(5), 451–460.

Hawkins, D. M., Qiu, P., & Kang, C. W. (2003). The changepoint model for statistical process control. *Journal of Quality Technology*, 35(4), 355–366. https://doi.org/10.1080/00224065.2003.11980233.

Li, J., & Ding, S. (2018). The several stratified methods of CAT in the presence of calibration error on GRM. *Journal of Jiangxi Normal University (Natural Science)*, 42(4), 374–378.

Luo, F., Wang, X., Xu, Y., & Feng, W. (2020). Research progress of cheating detection technology in examinations: Detection of group cheating. *China Examinations*, (11), 37–41.

McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, 27(2), 121–137. https://doi.org/10.1177/0146621602250534.

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, 39(6), 426–451. https://doi.org/10.3102/1076998614559412.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, 31(3), 200–219. https://doi.org/10.1111/j.1745-3984.1994.tb00443.x.

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4), 523–527. https://doi.org/10.2307/2333401.

Pan, Y., & Wollack, J. A. (2021). An unsupervised-learning-based approach to compromised item detection. *Journal of Educational Measurement*, 58(3), 413–433. https://doi.org/10.1111/jedm.12299.

Patton, J. M. (2015). *Some consequences of response time model misspecification in educational measurement* (Doctoral dissertation). University of Notre Dame.

Patton, J., Cheng, Y., Hong, M., & Diao, Q. (2019). Detection and treatment of careless responses to improve item parameter estimation. *Journal of Educational and Behavioral Statistics*, 44(3), 309–341. https://doi.org/10.3102/1076998618825116.

Shao, C. (2016). *Aberrant response detection using change-point analysis* (Doctoral dissertation). University of Notre Dame. https://curate.nd.edu/show/5425k932c5j.

Shao, C., Li, J., & Cheng, Y. (2016). 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(5), 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(1), 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(4), 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(6), 469–482. https://doi.org/10.1177/0146621617698453.

Stefan, Z., Dietrich, K., & Wolfgang, H. (2016). Are exam questions known in advance? Using local dependence to detect cheating. *PLOS ONE*, 11(12), e0167545. https://doi.org/10.1371/journal.pone.0167545.

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. https://doi.org/10.1111/j.1745-3984.2012.00176.x.

Toton, S. L., & Maynes, D. D. (2019). Detecting examinees with pre-knowledge in experimental data using conditional scaling of response times. *Frontiers in Education*, 4, 49. https://doi.org/10.3389/feduc.2019.00049.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2), 181–204. https://doi.org/10.3102/10769986031002181.

van der Linden, W. J. (2011). Test design and speededness. *Journal of Educational Measurement*, 48(1), 44–60. https://doi.org/10.1111/j.1745-3984.2010.00130.x.

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. https://doi.org/10.1007/BF02294800.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, 73(3), 365–384. https://doi.org/10.1007/s11336-007-9046-8.

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. https://doi.org/10.1111/bmsp.12054.

Wang, C., Xu, G., & Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, 83(1), 223–254. https://doi.org/10.1007/s11336-016-9525-x.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, 29(5), 323–339. https://doi.org/10.1177/0146621605275984.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2.

Wollack, J. A., & Cohen, A. S. (2004). A model for simulating speeded test data. In *Annual meeting of the American Educational Research Association*, San Diego, CA.

Woodall, W. H., & Montgomery, D. C. (1999). Research issues and ideas in statistical process control. *Journal of Quality Technology*, 31(4), 376–386. https://doi.org/10.1080/00224065.1999.11979944.

Worsley, K. J. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association*, 74(366a), 365–367. https://doi.org/10.1080/01621459.1979.10482519.

Xiong, J., Luo, H., Wang, X., & Ding, S. (2018). The online calibration based on graded response model. *Journal of Jiangxi Normal University (Natural Science)*, 42(1), 62–66.

Yu, X. F., & Cheng, Y. (2019). A change-point analysis procedure based on weighted residuals to detect back random responding. *Psychological Methods*, 24(5), 658–674. https://doi.org/10.1037/met0000212.

Yu, X. F., & Cheng, Y. (2020). A comprehensive review and comparison of CUSUM and change-point-analysis methods to detect speededness. *Multivariate Behavioral Research*, 55(5), 720–741. https://doi.org/10.1080/00273171.2020.1809981.

Zhang, J. (2014). A sequential procedure for detecting compromised items in the item pool of a CAT system. *Applied Psychological Measurement*, 38(2), 87–104. https://doi.org/10.1177/0146621613510062.

Zhang, L., Wang, X., Cai, Y., & Tu, D. (2020). Change point analysis: A new method to detect aberrant responses in psychological and educational testing. *Advances in Psychological Science*, 28(9), 1462–1477.

Zhan, P. D. (2019). Joint modeling for response times and response accuracy in computer-based multidimensional assessments. *Journal of Psychological Science*, 42(1), 170–178.

Zhan, P. D., Hong, J., & Man, K. W. (2020). The multidimensional log-normal response time model: An exploration of the multidimensionality of latent processing speed. *Acta Psychologica Sinica*, 52(9), 1132–1142.

Note: Figure translations are in progress. See original paper for figures.