Multivariate Behavioral Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

A Comprehensive Review and Comparison of CUSUM and Change-Point-Analysis Methods to Detect Test Speededness

Xiaofeng Yu & Ying Cheng

To cite this article: Xiaofeng Yu & Ying Cheng (2020): A Comprehensive Review and Comparison of CUSUM and Change-Point-Analysis Methods to Detect Test Speededness, Multivariate Behavioral Research, DOI: 10.1080/00273171.2020.1809981

To link to this article: https://doi.org/10.1080/00273171.2020.1809981

MULTIVARIATE BEHAVIORAL RESEARCH
https://doi.org/10.1080/00273171.2020.1809981

A Comprehensive Review and Comparison of CUSUM and Change-Point-Analysis Methods to Detect Test Speededness

Xiaofeng Yu[a,b] and Ying Cheng[a]

[a]Department of Psychology, University of Notre Dame; [b]Department of Psychology, Jiangxi Normal University

ABSTRACT

Cumulative sum (CUSUM) and change-point analysis (CPA) are two well-established statistical process control methods to detect changes in a sequence. Both have been used in psychometric research to detect aberrant responses in a response sequence, e.g., test speededness, inattentiveness, or cheating. However, the pros and cons of CUSUM and CPA in different testing settings still remain unclear. In this paper, we conduct a comprehensive comparison of the performance of twelve CUSUM-based statistics and three CPA-based procedures in detecting test speededness. Two speededness mechanisms are considered, namely the graduate change model (GCM) and the hybrid model (HM), to test the robustness and flexibility of the two methods. Simulation studies show that the performances of the statistics are affected by the underlying data generating model, the severity of speededness, and the test length. Generally, under HM some CUSUM statistics perform much better than the CPA-based statistics. Under the GCM, the performance of the CPA statistics is dramatically improved. Taken together, due to the unknown mechanism of speededness in real applications, two CUSUM-based statistics are recommended when the test length is long (e.g., 80 items), regardless of the underlying mechanism being HM or GCM. In a relatively short (e.g., 40 items) or medium-length (e.g., 60 items) test, no statistic always ends up in the top three under both HM and GCM. In those cases, either one of the two CUSUM-based statistics mentioned above can be a reasonable choice because of their good (though not necessarily the best) performance in a wide range of conditions.

KEYWORDS

Speededness; cumulative sum control chart; change point analysis; item response theory; intra-individual change; gradual change

Test speededness occurs when not all respondents have sufficient time to fully consider the answer for each question on a test within a fixed time limit (Bejar, 1985). According to Schnipke and Scrams (1997), speededness refers to the extent to which respondents' test scores are affected by time limits, as often measured by calculating the proportion of respondents who cannot complete a certain percentage of test items. According to van der Linden (2011), speededness in testing is the end results of the interaction between three important factors: the cognitive speed at which the test taker works during the test, the amount of labor required by the items, and the time limit on the test.

Speededness has been a long-standing issue in test theory (Schnipke & Scrams, 1997). Glliksen (1950, p. 367) pointed out that the item indices of classical test theory developed for power tests might not be appropriate if a test is speeded. For example, split-half reliability indices are abnormally high in speeded tests (Gulliksen, 1950, p. 236). Lu and Sireci (2007) noted that test speededness affects the accuracy of the reliability and validity estimates, as well as the accuracy of identifying the correct factor structure.

Under the item response theory (IRT) framework, Hambleton and Swaminathan (1985) pointed out that unidimensional IRT models implicitly assume that the test is unspeeded. If a test is speeded, the unidimensionality assumption might be violated. This was echoed in Yen (1993), which identified test speededness as one of the most prevalent causes of local dependence in educational testing. Oshima (1994), Schnipke (1996), and Shao et al. (2016) showed that IRT item and person parameter estimates are distorted by speededness. Wollack et al. (2003) studied the effect of test speededness on equating. Schlemer (2007) showed that in a computerized adaptive test (CAT) speededness could have a more detrimental impact on ability estimates than in a paper-and-pencil (P&P)

CONTACT Ying Cheng ycheng4@nd.edu Department of Psychology, University of Notre Dame, 390 Corbett Hall, Notre Dame, IN 46556, USA.
Supplemental data for this article can be accessed at https://doi.org/10.1080/00273171.2020.1809981.

test. Han (2013) also suggested that speededness might increase measurement errors in a CAT.

All these studies pointed to the critical importance of controlling test speededness. However, it is nearly impossible to completely eliminate speededness in a high-stakes test due to the necessity of time limits. Over the years, two approaches have been developed to address test speededness. One is to directly model it. In other words, instead of using the regular, unidimensional IRT models such as the one-, two-, or three-parameter logistic (1PL, 2PL, 3PL) models that assume the tests are non-speeded, new models directly take test speededness into account. These models treat speededness as a decrease in ability during the testing process that abruptly changes the probability of an examinee giving correct responses (Yamamoto, 1989), or as a process that gradually reduces the probability of correct responses as test progresses (Wollack & Cohen, 2004; Goegebeur et al., 2008), or as an auxiliary dimension in addition to the dimension of ability (van der Linden et al., 1999; van der Linden & Xiong, 2013). Some others model the testing process as a mixture of normal responding behavior and speededness (Bolt et al., 2002; Rost, 1990; Schnipke & Scrams, 1997; Wang & Xu, 2015; Yamamoto, 1989; Yamamoto & Everson, 1997). By accounting for test speededness in the model, one can obtain better item parameter estimates (Oshima, 1994; Suh et al., 2012), and more accurate ability estimates (Shao et al., 2016; van der Linden, 2009).

The second approach is to detect (examinees with) speeded responses and that will be the focus of our study. This is typically done by modeling the regular responses, and flagging responses or response patterns that do not conform to the regular response model. Many studies of this second approach fall under the umbrella of detection of aberrant responses. In this paper, we compare the performance of two statistical process control (SPC, which is a collection of methods for monitoring, controlling and improving a random process through statistical analysis) methods: CUSUM (Cumulative SUM; Page, 1954) and CPA. CUSUM procedure has been widely used in SPC, and it is effective in detecting small shifts in the mean of the variable being measured, originally proposed by Page (1954). In educational and psychological testing, researchers have proposed various CUSUM indices to detect aberrant response behaviors, such as Armstrong and Shi (2009a, 2009b), Bradlow et al. (1998), Egberink et al. (2010), Meijer (2002), Shi (2007), Tendeiro and Meijer (2012), Tendeiro et al. (2013), and van Krimpen-Stoop and Meijer (2000, 2001). This paper will briefly summarize these indices.

Another SPC procedure that can be used to detect aberrant response behaviors is the CPA. CPA is also a flexible tool to detect abrupt changes in a sequence of data which can be dated back to the 1950s. In recent years, CPA has also been used in educational testing to detect the unusual change in the mean score of international testing programs over time (Lee & von Davier, 2013), or change in the behavior of test items over time due to potentially compromised item pool (Zhang, 2014), or to detect response anomaly during the test taking process due to item preknowledge, speededness, person misfit (Shao et al., 2014, 2016; Sinharay, 2016, 2017a, 2017b) or carelessness (Yu & Cheng, 2019).

Even though both CUSUM and CPA have been applied to educational testing data to detect aberrant responses, all existing research focused only on one of these two methods, with Sinharay (2016, 2017b) being the two exceptions. Sinharay (2016) proposed three CPA-based statistics and compared their performance against four CUSUM-based indices in evaluating person-fit. He found that the CPA-based statistics were more powerful under his simulation conditions. However, the primary purpose of Sinharay (2016) was still to investigate and establish the effectiveness of CPA in evaluating person fit in the context of computerized adaptive testing. In order to provide a general review of tests of a change point in psychometric problems, Sinharay (2017b) answered three basic questions of applying the two methods. In the review below, we have identified at least twelve CUSUM-based statistics in the existing literature, which includes the four used in Sinharay (2016). Little is known on how these statistics compare against each other, let alone how they compare against the CPA-based methods in detecting test speededness. In other words, a comprehensive comparison of CUSUM- and CPA-based methods has been lacking.

More importantly, it is well known that the CUSUM procedures are the most appropriate (in the sense of being the most powerful) when the parameters of the underlying statistical model before and after the change are known (Hawkins et al., 2003; Montgomery, 2013). If one or more of the parameters are unknown, the application of tests based on CPA may be more appropriate than that of the CUSUM procedures (Sinharay, 2016, 2017b). For this reason, we suspect that the relative advantage of CPA over the CUSUM-based procedures found in Sinharay (2016) may or may not hold depending on the

underlying mechanism of speededness. Given this consideration, in this study we include two well-known models of speededness: the gradual change model (GCM; Suh et al., 2012; Wollack & Cohen, 2004) and the hybrid model (HM; Yamamoto, 1989; Yamamoto & Everson, 1997). The two models reflect two distinct mechanisms of speededness. The GCM, first proposed by Wollack and Cohen (2004), allows for a gradual decline in the probability of correct responses, mimicking the effect of increasingly felt time pressure on the part of test takers. Yamamoto and Everson (1997) proposed a hybrid model, which assumed that every item before the speeding point could be modeled by the ordinary 2PL model, and after that point, by random guessing. The model represents the situation of a student randomly choosing a response option when running out of time. The difference is that the GCM models a gradual decline in the probability of a correct answer, while the HM models an abrupt change. In particular, for the HM, the probability of answering an item correctly is fixed to a pre-determined constant. Hence CUSUM procedures may work better under the HM. In this paper we will conduct a comprehensive examination of the performance of CUSUM and CPA statistics, and evaluate each method's robustness and flexibility under both GCM and HM.

Methods

In educational and psychological measurement, test taking is often a sequential process, resembling an industrial process. This is more true than ever when many tests are given through the computer, and items are presented sequentially to the test takers, and one has to respond to the current item before moving on to the next one. In statistical process control literature, the CUSUM control chart is effective in detecting small shifts in the mean of the variable being measured, while the CPA is adept at detecting abrupt changes in a sequence of data. The use of CUSUM procedure to detect aberrant response patterns can be traced back to Bradlow et al. (1998), while the application of CPA in detecting intra-individual change in test taking is much more recent (e.g., Shao et al., 2016). Our literature review covers at least twelve CUSUM indices and three CPA statistics and they are summarized below.

CUSUM and existing CUSUM indices

Consider a manufacturing process from which we collected n products at sequential time points. The observation is denoted as x_j, $j = 1, 2, \ldots, n$. Then the CUSUM control chart is created based on the scatter plot of C^+ and C^- over time, where C^+ and C^- are the cumulative sum of the consecutive positive and negative residuals, respectively. Mathematically, C^+ and C^- are oftentimes defined as $max\{0, (x_j - \hat{\mu})\}$ and $min\{0, (x_j - \hat{\mu})\}$, where $\hat{\mu}$ is the estimate of the in-control mean. As long as the production process remains in control, x_j should be centered at $\hat{\mu}$, and the control chart should fluctuate in a random pattern centering around zero. If the mean shifts upward, the curve drawn by the value of C^+ points will eventually drift upward; or if the process mean continues decreasing, the curve drawn by the value of C^- point will drift downward. Detection of a mean change can be triggered when C^+ or C^- reaches a predefined critical value.

A general formula of the CUSUM can be presented as

$$C_j^+ = max\left\{0, T_j + C_{j-1}^+\right\}, \quad (1)$$

$$C_j^- = min\left\{0, T_j + C_{j-1}^-\right\}, \quad (2)$$

$$C_0^+ = C_0^- = 0, \quad (3)$$

where T_j is the j^{th} observed residual in the sequential process. Let UB and LB denote pre-specified upper and lower bound (van Krimpen-Stoop & Meijer, 2000), respectively. In an educational test, T_j can be some kind of residual (weighted or unweighted) between the expected and observed score of item j. Existing CUSUM statistics for detecting aberrant behaviors mostly differ in their definitions of T_j. Based on the logic of CUSUM, a response pattern would be identified as aberrant response pattern when $C_j^+ > UB$ or $C_j^- < LB$. A response pattern would be classified as normal otherwise.

To illustrate the application of CUSUM chart to testing, consider a response pattern of someone taking a 20-item dichotomous test given in Table 1 (column "u"). This response pattern was observed in one fixed-form module of a multistage computerized test. The ability of the examinee was estimated with $\hat{\theta} = -0.06$ based on his/her responses to these 20 items, based on 3PLM item parameters that were known in advance (also shown in Table 1). The definition of T_j used in this example is $\frac{1}{n}\left[u_j - P_j(\hat{\theta}_n)\right]$, which is the T_{5j} statistic introduced in the next section. It is a residual defined as the difference between an observed item response and the model-based expected item response.

Table 1. CUSUM procedure for an observed response pattern.

Item	a	*b*	*c*	*u*	*P*	*T*	C^+	C^-
1	0.976	−0.693	0.371	0	0.163	−0.008	0	−0.008
2	0.973	0.600	0.224	1	0.419	0.029	0.029	0
3	0.871	−0.607	0.159	1	0.741	0.013	0.042	0
4	0.768	−0.637	0.377	1	0.800	0.010	0.052	0
5	0.940	−1.095	0.159	1	0.865	0.007	0.059	0
6	1.109	−0.202	0.146	1	0.630	0.019	0.077	0
7	1.063	−0.679	0.181	1	0.798	0.010	0.087	0
8	0.888	0.058	0.251	1	0.592	0.020	0.108	0
9	0.648	−0.822	0.179	1	0.752	0.012	0.120	0
10	0.733	−0.768	0.214	1	0.770	0.012	0.132	0
11	0.800	−0.737	0.312	1	0.804	0.010	0.141	0
12	0.823	−1.158	0.224	0	0.138	−0.007	0.135	−0.007
13	0.611	−0.294	0.246	1	0.669	0.017	0.151	0
14	0.965	−0.856	0.225	1	0.835	0.008	0.159	0
15	1.052	−0.833	0.155	1	0.830	0.009	0.168	0
16	0.937	−0.613	0.166	1	0.756	0.012	0.180	0
17	0.894	−0.151	0.456	1	0.747	0.013	0.193	0
18	0.720	−0.614	0.327	0	0.227	−0.011	0.181	−0.011
19	0.686	−0.070	0.112	0	0.442	−0.022	0.159	−0.033
20	0.608	−0.806	0.169	1	0.737	0.013	0.173	−0.020

Note: u = item scores, P = probability of endorsing an item given $\hat{\theta}_n$, T =weighted difference between the observed and expected score, C^+ = maximum value of the CUSUM and C^- = minimum value of the CUSUM. The column shaded is the test-taker's response pattern.

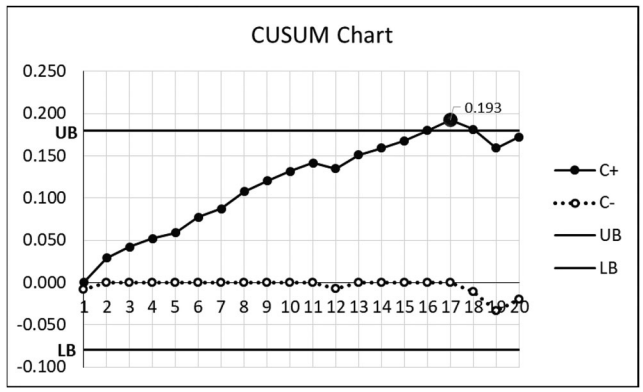

Figure 1. The CUSUM chart for a respondent with ability estimation: -0.06.

In Figure 1, the item number and the cumulative residual are represented by the horizontal and the vertical axis, respectively. Two horizontal lines show the upper bound (0.18) and the lower bound (-0.08), respectively. The two bounds here were obtained based on a Monte Carlo simulation (details are provided in the section "Simulation Studies"). The solid dot curve exceeds the upper bound 0.18, which indicates that the respondent has some kind of aberrant behavior that leads to the aberrant upward shift, which may be possibly indicative of the test taker's item preknowledge.

Bradlow et al. (1998) first adopted CUSUM to detect outliers in a CAT. They were interested in four types of outliers, representing effect of warm-up, fatigue, sub-expertise, and lack-of-fit. Here sub-expertise may be caused by extra or pre-knowledge or training in certain content areas of the test or on some items. Lack-of-fit refers to those people with implausible response patterns, e.g., answering difficult items correctly but easy ones wrong. van Krimpen-Stoop and Meijer (2000) reviewed some person-fit research in the context of P&P tests, and proposed eight statistics to investigate person fit in CAT based on CUSUM. They conducted a simulation study to investigate the numerical values of the upper and lower thresholds for the statistics, and investigated the power of these statistics. The results showed that the bounds of most of the eight CUSUM procedures were stable across θ-values. van Krimpen-Stoop and Meijer (2001) introduced two strategies for the proposed Z-statistics, in which they used all administered items and the final ability $\hat{\theta}_n$ to classify each examinee in the first strategy, while dividing each item response pattern into disjoint subsets in the second strategy. The bound of the CUSUM procedure was obtained by solving Siegmund's approximation (Siegmund, 1985, pp. 24–30), and their results showed the proposed statistic had a relatively high detection rate for misfitting item-score patterns. Both in the context of paper-and-pencil (P&P) testing and computerized adaptive testing, van Krimpen-Stoop and Meijer (2002) evaluated the performances of the popular person-fit statistic $l_z^{(p)}$ (Drasgow et al., 1985) and a CUSUM statistic to detect person misfit with polytomous items. Results showed the detection rates for the CUSUM statistic were reasonably high. Meijer (2002) applied some CUSUM statistics to classify a response pattern as fitting or misfitting the underlying item response theory model in CAT based on empirical data, and the analysis showed that different types of misfit involved could be distinguished.

Shi (2007) proposed a likelihood-based statistic to detect the aberrant response patterns, and extended the statistics to model-free detection, which means that the statistics do not rely on any pre-specified item response function. Armstrong and Shi (2009a, 2009b) evaluated and compared some non-likelihood-based person-fit statistic with the likelihood-based statistic, and used a quadratic curve to represent the aberrant shifts: the aberrant upward shift and the aberrant downward shift. In these studies, the statistical critical values used for hypothesis testing were calculated empirically with Monte Carlo simulations. Based on empirical data, Egberink et al. (2010) considered applying the CUSUM procedure to detect the inconsistent item score patterns in a polytomous CAT. Armstrong and Kung (2011) developed a CUSUM statistic to identify aberrant behavior in a sequential, multiple-choice standardized examination.

In their research, the responses were taken as finite Poisson trials, the significance level was computed with Markov chains, and they examined the performance of the statistic based on both simulated and empirical data. Further, Tendeiro and Meijer (2012) extended the procedure of Armstrong and Shi (2009a) and presented the theoretical ground of the CUSUM statistic based on likelihood ratio. They proposed a different version of the statistic and compared the detection rates of some statistics based on simulation data. Tendeiro et al. (2013) applied several CUSUM statistics besides the likelihood-based CUSUM statistic to detect the inconsistencies in an unproctored internet test.

Twelve CUSUM statistics

Through literature review we identified 12 CUSUM statistics for detection of intra-individual change during test-taking process. As we mentioned before, most of them differ in their definition of T_j. Eight of the CUSUM indices were developed in van Krimpen-Stoop and Meijer (2000), and they all defined their T_j's as some variations of the difference between observed and model-implied responses, which are labeled T_{1j} to T_{8j} :

$$T_{1j} = \frac{1}{n}\left[u_j - P_j(\hat{\theta}_j)\right], \tag{4}$$

$$T_{2j} = T_{1j} \times \left\{P_j\left(\hat{\theta}_j\right)\left[1 - P_j(\hat{\theta}_j)\right]\right\}^{-\frac{1}{2}}, \tag{5}$$

$$T_{3j} = T_{1j} \times [I\left(\hat{\theta}_j\right)]^{-\frac{1}{2}}, \tag{6}$$

$$T_{4j} = T_{1j} \times \sqrt{j}, \tag{7}$$

$$T_{5j} = \frac{1}{n}\left[u_j - P_j(\hat{\theta}_n)\right], \tag{8}$$

$$T_{6j} = T_{5j} \times \left\{P_j\left(\hat{\theta}_n\right)\left[1 - P_j(\hat{\theta}_n)\right]\right\}^{-\frac{1}{2}}, \tag{9}$$

$$T_{7j} = T_{5j} \times [I\left(\hat{\theta}_j\right)]^{-\frac{1}{2}}, \tag{10}$$

$$T_{8j} = T_{5j} \times \sqrt{j}, \tag{11}$$

where n is the test length, and j denotes the j^{th} item in the test, $j > 1$. $\hat{\theta}_j$ is the estimated θ based on the first j answered items, and $\hat{\theta}_n$ refers to the θ estimate based on the whole test. Without loss of generality, the subscript of the examinee is omitted. $I(\hat{\theta}_j)$ and $I(\hat{\theta}_n)$ are the test information evaluated at $\hat{\theta}_j$ and $\hat{\theta}_n$, respectively. By plugging in a different T_j into Eqs 1 and 2, one obtains a different CUSUM procedure. Here we label them as C_{T_1}- C_{T_8}.

In addition to these eight statistics, for analyzing person fit in adaptive testing Sinharay (2016) considered four additional CUSUM statistics, C^{LR}, *LARD*, C^{T} and $C^{l_z^*}$. In our study, these four statistics are denoted as C_{T_9}, $C_{T_{10}}$, $C_{T_{11}}$ and $C_{T_{12}}$, respectively. The corresponding definitions of T_j are named as T_{9j}, T_{10j}, T_{11j}, T_{12j}. T_{9j} was first defined in Bradlow et al. (1998) as

$$T_{9j} = \frac{\left|\sum_{j=1}^{n}\left[u_j - P_j(\hat{\theta}_j)\right]\right|}{\sqrt{\sum_{j=1}^{n} P_j\left(\hat{\theta}_j\right)\left(1 - P_j(\hat{\theta}_j)\right)}}. \tag{12}$$

This statistic in form has some resemblance to T_{2j} and T_{6j}.

Armstrong and Shi (2009a) suggested a statistic based on the CUSUM and a likelihood ratio statistic. Their statistic defines T_j differently in C^+ and C^-, and the associated two T_j definitions are named as T_{10j}^U and T_{10j}^L, which denote two likelihood ratio statistics for testing whether there is an "aberrant upward shift" and "aberrant downward shift" of the probability of a correct answer, respectively:

$$T_{10j}^U = ln\frac{g_j^U\left(p_j(\theta)\right)}{p_j(\theta)}, \tag{13}$$

$$T_{10j}^L = ln\frac{g_j^L\left(p_j(\theta)\right)}{p_j(\theta)}, \tag{14}$$

where $g_j^U\left(p_j(\theta)\right)$ and $g_j^L\left(p_j(\theta)\right)$ are two continuous curves that match an aberrant upward shift and an aberrant downward shift, respectively. Details for obtaining $g_j^U\left(p_j(\theta)\right)$ and $g_j^L\left(p_j(\theta)\right)$ can be found in Armstrong and Shi (2009a).

Based on the definition of T_{1j}, Sinharay (2016) suggested another CUSUM statistic, which is referred to as $C_{T_{11}}$ in this article. Note that C_{T_1} is also defined on T_{1j} – it is obtained by plugging T_{1j} into Eqs 1 and 2. Meanwhile, $C_{T_{11}}$ is obtained by plugging T_{1j} into Eq. 19 (see below).van Krimpen-Stoop and Meijer (2001) also suggested a CUSUM statistic based on the l_Z^* statistic (Snijders, 2001), which is denoted as T_{12j} as follows:

$$T_{12j}^U = l_Z^* + 0.5, \tag{15}$$

$$T_{12j}^L = l_Z^* - 0.5, \tag{16}$$

where l_Z^* is a variation of the person-fit statistics l_z proposed in Snijders (2001).

Different from the first eight CUSUM statistics, C_{T_9}, $C_{T_{10}}$, $C_{T_{11}}$ and $C_{T_{12}}$ take different forms, which are provided as follows:

$$C_{T_9} = \max_{1\le j\le n} T_{9j}, \tag{17}$$

$$C_{T_{10}} = \max_{1\le j\le n}\left(C^{+}_{T_{10,j}}\right) - \min_{1\le j\le n}\left(C^{-}_{T_{10,j}}\right), \tag{18}$$

where $C^{+}_{T_{10,j}} = \max_{1\le j\le n}\left(0, C^{+}_{T_{10,j-1}} + T^{U}_{10j}\right)$ and $C^{-}_{T_{10,j}} = \min_{1\le j\le n}\left(0, C^{-}_{T_{10,j-1}} + T^{L}_{10j}\right)$.

$$C_{T_{11}} = \max_{1\le j\le n}\left(C^{+}_{T_{11,j}}\right) - \min_{1\le j\le n}\left(C^{-}_{T_{11,j}}\right), \tag{19}$$

where $C^{+}_{T_{11,j}}$ and $C^{-}_{T_{11,j}}$ can be obtained based on the general formulas of CUSUM, that is, $C^{+}_{T_{11,j}} = max\{0, T_{11j} + C^{+}_{T_{11,j-1}}\}$, $C^{-}_{T_{11,j}} = min\left\{0, T_{11j} + C^{-}_{T_{11,j-1}}\right\}$. Note that $C_{T_{11}}$ only has one single value which measures the dispersion of an individual's performance. In contrast, C_{T_1} has the typical form of a CUSUM statistic, which has two values at the point j: $C^{+}_{T_{1,j}}$ and $C^{-}_{T_{1,j}}$.

Similarly, $C_{T_{12}}$ is defined as follows:

$$C_{T_{12}} = \max_{1\le j\le n}\left(C^{+}_{T_{12,j}}\right) - \min_{1\le j\le n}\left(C^{-}_{T_{12,j}}\right), \tag{20}$$

where $C^{+}_{T_{12,j}} = \max_{1\le j\le n}\left(0, T^{U}_{12j} + C^{+}_{T_{12,j-1}}\right)$, and $C^{-}_{T_{12,j}} = \min_{1\le j\le n}\left(0, T^{L}_{12j} + C^{-}_{T_{12,j-1}}\right)$.

Note that C_{T_1}- C_{T_8} involve a positive or negative residual for each item; that is, a large (in absolute value) CUSUM statistic suggests an aberrant response pattern. However, each of the last four CUSUM statistics (C_{T_9} - $C_{T_{12}}$) only involves a single residual, by which C_{T_9} measures the largest absolute realized deviation, while $C_{T_{10}}$, $C_{T_{11}}$ and $C_{T_{12}}$ measure the dispersion of the individual's performance during the test.

CPA and previous studies

Recently there have been a handful of studies that use CPA to detect intra-individual change during a test-taking process. Shao et al. (2016) used CPA to classify examinees into speeded and non-speeded groups, and estimated the point at which an examinee starts to speed. They conducted a simulation study to evaluate the performance of the CPA to detect the speededness when the speededness mechanism followed the GCM. Results showed that the CPA is efficient in detecting both speeded examinees and the speeding point. They adopted a permutation method to generate the null distribution of the CPA test statistic. Shao (2016) investigated whether CPA can help improve item calibration in the presence of speededness, applied the CPA method to warm-up effect detection, and conducted the CPA to detect the speededness based on response-time data. Sinharay (2016) suggested three statistics based on CPA to detect any abrupt changes during CAT, and compared the performances of the new statistics with four aforementioned CUSUM statistics (i.e., C_{T_9} - $C_{T_{12}}$), and used the asymptotic critical values for the significance level of the null distribution. All these studies used some or all of the three statistics that are introduced below.

Three CPA statistics

The Wald Test. Splitting a test into two subtests, subtest1 and subtest2 at point j, the Wald test can be used to detect whether a respondent's ability (denoted by θ) has changed from subtest1 to subtest2. The corresponding Wald test statistic can be formulated as

$$W_j = \frac{\left(\hat{\theta}_{1j} - \hat{\theta}_{2j}\right)^2}{\frac{1}{I_{1j}(\hat{\theta}_n)} + \frac{1}{I_{2j}(\hat{\theta}_n)}}, \tag{21}$$

where $\hat{\theta}_{1j}$ and $\hat{\theta}_{2j}$ denote the maximum likelihood estimate of ability based on subtest1 (from item 1 to item j) and subtest2 (from item $j+1$ to n), respectively. $I_{1j}(\hat{\theta}_n)$ and $I_{2j}(\hat{\theta}_n)$ are the test information for subtest1 and subtest2, respectively, both evaluated at $\hat{\theta}_n$, which is the ability estimate based on the whole test.

In practice, the change point is unknown, and there are (n-1) possible change positions. Therefore, the corresponding statistic is taken as the maximum over all possible change points, i.e.,

$$W_{max} = \max_{1\le j\le n-1}\{W_j\}. \tag{22}$$

Any respondent will be flagged as aberrant if his or her associated W_{max} is significantly larger than 0, which signals a significant change in the underlying θ in the test taking process. The change point is estimated to be the point where W_{max} is achieved.

The likelihood ratio test

According to Shao et al. (2016) and Sinharay (2016), the likelihood ratio test statistic can be formulated as follows:

$$\Delta l_j = -2\Big\{l\left(\hat{\theta}_n; Y_1, Y_2, \ldots, Y_n\right) - \Big[l\left(\hat{\theta}_{1j}; Y_1, Y_2, \ldots, Y_j\right) + l\left(\hat{\theta}_{2j}; Y_{j+1}, Y_{j+2}, \ldots, Y_n\right)\Big]\Big\}, \tag{23}$$

where $l\left(\hat{\theta}_n; Y_1, Y_2, \ldots, Y_n\right)$ refers to the log-likelihood for the whole test evaluated at $\hat{\theta}_n$, while

Table 2. The summaries of the fifteen statistics.

Statistic	Mathematical definition	Source of reference
W	$\frac{\left(\hat{\theta}_{1j}-\hat{\theta}_{2j}\right)^2}{\frac{1}{I_{1j}(\hat{\theta}_n)}+\frac{1}{I_{2j}(\hat{\theta}_n)}}$	Sinharay (2016)
L	$-2\left\{l(\hat{\theta}_n; Y_1, Y_2, ..., Y_n) - \left[l\left(\hat{\theta}_{1j}; Y_1, Y_2, ..., Y_j\right) + l\left(\hat{\theta}_{2j}; Y_{j+1}, Y_{j+2}, ..., Y_n\right)\right]\right\}$	Shao et al. (2016) Sinharay (2016, 2017a)
S	$\frac{\left(\nabla\left(\hat{\theta}_n; Y_1, Y_2, ..., Y_j\right)\right)^2}{I_{1j}(\hat{\theta}_n)} + \frac{\left(\nabla\left(\hat{\theta}_n; Y_{j+1}, Y_{j+2}, ..., Y_n\right)\right)^2}{I_{2j}(\hat{\theta}_n)}$	Sinharay (2016, 2017a)
C_{T_1}	$T_1 = \frac{1}{n}\left[u_j - P_j(\hat{\theta}_j)\right]$	van Krimpen-Stoop and Meijer (2000)
C_{T_2}	$T_2 = T_{1j} \times \left\{P_j(\hat{\theta}_j)\left[1 - P_j(\hat{\theta}_j)\right]\right\}^{-\frac{1}{2}}$	
C_{T_3}	$T_3 = T_{1j} \times [I\left(\hat{\theta}_j\right)]^{-\frac{1}{2}}$	
C_{T_4}	$T_4 = T_{1j} \times \sqrt{j}$	
C_{T_5}	$T_5 = \frac{1}{n}\left[u_j - P_j(\hat{\theta}_n)\right]$	
C_{T_6}	$T_6 = T_{5j} \times \left\{P_j(\hat{\theta}_n)\left[1 - P_j(\hat{\theta}_n)\right]\right\}^{-\frac{1}{2}}$	
C_{T_7}	$T_7 = T_{5j} \times [I\left(\hat{\theta}_j\right)]^{-\frac{1}{2}}$	
C_{T_8}	$T_8 = T_{5j} \times \sqrt{j}$	
C_{T_9}	$T_9 = \frac{\left\lvert\sum_{j=1}^{n}\left[u_j - P_j(\hat{\theta}_j)\right]\right\rvert}{\sqrt{\sum_{j=1}^{n} P_j(\hat{\theta}_j)\left(1-P_j(\hat{\theta}_j)\right)}}$	Bradlow et al. (1998); Sinharay (2016)
$C_{T_{10}}$	$T^U_{10j} = ln\frac{g^U_j(p_j(\theta))}{p_j(\theta)},\ T^L_{10j} = ln\frac{g^L_j(p_j(\theta))}{p_j(\theta)}$	Armstrong and Shi (2009a) Sinharay (2016)
$C_{T_{11}}$	$\frac{1}{n}\left[u_j - P_j(\hat{\theta}_j)\right]/\lambda_i$	Sinharay (2016)
$C_{T_{12}}$	$T^U_{12j} = I^*_Z + 0.5,\ T^L_{12j} = I^*_Z - 0.5$	Sinharay (2016); van Krimpen-Stoop and Meijer (2001)

Note: The first three shaded entries are CPA-based statistics, and the rest are the CUSUM-based statistics. For the first eight CUSUM-based statistics, C_{T_1}-C_{T_8}, they use the general formula of CUSUM statistics, which are shown in Eqs. 1–3, while the latter four take different forms, which are provided in Eqs. 17–20.

$l\left(\hat{\theta}_{1j}; Y_1, Y_2, ..., Y_j\right) + l\left(\hat{\theta}_{2j}; Y_{j+1}, Y_{j+2}, ..., Y_n\right)$ is the alternative likelihood that allows separate θ estimates before and after item j. As encountered previously, the test statistic is taken as the maximum of the likelihood-ratio test statistic due to the unknown change position in a real test:

$$L_{max} = \max_{1\le j\le n-1}\left\{\Delta l_j\right\}. \tag{24}$$

Similar to W_{max}, if L_{max} is significantly larger than 0, the respondent will be flagged as aberrant with the change point estimated to be the point where L_{max} is achieved.

The score test

The score test can also be used to test the significant difference in ability between subtest1 and subtest2. It is defined as follows:

$$S_j = \frac{\left(\nabla\left(\hat{\theta}_n; Y_1, Y_2, ..., Y_j\right)\right)^2}{I_{1j}(\hat{\theta}_n)} + \frac{\left(\nabla\left(\hat{\theta}_n; Y_{j+1}, Y_{j+2}, ..., Y_n\right)\right)^2}{I_{2j}(\hat{\theta}_n)}, \tag{25}$$

where $\nabla\left(\hat{\theta}_0; Y_1, Y_2, ..., Y_j\right)$ and $\nabla\left(\hat{\theta}_0; Y_{j+1}, Y_{j+2}, ..., Y_n\right)$ are the first-order derivatives with respect to θ of the log likelihood of subtest1 and subtest2, respectively. The test statistic is the maximum of the S_j :

$$S_{max} = \max_{1\le j\le n-1}\left\{S_j\right\}. \tag{26}$$

Similar to W_{max} and L_{max}, a response pattern will be flagged if S_{max} is significantly larger than 0, and the change point is estimated to be the point where S_{max} is achieved.

Shao et al. (2016) proposed to use the likelihood ratio test and to obtain critical values for this test statistic through permutation of the data. Sinharay (2016) suggested that it is possible to use the asymptotic critical values provided in Andrews (1993), and that the asymptotic critical values for the three statistics are the same. In this study, we propose to use Monte Carlo simulations to obtain critical values as done in Worsley (1979) and Shao (2016). This is because permutation can be computationally inefficient for long tests, and the asymptotic values may not be applicable if the change point appears early or late on a test (Andrews, 1993; Sinharay, 2016). They are not

Table 3. The parameters setting for simulation studies.

Type	Parameter	Distribution			
			Condition	α	β
GCM parameters	η	Beta distribution (α, β)	C1	146.345	62.910
			C2	14.048	6.211
			C3	3.033	1.490
			C4	143.367	95.689
			C5	13.768	9.290
			C6	2.970	2.091
			C7	124.500	124.500
			C8	12.000	12.000
			C9	2.625	2.625
	λ	Log normal (3.912,1)			
Item parameters	a	Log normal (0,0.05)			
	b	Normal (0,1)			
Ability parameters	θ	Normal (0,1)			

Note. C1 means Condition 1 ($\eta_{median} = 0.7$, $\eta_{variance} = 0.001$), C2 means Condition 2 ($\eta_{median} = 0.7$, $\eta_{variance} = 0.01$). The settings of item parameters and ability parameters are common in GCM and HM.

applicable for CUSUM statistics, either. Hence all critical values in this study are obtained through Monte Carlo simulations. All the fifteen statistics of interest in this article are summarized in Table 2.

Simulation studies

Different statistics may be sensitive to different types of aberrant response behaviors. There is no known method that can always guarantee identification of all types of aberrant item response patterns with a high success rate (Tendeiro & Meijer, 2012). The focus of this study is test speededness. The same as the above literature, we focus on dichotomous items with known parameters from the IRT model that the regular response data are assumed to follow.

Underlying mechanism of speededness

As stated previously, we consider two models with distinct assumptions regarding the underlying mechanism of speededness, GCM and HM, the former capturing gradual change while the latter abrupt change. The HM assumes that all responses after the speeding point will be completely random which is a rather strict assumption. The GCM, on the other hand, considers that each speeded examinee has a unique speeding point, and the probabilities for answering those speeded questions will gradually decrease toward the end of the test.

The HM models examinees who randomly guess on some items, typically items toward the end of the test in the context of speededness. The 2PL version of the HM takes the following form:

$$P_{ij}^{*} = \begin{cases} P\left(u_{ij}=1|\theta_i,\gamma_j\right) & ,before\ the\ change\ point \\ r & ,\ after\ the\ change\ point \end{cases}, \tag{27}$$

where $P\left(u_{ij}=1|\theta_i,\gamma_j\right)$ is the ordinary 2PL IRT model, and r is the random guessing probability, which is independent to the examinee's true ability, and often assumed to equal the reciprocal of the number of response options of a multiple-choice item. γ_j refers to the item parameters for item j. In the case of 2PL, $\gamma_j = \left(a_j,\ b_j\right)'$, where a_j, b_j are the item discrimination and difficulty parameter for item j, respectively. $P\left(u_{ij}=1|\theta_i,\gamma_j\right)$, the response function of the 2PL IRT model takes the following form:

$$P\left(u_{ij}=1|\theta_i,\gamma_j\right) = \frac{1}{1+e^{-1.7a_j\left(\theta_i-b_j\right)}}, \tag{28}$$

The GCM was first proposed by Wollack and Cohen (2004) to generate speeded responses. Goegebeur et al. (2008, 2010) showed how to estimate the model parameters, and evaluated the model fit by comparing the model-implied proportions of the correct responses against the observed proportions. Suh et al. (2012) fitted the GCM to classify examinees into speeded and non-speeded classes. The 2PL version of the GCM takes the following form:

$$P_{ij}^{*} = P\left(u_{ij}=1|\theta_i,\gamma_j\ \right) \times min\left(1,\left[1-\left(\frac{j}{n}-\eta_i\right)\right]\right)^{\lambda_i}, \tag{29}$$

where the speeding point is modeled by the parameter $(0 \le \eta_i \le 1)$, and the speededness rate is modeled by the parameter λ_i $(\lambda_i \ge 0)$. Because of speededness, a larger η_i means a later speeding point which translates to less severe speededness, and a larger λ_i suggests a faster decline of P_{ij}^{*}.

Simulation design

For the comparison of speededness detection based on CUSUM and CPA under different situations, three

Table 4. Empirical critical values for CPA-based statistics.

	Critical Value		
Test length	L	W	S
40	8.087	27.194	6.978
60	8.481	32.979	8.972
80	8.874	58.898	13.13

simulation studies were conducted to evaluate the performance of 15 statistics, three based on CPA: W_{max}, L_{max}, S_{max}; and twelve based on CUSUM: C_{T_1} - $C_{T_{12}}$. In the first two studies, the lower and upper bounds of the CUSUM and the critical values of the CPA statistics are obtained by Monte Carlo simulations based on 10,000 normal response patterns. In the third study (presented in the appendix), we use existing asymptotic critical values for the CPA statistics. In each study, two hundred data sets, each consisting of 10,000 normal score patterns, were generated based on the 2PL IRT model. The item and ability parameters were generated based on the distributions summarized in Table 3.

For each response pattern, twelve different CUSUM statistics plus three CPA statistics were calculated. Detection of speededness was done using a 5% nominal type-I error level for each statistic. That is, the statistics were compared to the critical values or upper/lower bounds that were obtained at a 5% nominal type-I error level. For the three CPA statistics and the last four CUSUM statistics (C_{T_9}- $C_{T_{12}}$), 10,000 values for each statistic were sorted in descending order. The corresponding values of the bound and critical values were taken to be the value in the 500^{th} position. For the first eight CUSUM statistics (C_{T_1}- C_{T_8}), the lower values of the bound were taken to be in the $9,750^{th}$ position, and the upper values of the bound were taken to be in the 250^{th} position. All the bound or critical values were shown in Tables 4 and 5.

Table 4 presents empirical critical values for the three CPA-based statistics. The critical values for L_{max} and S_{max} show smaller fluctuation than the W_{max}, and they are closer to the asymptotic critical values provided in Sinharay (2016). Table 5 provides the values of the bound for the twelve CUSUM-based statistics. For the first eight CUSUM statistics C_{T_1} - C_{T_8}, the values of UB and LB are almost symmetric around 0. Because the alternative hypothesis of interest for C_{T_9}, $C_{T_{10}}$, $C_{T_{11}}$, $C_{T_{12}}$ is one-sided, the lower values of the bound for these four statistics were left blank in Table 5.

Item response data affected by speednedness are generated following the HM and the GCM. As described in Shao (2016), examinees will be classified into two groups, speeded vs. non-speeded under the mixture-modeling framework. In all three studies in this paper, different prevalence and severity rates of speededness were considered. Prevalence refers to the proportion of examinees who have exhibited speeded response behaviors in their tests, and severity refers to the proportion of items affected by speededness of those examinees with speeded behaviors, which is regulated by the speeding point. In Study 1, speededness detection based on Monte Carlo critical values under HM was conducted. Different from Study 1, the GCM was adopted in Study 2[1]. The maximum likelihood estimation (MLE; Baker & Kim, 2004) algorithm was used to obtain the interim θ estimates for each respondent during the test. Respondents with all correct or all wrong answers were assigned ability estimates 3 or −3.

Table 5. Empirical values of the bound for CUSUM-based statistics.

	40 items		60 items		80 items	
Statistic	LB	UB	LB	UB	LB	UB
C_{T_1}	−0.131	0.127	−0.110	0.110	−0.097	0.097
C_{T_2}	−0.329	0.322	−0.283	0.276	−0.251	0.244
C_{T_3}	−0.046	0.046	−0.030	0.032	−0.029	0.030
C_{T_4}	−0.593	0.582	−0.626	0.632	−0.643	0.636
C_{T_5}	−0.103	0.101	−0.087	0.086	−0.076	0.074
C_{T_6}	−0.271	0.274	−0.232	0.227	−0.200	0.195
C_{T_7}	−0.022	0.023	−0.016	0.016	−0.013	0.013
C_{T_8}	−0.479	0.467	−0.508	0.504	−0.526	0.513
C_{T_9}		*2.517*		*2.878*		*3.524*
$C_{T_{10}}$		*2.545*		*2.790*		*2.564*
$C_{T_{11}}$		*20.558*		*39.109*		*53.970*
$C_{T_{12}}$		*0.170*		*0.144*		*0.127*

Note. All the values are based on the average of 200 replications.

Study 1

The first study used the HM as the true model to generate item responses affected by speededness. More specifically, respondents have a probability of .2 to respond correctly to items that are affected by speededness, or r is set to .2 in Eq. 27. For the respondents without speededness, their response patterns were generated based on the ordinary 2PL IRT model. Then we applied the 12 CUSUM procedures and 3 CPA statistics to detect the speededness with empirically derived critical values. For the prevalence of speededness, 10% and 30% of the examinees were randomly chosen to be affected by speededness. The last 30%, 40%, and 50% of the items in the test, which are referred to as low, medium, and high level of severity

[1]We also conducted a Study 3 where the performance of the three CPA statistics was evaluated based on asymptotic critical values (see Table 2 of Sinharay, 2016) instead of empirically derived critical values under HM and GCM. Everything was kept the same as in Studies 1 and 2 except for the critical values. Results of Study 3 are summarized in the Appendix.

of speededness, were chosen to be answered aberrantly for those affected by speededness. The simulated sample consisted of 1,000 respondents with their abilities randomly generated from the standard normal distribution $N(0,1)$. 40, 60 and 80 pairs of 2PL item parameters were generated with the distributions specified in Table 3. Based on the simulated parameters and the HM, the response matrix can be generated. Overall there are 18 conditions: 3 test lengths (40, 60 and 80)× 2 prevalence rates (10% and 30%)× 3 severity rates (30%, 40%, and 50%). Each condition was replicated 200 times. For every response pattern in each dataset, the values for the 15 statistics can be calculated based on the corresponding formulas and compared to their respective bound or critical values.

Study 2

Different from the Study 1, Study 2 generated data from the GCM. Everything else is kept the same as Study 1. The same distributions of η, λ as in Shao et al. (2016) were used in this study (see Table 3). η determines the change point at which an examinee starts his/her speededness behavior – this dictates the severity level. Its distribution was manipulated through a Beta distribution in a 3×3 design, with three levels of the median of η (0.7, 0.6 and 0.5), and three levels of variance (0.001, 10×0.001, 40×0.001) (see the corresponding α and β parameter values of the Beta distribution in Table 3). This resulted in nine different η conditions denoted as C1-C9. The Same as in study 1, we included three levels of test length: 40, 60 and 80, the sample size was fixed at 1,000, and true abilities were generated following the standard normal distribution. In total there were 54 conditions: 3 test lengths (40, 60 and 80)×3 η medians (0.7, 0.6 and 0.5) ×3 η variances (original, 10×, and 40×)×2 prevalence levels (10% and 30%). For the respondents without speededness, their response patterns were generated based on the ordinary 2PL IRT model. Each condition was replicated 200 times. The same as Study 1, values for the 15 statistics for each respondent was calculated and compared to their respective bound/critical values.

Evaluative measures

While obtaining the values of the 15 statistics for the response pattern of each respondent, the correct classification rates (CCR), type I error and power rates were obtained by comparing their values with the associated cutoffs presented in Tables 4 and 5, where CCR or the hit rate refers to the chance of classifying each respondent based on his/her behavior correctly into the speeded or non-speeded group. Correct classification occurs when a respondent is classified into the group to which he or she truly belongs. All results of Studies 1 and 2 are summarized in Tables 6–13, while results for Study 3 are presented in Tables A1 and A2 in the appendix.

Results

The results are presented and organized based on the generating model. Two levels (10% and 30%) for speededness prevalence were simulated for both the HM and GCM. Under the HM, three tables were created for results for test lengths of 40, 60, and 80 with three levels of speededness severity (30%, 40%, and 50%), respectively. Under the GCM, there are nine tables (nine conditions for speededness severity, C1-C9) for each test length. Due to the space limit and the large number of conditions, only results of the high-severity conditions, C7-C9, were presented in this paper. Results of the low-severity conditions (C1-C3) and the middle-severity conditions (C4-C6) show the same trend as C7 – C9 except with generally lower power. These results are available from the authors upon request.

Generating model: HM

Tables 6–8 showed the results of speededness detection under different levels of prevalence and severity when the speeded responses were generated based on HM. The outputs were generated based on the fifteen statistics: W_{max}, L_{max}, S_{max}, and C_{T_1}- $C_{T_{12}}$. Considering power, L_{max} and S_{max} always perform better than W_{max}, which has the smallest detection rate. For example, under HM, 40 items, 10% speededness prevalence, 30% speededness severity, the CCR of W_{max}, L_{max}, S_{max} are 0.872, 0.918 and 0.912, respectively (see Table 6). The power for the three statistics are: 0.202, 0.628 and 0.556. The CCR is generally high in spite of the low power because the speeded prevalence (i.e., the base rate) is low. The type I error rates for the CPA-based statistics are always close to the corresponding nominal level. Meanwhile, among 12 CUSUM statistics, C_{T_2}, C_{T_6} and C_{T_9} perform better in the 40-item and 60-item tests in terms of power, while in the 80-item test, C_{T_6}, C_{T_7}, and C_{T_9} are the top three. The same as CPA-based statistics, the type I error rates for the 12 CUSUM-based statistics are also close to the nominal level. For each statistic of

Table 6. Results based on HM, speededness severity = 30%, speededness prevalence = 10%, 30%.

			40 Items			60 Items			80 Items		
Severity	Prevalence	Statistics	CCR	Power	Type-I Error	CCR	Power	Type-I Error	CCR	Power	Type-I Error
30%	10%	W	0.872	0.202	0.054	0.894	0.396	0.051	0.899	0.470	0.054
		L	0.918	0.628	0.050	0.921	0.664	0.051	0.930	0.741	0.049
		S	0.912	0.556	0.049	0.915	0.585	0.048	0.931	0.741	0.047
		C_{T_1}	0.897	0.430	0.051	0.908	0.524	0.049	0.926	0.729	0.052
		C_{T_2}	0.925	**0.701**	0.050	0.928	**0.734**	0.051	0.926	0.735	0.053
		C_{T_3}	0.878	0.225	0.049	0.889	0.345	0.051	0.901	0.469	0.05
		C_{T_4}	0.906	0.514	0.050	0.915	0.588	0.049	0.933	0.775	0.05
		C_{T_5}	0.903	0.477	0.050	0.913	0.576	0.049	0.933	0.748	0.047
		C_{T_6}	0.930	**0.746**	0.050	0.930	**0.749**	0.050	0.931	**0.788**	0.053
		C_{T_7}	0.899	0.427	0.049	0.910	0.556	0.051	0.94	**0.797**	0.044
		C_{T_8}	0.911	0.545	0.049	0.916	0.602	0.049	0.934	0.786	0.05
		C_{T_9}	0.934	**0.767**	0.048	0.932	**0.803**	0.054	0.937	**0.826**	0.050
		$C_{T_{10}}$	0.882	0.235	0.046	0.898	0.427	0.050	0.901	0.430	0.047
		$C_{T_{11}}$	0.868	0.147	0.052	0.877	0.237	0.052	0.874	0.193	0.051
		$C_{T_{12}}$	0.907	0.527	0.051	0.915	0.600	0.050	0.931	0.750	0.049
	30%	W	0.719	0.194	0.055	0.778	0.383	0.053	0.800	0.460	0.055
		L	0.850	0.618	0.051	0.863	0.663	0.051	0.888	0.746	0.052
		S	0.831	0.546	0.047	0.838	0.569	0.047	0.890	0.741	0.047
		C_{T_1}	0.791	0.424	0.051	0.818	0.513	0.051	0.886	0.733	0.049
		C_{T_2}	0.867	**0.675**	0.050	0.882	**0.730**	0.053	0.886	0.73	0.047
		C_{T_3}	0.728	0.214	0.051	0.763	0.323	0.048	0.807	0.457	0.044
		C_{T_4}	0.819	0.516	0.051	0.839	0.583	0.050	0.899	0.777	0.048
		C_{T_5}	0.807	0.466	0.047	0.838	0.572	0.048	0.894	0.751	0.045
		C_{T_6}	0.880	**0.721**	0.051	0.892	**0.745**	0.045	0.9	**0.782**	0.05
		C_{T_7}	0.798	0.429	0.044	0.836	0.556	0.044	0.91	**0.793**	0.04
		C_{T_8}	0.827	0.539	0.050	0.845	0.595	0.048	0.895	0.753	0.045
		C_{T_9}	0.890	**0.743**	0.047	0.902	**0.793**	0.052	0.906	**0.808**	0.052
		$C_{T_{10}}$	0.733	0.224	0.049	0.794	0.427	0.048	0.788	0.407	0.049
		$C_{T_{11}}$	0.703	0.141	0.056	0.732	0.225	0.051	0.723	0.195	0.051
		$C_{T_{12}}$	0.819	0.514	0.050	0.846	0.607	0.051	0.893	0.761	0.050

Note: W, L, and S refer to the statistics of Wald test, likelihood ratio test, and score test, respectively. C_{T_1} - $C_{T_{12}}$ refer to the twelve CUSUM-based statistics.

interest, with the increase of test length or speededness severity, the power generally increases. We also conducted a simulation with lower speededness severity (10% and 20%), and the power for some statistics decreased quickly. This indicates that more serious speeded response behavior can be more easily detected. For the comparison of all methods, at least three CUSUM-based statistics (e.g., C_{T_2}, C_{T_6}, C_{T_9} for 40- and 60-item tests, and C_{T_6}, C_{T_7}, C_{T_9} for 80-item test) outperformed the three CPA-based statistics. This indicates that CUSUM-based statistics are favored when the respondents affected by speededness apply the random response strategy to the end of the test, which is consistent with the expectation that CUSUM-based methods are more powerful when the parameters before/after the change is known. In this study, the probability of answering an item correctly after the change point under HM is known to be .2. In general, some CUSUM procedures such as C_{T_9} consistently shows a promising power rate at .75 or above across a variety of conditions in Tables 6–8.

Generating model: GCM

Compared to HM, most statistics have higher detection rates under the GCM. Tables 9–11 present the corresponding results under the GCM, high-severity conditions (C7-C9) with three test lengths, respectively. Under GCM, for the twelve CUSUM-based statistics, C_{T_7} is the most powerful. For the three CPA-based statistics, the same as under HM, L_{max}, S_{max} outperform W_{max}. For all statistics, the type I error rates are well controlled. It also can be seen that the test length has a positive effect on power. Nine conditions for the speededness severity, C1-C9, can be binned into three groups based on the median or the variance of the starting position of speededness. Given a fixed median of the speededness starting position, the power will decrease with the increase of the position variance. Given a variance of the speededness starting position, the power will decrease with the increase of the speed position median. In general, L_{max}, S_{max}, and C_{T_7} are the top three statistics in terms of power in the 40-item and 60-item test, while S_{max}, C_{T_7}, and C_{T_9} are the top three for the 80-item test. This shows clear contrast with results under the HM, where none of the CPA statistics lead to comparable performance to the top CUSUM procedures, corroborating with previous findings that CPA is more helpful when the parameters before and/or after the change point are unknown.

Table 7. Results based on HM, speededness severity = 40%, speededness prevalence = 10%, 30%.

			40 Items			60 Items			80 Items		
Severity	Prevalence	Statistics	CCR	Power	Type-I Error	CCR	Power	Type-I Error	CCR	Power	Type-I Error
40%	10%	W	0.872	0.206	0.054	0.894	0.397	0.051	0.905	0.532	0.054
		L	0.919	0.636	0.049	0.923	0.670	0.049	0.931	0.761	0.050
		S	0.915	0.594	0.049	0.917	0.604	0.048	0.930	0.741	0.049
		C_{T_1}	0.906	0.520	0.051	0.914	0.582	0.049	0.932	0.763	0.049
		C_{T_2}	0.926	**0.698**	0.049	0.929	**0.750**	0.051	0.93	0.779	0.053
		C_{T_3}	0.887	0.322	0.050	0.901	0.449	0.049	0.914	0.576	0.048
		C_{T_4}	0.911	0.584	0.053	0.919	0.629	0.049	0.936	0.811	0.05
		C_{T_5}	0.909	0.524	0.048	0.916	0.590	0.048	0.936	0.788	0.047
		C_{T_6}	0.928	**0.741**	0.051	0.937	**0.801**	0.048	0.937	**0.842**	0.053
		C_{T_7}	0.913	0.540	0.046	0.917	0.590	0.047	0.945	**0.843**	0.043
		C_{T_8}	0.913	0.596	0.052	0.918	0.624	0.050	0.936	0.814	0.05
		C_{T_9}	0.932	**0.775**	0.050	0.937	**0.814**	0.050	0.943	**0.873**	0.050
		$C_{T_{10}}$	0.890	0.319	0.046	0.907	0.504	0.049	0.909	0.546	0.050
		$C_{T_{11}}$	0.879	0.266	0.053	0.891	0.372	0.051	0.902	0.474	0.050
		$C_{T_{12}}$	0.909	0.550	0.051	0.918	0.636	0.051	0.933	0.748	0.046
	30%	W	0.720	0.195	0.055	0.778	0.384	0.052	0.818	0.517	0.053
		L	0.854	0.630	0.050	0.863	0.663	0.052	0.895	0.768	0.051
		S	0.842	0.584	0.048	0.844	0.590	0.047	0.891	0.743	0.046
		C_{T_1}	0.811	0.495	0.054	0.839	0.575	0.048	0.898	0.765	0.046
		C_{T_2}	0.874	**0.690**	0.048	0.882	**0.732**	0.054	0.903	0.785	0.047
		C_{T_3}	0.758	0.313	0.051	0.803	0.451	0.047	0.845	0.584	0.044
		C_{T_4}	0.830	0.559	0.054	0.854	0.625	0.048	0.909	0.806	0.047
		C_{T_5}	0.817	0.502	0.049	0.846	0.587	0.044	0.905	0.787	0.044
		C_{T_6}	0.889	**0.739**	0.047	0.899	**0.780**	0.050	0.92	**0.841**	0.046
		C_{T_7}	0.823	0.521	0.048	0.846	0.587	0.044	0.923	**0.842**	0.042
		C_{T_8}	0.835	0.564	0.049	0.853	0.615	0.046	0.913	0.82	0.047
		C_{T_9}	0.893	**0.759**	0.050	0.905	**0.809**	0.054	0.929	**0.877**	0.049
		$C_{T_{10}}$	0.758	0.316	0.053	0.812	0.490	0.049	0.829	0.553	0.052
		$C_{T_{11}}$	0.741	0.261	0.053	0.772	0.363	0.053	0.803	0.465	0.052
		$C_{T_{12}}$	0.821	0.539	0.057	0.851	0.627	0.053	0.898	0.774	0.048

Note: W, L, and S refer to the statistics of Wald test, likelihood ratio test, and score test, respectively. C_{T_1} - $C_{T_{12}}$ refer to the twelve CUSUM-based statistics.

Table 8. Results based on HM, speededness severity = 50%, speededness prevalence = 10%, 30%.

			40 Items			60 Items			80 Items		
Severity	Prevalence	Statistics	CCR	Power	Type-I Error	CCR	Power	Type-I Error	CCR	Power	Type-I Error
50%	10%	W	0.872	0.200	0.054	0.892	0.388	0.052	0.900	0.488	0.054
		L	0.920	0.647	0.049	0.923	0.683	0.050	0.931	0.759	0.050
		S	0.918	0.613	0.048	0.921	0.641	0.048	0.929	0.725	0.048
		C_{T_1}	0.908	0.551	0.052	0.914	0.613	0.052	0.931	0.763	0.051
		C_{T_2}	0.929	**0.757**	0.052	0.928	**0.768**	0.054	0.929	0.786	0.055
		C_{T_3}	0.898	0.424	0.050	0.909	0.543	0.050	0.92	0.655	0.051
		C_{T_4}	0.913	0.601	0.053	0.920	0.648	0.050	0.935	0.802	0.05
		C_{T_5}	0.909	0.539	0.050	0.918	0.623	0.049	0.935	0.784	0.049
		C_{T_6}	0.932	**0.802**	0.054	0.935	**0.814**	0.051	0.936	**0.840**	0.054
		C_{T_7}	0.915	0.590	0.048	0.922	0.638	0.047	0.945	**0.855**	0.045
		C_{T_8}	0.913	0.593	0.052	0.921	0.640	0.048	0.937	0.825	0.05
		C_{T_9}	0.938	**0.820**	0.049	0.941	**0.830**	0.047	0.945	**0.887**	0.049
		$C_{T_{10}}$	0.894	0.408	0.052	0.916	0.562	0.045	0.917	0.618	0.050
		$C_{T_{11}}$	0.889	0.369	0.053	0.902	0.471	0.050	0.906	0.497	0.049
		$C_{T_{12}}$	0.911	0.575	0.052	0.921	0.651	0.049	0.933	0.764	0.049
	30%	W	0.720	0.195	0.056	0.775	0.370	0.052	0.803	0.469	0.054
		L	0.856	0.635	0.050	0.868	0.677	0.050	0.894	0.763	0.051
		S	0.849	0.606	0.047	0.854	0.626	0.048	0.885	0.724	0.047
		C_{T_1}	0.827	0.536	0.049	0.847	0.606	0.050	0.893	0.769	0.054
		C_{T_2}	0.889	**0.746**	0.049	0.889	**0.747**	0.050	0.896	0.774	0.052
		C_{T_3}	0.784	0.395	0.049	0.824	0.526	0.048	0.857	0.635	0.048
		C_{T_4}	0.839	0.580	0.050	0.859	0.646	0.050	0.905	0.81	0.054
		C_{T_5}	0.821	0.520	0.050	0.853	0.613	0.045	0.903	0.789	0.048
		C_{T_6}	0.902	**0.788**	0.049	0.907	**0.809**	0.051	0.916	**0.843**	0.052
		C_{T_7}	0.840	0.571	0.045	0.858	0.629	0.044	0.926	**0.851**	0.042
		C_{T_8}	0.836	0.575	0.052	0.858	0.637	0.047	0.911	0.828	0.054
		C_{T_9}	0.910	**0.824**	0.054	0.913	**0.829**	0.051	0.934	**0.892**	0.048
		$C_{T_{10}}$	0.783	0.395	0.051	0.833	0.553	0.047	0.848	0.614	0.051
		$C_{T_{11}}$	0.766	0.347	0.054	0.802	0.461	0.051	0.807	0.481	0.053
		$C_{T_{12}}$	0.826	0.549	0.056	0.859	0.651	0.051	0.901	0.780	0.047

Note: W, L, and S refer to the statistics of Wald test, likelihood ratio test, and score test, respectively. C_{T_1} - $C_{T_{12}}$ refer to the twelve CUSUM-based statistics.

Table 9. Results based on GCM, speededness severity: C7, speededness prevalence = 10%, 30%.

			40 Items			60 Items			80 Items		
Severity	Prevalence	Statistics	CCR	Power	Type-I Error	CCR	Power	Type-I Error	CCR	Power	Type-I Error
C7	10%	W	0.907	0.561	0.054	0.913	0.594	0.052	0.921	0.687	0.053
		L	0.938	**0.823**	0.050	0.943	**0.892**	0.051	0.948	0.933	0.050
		S	0.943	**0.865**	0.049	0.949	**0.931**	0.049	0.956	**0.987**	0.048
		C_{T_1}	0.924	0.698	0.051	0.927	0.77	0.056	0.945	0.895	0.050
		C_{T_2}	0.926	0.708	0.050	0.927	0.767	0.055	0.943	0.888	0.051
		C_{T_3}	0.924	0.676	0.049	0.93	0.766	0.052	0.952	0.956	0.049
		C_{T_4}	0.928	0.740	0.051	0.932	0.799	0.053	0.947	0.913	0.050
		C_{T_5}	0.915	0.622	0.052	0.925	0.725	0.053	0.939	0.853	0.052
		C_{T_6}	0.922	0.684	0.051	0.928	0.752	0.052	0.942	0.899	0.054
		C_{T_7}	0.943	**0.874**	0.049	0.947	**0.904**	0.049	0.958	**1.000**	0.047
		C_{T_8}	0.920	0.671	0.052	0.926	0.753	0.054	0.942	0.891	0.052
		C_{T_9}	0.952	0.773	0.028	0.961	0.890	0.031	0.954	**0.975**	0.048
		$C_{T_{10}}$	0.931	0.745	0.048	0.935	0.769	0.047	0.941	0.866	0.050
		$C_{T_{11}}$	0.885	0.300	0.050	0.891	0.377	0.052	0.898	0.415	0.048
		$C_{T_{12}}$	0.916	0.652	0.054	0.934	0.768	0.048	0.944	0.870	0.048
	30%	W	0.823	0.544	0.057	0.838	0.584	0.053	0.873	0.700	0.054
		L	0.915	**0.832**	0.050	0.932	**0.892**	0.051	0.944	0.930	0.050
		S	0.929	**0.872**	0.047	0.948	**0.934**	0.046	0.963	**0.987**	0.047
		C_{T_1}	0.876	0.707	0.052	0.899	0.781	0.05	0.934	0.899	0.051
		C_{T_2}	0.877	0.715	0.054	0.899	0.779	0.049	0.934	0.896	0.049
		C_{T_3}	0.868	0.675	0.049	0.899	0.779	0.05	0.957	0.965	0.047
		C_{T_4}	0.894	0.763	0.049	0.909	0.813	0.05	0.942	0.920	0.049
		C_{T_5}	0.850	0.619	0.051	0.884	0.731	0.05	0.925	0.859	0.047
		C_{T_6}	0.873	0.692	0.049	0.892	0.761	0.052	0.932	0.897	0.053
		C_{T_7}	0.931	**0.886**	0.050	0.936	**0.901**	0.048	0.966	**1.000**	0.048
		C_{T_8}	0.867	0.675	0.051	0.894	0.766	0.051	0.932	0.887	0.049
		C_{T_9}	0.918	0.793	0.029	0.945	0.890	0.031	0.960	**0.970**	0.044
		$C_{T_{10}}$	0.893	0.758	0.050	0.896	0.766	0.049	0.927	0.867	0.047
		$C_{T_{11}}$	0.746	0.275	0.053	0.765	0.338	0.052	0.775	0.364	0.049
		$C_{T_{12}}$	0.862	0.660	0.051	0.902	0.787	0.049	0.924	0.868	0.053

Note. W, L, and S refer to the statistics of Wald test, likelihood ratio test, and score test, respectively. C_{T_1} - $C_{T_{12}}$ refer to the twelve CUSUM-based statistics.

Table 10. Results based on GCM, speededness severity: C8, speededness prevalence = 10%, 30%.

			40 Items			60 Items			80 Items		
Severity	Prevalence	Statistics	CCR	Power	Type-I Error	CCR	Power	Type-I Error	CCR	Power	Type-I Error
C8	10%	W	0.902	0.503	0.053	0.910	0.564	0.052	0.918	0.665	0.054
		L	0.937	**0.817**	0.050	0.941	**0.869**	0.051	0.947	0.926	0.051
		S	0.940	**0.848**	0.050	0.947	**0.913**	0.049	0.955	**0.985**	0.049
		C_{T_1}	0.921	0.687	0.053	0.927	0.752	0.053	0.943	0.888	0.051
		C_{T_2}	0.924	0.694	0.050	0.928	0.752	0.053	0.942	0.877	0.051
		C_{T_3}	0.919	0.656	0.052	0.931	0.753	0.049	0.950	0.930	0.048
		C_{T_4}	0.928	0.735	0.051	0.93	0.786	0.054	0.945	0.903	0.050
		C_{T_5}	0.917	0.613	0.049	0.923	0.704	0.053	0.940	0.847	0.050
		C_{T_6}	0.920	0.680	0.053	0.926	0.746	0.054	0.941	0.882	0.052
		C_{T_7}	0.940	**0.849**	0.050	0.947	**0.877**	0.045	0.958	**0.998**	0.046
		C_{T_8}	0.921	0.664	0.050	0.927	0.735	0.052	0.943	0.868	0.049
		C_{T_9}	0.952	0.768	0.027	0.960	0.869	0.030	0.952	**0.962**	0.049
		$C_{T_{10}}$	0.929	0.729	0.049	0.934	0.753	0.046	0.939	0.836	0.050
		$C_{T_{11}}$	0.879	0.260	0.052	0.885	0.316	0.052	0.891	0.341	0.048
		$C_{T_{12}}$	0.919	0.640	0.050	0.928	0.744	0.051	0.940	0.854	0.050
	30%	W	0.813	0.504	0.055	0.833	0.562	0.051	0.865	0.677	0.054
		L	0.913	**0.826**	0.051	0.928	**0.880**	0.051	0.942	0.926	0.052
		S	0.924	**0.858**	0.047	0.943	**0.921**	0.047	0.962	**0.983**	0.047
		C_{T_1}	0.870	0.687	0.052	0.894	0.767	0.052	0.930	0.881	0.048
		C_{T_2}	0.876	0.702	0.049	0.892	0.761	0.051	0.928	0.879	0.050
		C_{T_3}	0.864	0.661	0.049	0.895	0.76	0.047	0.946	0.934	0.048
		C_{T_4}	0.888	0.744	0.050	0.9	0.792	0.053	0.937	0.906	0.050
		C_{T_5}	0.843	0.604	0.055	0.876	0.708	0.052	0.918	0.844	0.050
		C_{T_6}	0.870	0.685	0.050	0.887	0.741	0.051	0.930	0.883	0.050
		C_{T_7}	0.924	**0.859**	0.047	0.931	**0.88**	0.047	0.966	**0.991**	0.045
		C_{T_8}	0.864	0.664	0.050	0.883	0.738	0.054	0.926	0.868	0.050
		C_{T_9}	0.915	0.778	0.027	0.941	0.874	0.031	0.955	**0.961**	0.047
		$C_{T_{10}}$	0.887	0.735	0.048	0.892	0.752	0.048	0.917	0.840	0.050
		$C_{T_{11}}$	0.736	0.243	0.053	0.754	0.309	0.055	0.761	0.329	0.053
		$C_{T_{12}}$	0.857	0.647	0.052	0.893	0.761	0.050	0.925	0.863	0.049

Note. W, L, and S refer to the statistics of Wald test, likelihood ratio test, and score test, respectively. C_{T_1} - $C_{T_{12}}$ refer to the twelve CUSUM-based statistics.

Table 11. Results based on GCM, speededness severity: C9, speededness prevalence = 10%, 30%.

			40 Items			60 Items			80 Items		
Severity	Prevalence	Statistics	CCR	Power	Type-I Error	CCR	Power	Type-I Error	CCR	Power	Type-I Error
C9	10%	W	0.896	0.433	0.053	0.907	0.535	0.052	0.912	0.595	0.052
		L	0.933	**0.786**	0.050	0.937	**0.819**	0.050	0.945	0.892	0.049
		S	0.936	**0.793**	0.048	0.942	**0.858**	0.049	0.952	**0.948**	0.048
		C_{T_1}	0.914	0.602	0.051	0.922	0.675	0.051	0.936	0.821	0.051
		C_{T_2}	0.915	0.652	0.055	0.919	0.67	0.054	0.935	0.814	0.052
		C_{T_3}	0.910	0.548	0.050	0.919	0.653	0.052	0.936	0.808	0.050
		C_{T_4}	0.920	0.668	0.052	0.926	0.712	0.05	0.939	0.852	0.052
		C_{T_5}	0.909	0.544	0.051	0.917	0.625	0.05	0.931	0.773	0.051
		C_{T_6}	0.915	0.642	0.055	0.919	0.66	0.052	0.937	0.821	0.050
		C_{T_7}	0.932	**0.774**	0.050	0.933	0.782	0.05	0.952	**0.936**	0.046
		C_{T_8}	0.912	0.604	0.054	0.918	0.644	0.052	0.936	0.822	0.052
		C_{T_9}	0.946	0.705	0.028	0.950	**0.787**	0.032	0.945	**0.899**	0.050
		$C_{T_{10}}$	0.915	0.619	0.052	0.921	0.664	0.050	0.925	0.715	0.052
		$C_{T_{11}}$	0.872	0.199	0.054	0.883	0.272	0.050	0.885	0.290	0.049
		$C_{T_{12}}$	0.912	0.579	0.051	0.924	0.677	0.049	0.932	0.793	0.052
	30%	W	0.782	0.404	0.056	0.817	0.513	0.053	0.842	0.603	0.055
		L	0.905	**0.801**	0.051	0.912	**0.827**	0.052	0.933	0.893	0.051
		S	0.909	**0.806**	0.047	0.927	**0.867**	0.047	0.952	**0.952**	0.047
		C_{T_1}	0.846	0.608	0.052	0.87	0.687	0.052	0.909	0.820	0.053
		C_{T_2}	0.863	0.660	0.050	0.872	0.688	0.05	0.909	0.819	0.053
		C_{T_3}	0.838	0.581	0.052	0.87	0.676	0.047	0.915	0.831	0.049
		C_{T_4}	0.865	0.676	0.054	0.883	0.727	0.051	0.918	0.848	0.052
		C_{T_5}	0.826	0.537	0.050	0.853	0.628	0.051	0.895	0.780	0.056
		C_{T_6}	0.859	0.648	0.051	0.865	0.673	0.052	0.912	0.831	0.054
		C_{T_7}	0.905	**0.787**	0.045	0.905	**0.798**	0.049	0.947	**0.928**	0.045
		C_{T_8}	0.845	0.605	0.052	0.865	0.667	0.051	0.908	0.813	0.051
		C_{T_9}	0.897	0.721	0.028	0.917	0.793	0.029	0.936	**0.907**	0.051
		$C_{T_{10}}$	0.862	0.651	0.048	0.869	0.679	0.049	0.888	0.739	0.049
		$C_{T_{11}}$	0.718	0.186	0.054	0.736	0.238	0.051	0.740	0.251	0.051
		$C_{T_{12}}$	0.842	0.591	0.051	0.871	0.684	0.049	0.904	0.798	0.051

Note. W, L, and S refer to the statistics of Wald test, likelihood ratio test, and score test, respectively. C_{T_1} – $C_{T_{12}}$ refer to the twelve CUSUM-based statistics.

Table 12. Relative frequency of each statistic ending up in the top three in terms of power across conditions under the HM.

	40 Items		60 Items		80 Items	
Statistics	Power	Type I Error	Power	Type I Error	Power	Type I Error
W	0	0	0	0	0	0
L	0	0	0	0	0	0
S	0	**0.278**	0	**0.222**	0	**0.167**
C_{T_1}	0	0.056	0	0.056	0	0
$\boldsymbol{C_{T_2}}$	**0.333**	0.056	**0.333**	0	0	0
C_{T_3}	0	0	0	0	0	0.111
C_{T_4}	0	0	0	0.056	0	0
C_{T_5}	0	0.056	0	**0.167**	0	**0.278**
$\boldsymbol{C_{T_6}}$	**0.333**	0.056	**0.333**	0.056	**0.333**	0
$\boldsymbol{C_{T_7}}$	0	**0.222**	0	**0.278**	**0.333**	**0.333**
C_{T_8}	0	0	0	0.111	0	0
$\boldsymbol{C_{T_9}}$	**0.333**	**0.167**	**0.333**	0	**0.333**	0
$C_{T_{10}}$	0	0.111	0	0.056	0	0
$C_{T_{11}}$	0	0	0	0	0	0
$C_{T_{12}}$	0	0	0	0	0	0.111

Note: (1) all the percentages were calculated based on Study 1. For the top three statistics in power for each test length, the cells are shaded with solid gray color. For the top three statistics in type I error rates for each test length, the cells are shaded with gray lines. Shaded entries mean empirical type-I error closest to 5%.
(2) W, L, and S refer to the statistics of Wald test, likelihood ratio test, and score test, respectively. C_{T_1} – $C_{T_{12}}$ refer to the twelve CUSUM-based statistics.

To obtain a general overview of the performances for these 15 statistics, a comprehensive comparison for CUSUM and CPA based on HM and GCM were conducted, respectively. Overall, both CUSUM-based and CPA-based statistics can get well-controlled type I error rates in detecting speededness. Therefore, the following comparison will primarily consider power. The number in each cell of Tables 12 and 13 means the percentage of times for the statistic to rank in the top three in terms of power among the 15 statistics in all the simulation settings under each test length. For example, the number 0.333 (frequency of being in the top three for the power of L_{max} statistic under GCM in the 40-item test) means L_{max} is in the top three highest power indices about one third of the times among all conditions under the GCM. Under the HM, C_{T_2}, C_{T_6}, and C_{T_9} are the most frequent winners for the 40- and 60-item test, while C_{T_6}, C_{T_7}, and C_{T_9} are for the 80-item tests. In most cases, C_{T_9} has the highest detection rate. Under the GCM, L_{max}, C_{T_7} and S_{max} are the most frequent winners for the 40- and 60-item test, while C_{T_7}, C_{T_9} and S_{max} are for the 80-item test. Some statistics with three zeros mean these statistics never end up in the top three in terms of power.

Figure 2 compares the top three statistics in power for each test length under HM. The figures show that C_{T_9} always outperforms the other statistics. In the 40- and 60-item tests, C_{T_6} has relatively higher power than C_{T_2}, while their powers become pretty close in

Table 13. Relative frequency of each statistic ending up in the top three in terms of power across conditions under the GCM.

	40 Items		60 Items		80 Items	
Statistics	Power	Type I Error	Power	Type I Error	Power	Type I Error
W	0	0	0	0	0	0
L	**0.333**	0.037	**0.315**	0	0.111	0.019
S	**0.333**	**0.241**	**0.333**	**0.241**	**0.333**	**0.296**
T1	0	0	0	0.019	0	0.019
T2	0	0.074	0	0	0	0
T3	0	**0.111**	0	**0.111**	0	0.093
T4	0	0	0	0.019	0	0.037
T5	0	0.019	0	0.019	0	0.019
T6	0	0.037	0.019	0	0	0
T7	**0.333**	0.074	**0.241**	0.074	**0.315**	**0.241**
T8	0	0.019	0	0.037	0	0.019
T9	0	**0.333**	0.093	**0.333**	**0.241**	**0.185**
T10	0	0.056	0	0.111	0	0.037
T11	0	0	0	0	0	0.019
T12	0	0	0	0.037	0	0.019

Note: (1) all the percentages were calculated based on Study 2. For the top three statistics in power for each test length, the cells are shaded with solid gray color. For the top three statistics in type I error rates for each test length, the cells are shaded with gray lines. Shaded entries mean empirical type-I error closest to 5%.

(2) Note. W, L, and S refer to the statistics of Wald test, likelihood ratio test, and score test, respectively. T1 - T12 refer to the twelve CUSUM-based statistics.

the 80-item test, suggested by the almost overlapping curves in Figure 2.

Figures 3 illustrates the power for the top three statistics under GCM. It can be seen that the three statistics corresponding to each test length show different patterns. Take the 40-item test as an example: the top three statistics, C_{T_7}, L_{max}, and S_{max} have relatively close power in the median- and high-severity conditions (median severity: from C4 to C6, high-severity: from C7 to C9), while L_{max} outperforms C_{T_7} and S_{max} in the low-severity conditions (low severity: from C1 to C3). However, S_{max} seems to yield relatively high detection rates in the median- and high-severity conditions in 40- and 60-item tests. When the test length goes up to 80, C_{T_7}, C_{T_9}, and S_{max} have close detection rates in the median- and high-severity conditions, while S_{max} shows an advantage in the low-severity conditions. Overall, S_{max} yields a power of .7 or above in all conditions. Taking everything together, it seems that C_{T_9} and S_{max} are the best statistic under HM and GCM, respectively.

From Figures 2 and 3, some interesting patters of power in relation to other manipulated factors can be observed. The power increases with the extent of the speededness severity (small median suggesting more responses being affected by speededness in a response sequence). Power also increases with the increase of test length (with fixed speededness severity). With known item parameters, speededness prevalence shows little effect on power as speededness detection is performed on each response pattern separately, independent of other response patterns.

Real data example

To further evaluate the performance of the speededness detection statistics, we applied four statistics that outperform others in the simulation studies (C_{T_7}, C_{T_9}, L_{max}, S_{max}) to an empirical dataset of a large-scale standardized state assessment. In this study, the data was collected from 2012–2013 in Indiana, and a total of 32 items were administered online. Responses of 49,767 respondents to thirty multiple-choice items were used in the analysis after removing the two non-multiple-choice items. The dataset was also used in Shao et al. (2016), in which they illustrated the use of the likelihood ratio test. The entire dataset was fitted based on the 2PL model, and the means of the difficulty parameter and the discrimination parameter are −.0575, 1.179, respectively. The variances of the difficulty parameter and the discrimination parameter are 0.739, 0.231, respectively.

Three thousand respondents were randomly sampled from the dataset. We then applied the four chosen statistics (L_{max}, S_{max}, C_{T_7}, C_{T_9}) to identify the respondents with speededness behavior. Their critical or bound values were obtained based on simulation. As summarized in Table 14, the critical values of L_{max} and S_{max} are 8.134 and 7.785, respectively, and the values of the bound for C_{T_7} are [-0.034, 0.035], and 2.713 for C_{T_9}.

After obtaining the associated critical values and the values of the bound, we calculated the value of each statistic for each possible position for each examinee, with the ability parameter updated by MLE. Then if the value surpasses the critical value or the bound, this means the examinee may have exhibited certain aberrant behavior beyond this item position, such as speededness, preknowledge, etc. Since speededness is generally considered detrimental to response accuracy, we compared the ability estimate to confirm if the aberrant behavior could be speededess or not. If the ability estimate drops after the change point, the examinee would be flagged for possible speededness.

Table 14 summarizes the number of flagged respondents for possible speededness by the four top statistics. The numbers vary from 111 to 159, with the CPA statistics L_{max} and S_{max} flagging more than the CUSUM statistics. Regardless of the statistic used, only between 110 to 160 respondents were flagged for possible speededness out of 3000 participants, suggesting that speededness is not a prevalent issue for this

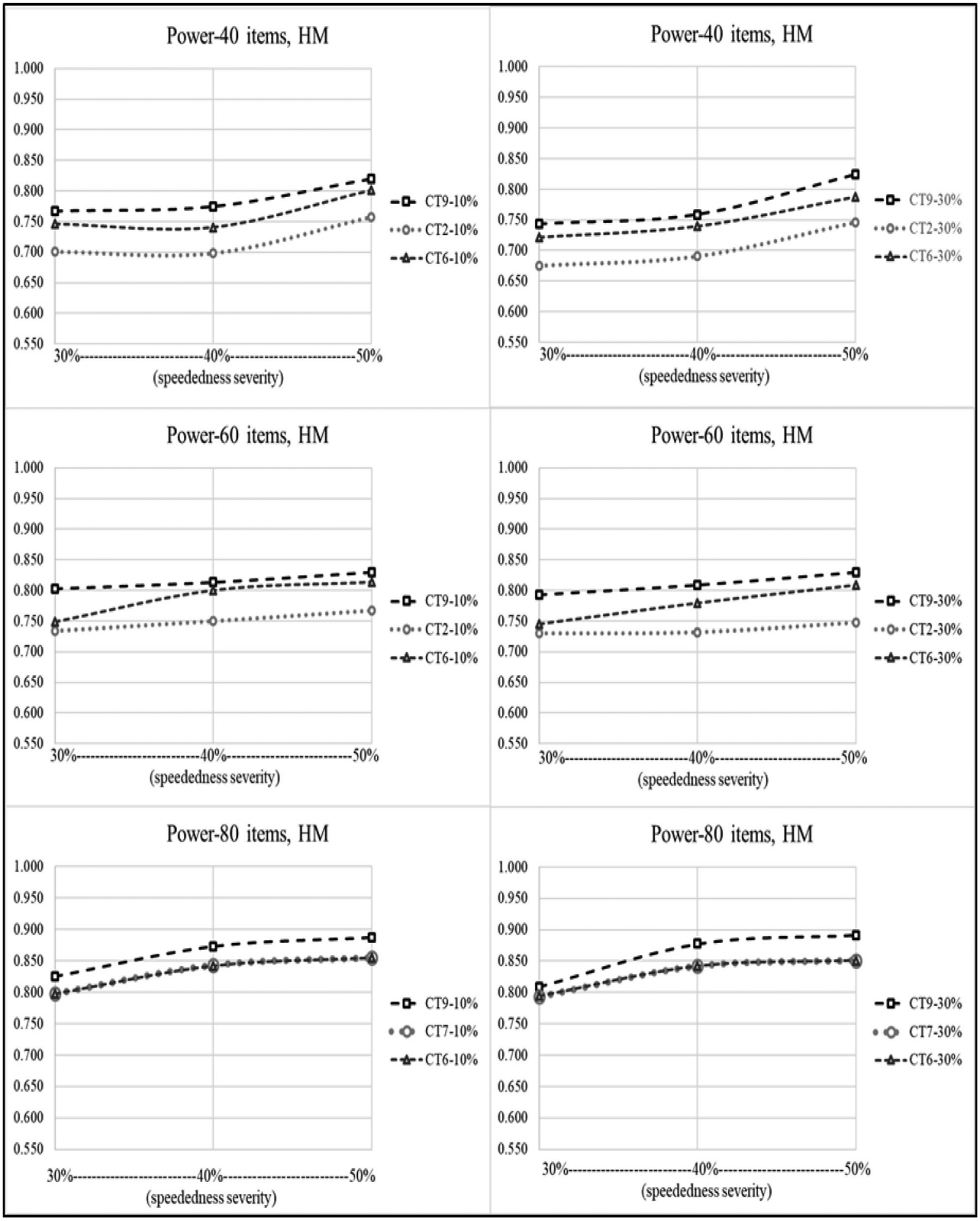

Figure 2. Comparison of power for top three indices under HM, speededness prevalence: 10% (left) and 30% (right). Speededness severity: low (30%), medium (40%), high (50%).

testing program. This is unsurprising because a standardized statewide assessment has undergone rigorous quality control process. We also looked into the agreement of the respondents flagged by the four statistics of interest. Results show that 61 respondents were flagged by both of the two CPA-based statistics

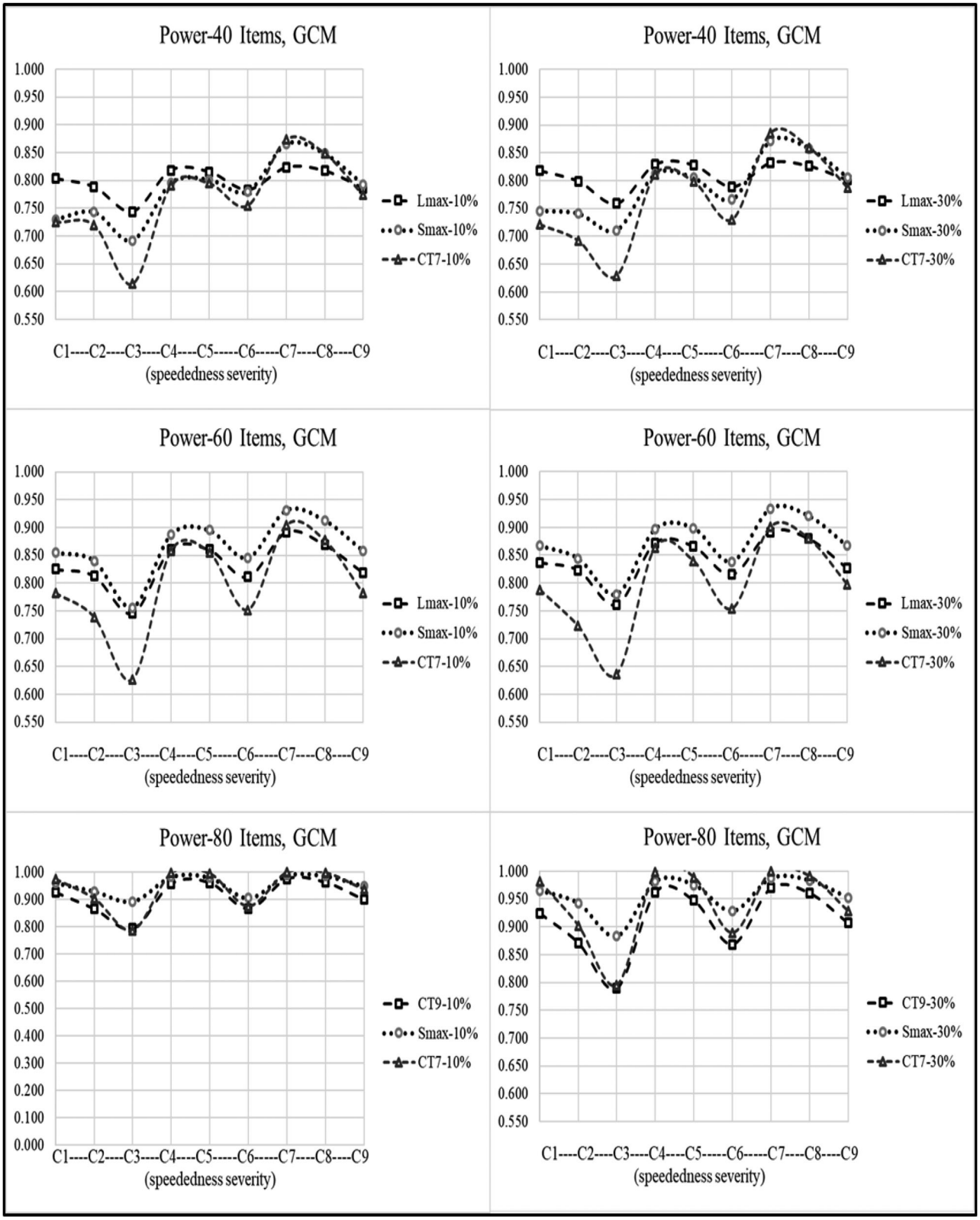

Figure 3. Comparison of power for top three indices under GCM, speededness prevalence: 10% (left) and 30% (right). Speededness severity: low (C1, C2, C3), medium (C4, C5, C6), high (C7, C8, C9).

and 48 for the two CUSUM-based statistics. 33 respondents were flagged by all the four statistics. Different sets of respondents flagged by these statistics suggest that: a) Different mechanisms of speededness might exist in the data; and b) We need to exercise caution before flagging a respondent as aberrant in

practice. Flagged response patterns require careful human review, ideally along with auxiliary information such as response time and log data, if available.

Figure 4 shows the response sequence and ability estimates of a respondent flagged by all four statistics. Its left and right panel shows how his or her responses and provisional ability estimates change over the item sequence, respectively. As we can see, there is a clear drop in ability estimates, and a consistent drop in proportion of correct responses at the end of the test. This indicates that the flagged respondents may have suffered from the time pressure toward the end of the test and their performance was negatively impacted.

Discussion

Results from our simulation studies show the performances of CUSUM procedures and CPA methods in detecting speededness depend on the underlying mechanism of speededness and the severity of the speeded behaviors. When the test is relatively short (e.g., 40 items) or of medium-length (e.g., 60 items), among all statistics, L_{max}, S_{max} and C_{T_7} are more powerful than the others under the GCM, while C_{T_2}, C_{T_7} and C_{T_9} perform the best under the HM. When the test is long (80), S_{max}, C_{T_7} and C_{T_9} outperform the others under the GCM, in contrast to C_{T_6}, C_{T_7} and C_{T_9} under HM. When comparing the two types of methods against each other, the advantage of CUSUM statistics are pronounced under the HM, while the performance of CPA statistics improve substantially under the GCM. In reality, due to the unknown mechanism of speededness, we recommend the use of C_{T_7} and C_{T_9} when the test length is 80, regardless of HM and GCM. They can reach respectable power (.75 or above) across a wide range of conditions under either model. In a short or medium-length test, no statistic always ends up in the top three under both HM and GCM. In this case, C_{T_9} is a reasonable choice because although it doesn't end up in the top three list under the GCM, it still has a relatively high power and comes in as the 4th.

In addition to the underlying mechanism of speededness, results also reveal that the test length needs to be considered as a factor for choosing the statistic to detect speededness in real applications. In general, the power increases as the test gets longer. If the test length reaches 80, some CUSUM-based statistics may be more effective. Meanwhile, L_{max} and S_{max} can always be used as options. The relative performance may change when the test length differs, and this result is consistent with what is found in Sinharay (2016). Meanwhile, some statistics show better performances in the median- and high-severity conditions, and some are favored in the low-severity condition. One thing we need to be aware of is that examinees with low abilities (such as θ less than or close to -2) pose challenges to the detection of speededness even when they are affected by speededness.

Table 14. Information for the flagged examinees with speededness.

Statistics	Critical Values or Values of the Bound	Numbers of Flagged Respondents
L	8.134	159
S	7.785	122
C_{T_7}	[-0.034, 0.035]	111
C_{T_9}	2.713	112

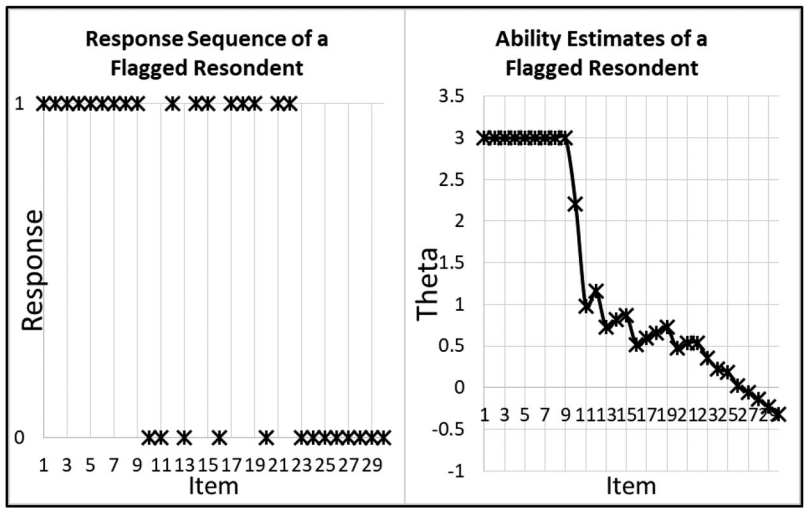

Figure 4. Responses and ability estiamtes of a flagged respondent.

This is because there may be only a very small difference between their normal behavior and speeded behavior in terms of probability of getting a correct response. In this case, response time information may be particularly helpful for detecting speededness (Shao & Cheng, 2017).

In spite of the very informative findings, there are some limitations of the current study. First, we only considered two mechanisms of test speededness represented by the HM and GCM, respectively, while aberrant response behavior is so complicated that it can never be fully captured by any single model. Second, we examine the situation when all speeded examinees only follow one underlying speededness mechanism or has only one change point. In reality, examinees may exhibit more complex patterns of aberrant response behavior. Further, in this study only P&P test is considered. Other important testing modes, for example, typical computerized adaptive test (CAT; van der Linden & Glas, 2000) and computerized multistage test (MST; Yan et al., 2014) should be considered in future studies, because aberrant response behavior could have a larger influence on CAT than on a P&P test. Moreover, the item parameters are assumed known in the current study. This is not an unreasonable assumption, as many testing programs calibrate their items in pretesting and treat the item parameters estimates as known in operations anyways. For example, many applications of IRT models rely on estimated item parameters such as scoring of test takers (Cheng & Yuan, 2010; Zwick et al., 1995), fixed- and variable-length computerized adaptive testing (Cheng et al., 2015; Meijer & Nering, 1999; Patton et al., 2019), estimation of classification accuracy and consistency (Lathrop & Cheng, 2013), equating (Kolen & Brennan, 2014; Skaggs & Lissitz, 1986), and so on so forth. However, given that the true item parameters are actually never known, the advantage of CUSUM statistics may diminish if estimation error is taken into account. Further research is warranted on this issue.

Finally, it is important to note that flagging an aberrant response pattern is always a controversial practice, especially in those high-stakes setting. For example, removing the flagged patterns may affect parameter estimation (Michaelides, 2010), and test information, reliability, validity (Hong et al., 2020), and representation of the overall test construct (Ozturk & Karabatsos, 2017). Therefore, in real applications, it is prudent to use caution before removing any responses.

Open practices statement

We have provided sample matlab codes for researchers to replicate real data analysis with two CPA statistics and two CUSUM statistics. Please see the supplemental file: SpeedednessDetectionDemo.m.

Article Information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

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

Funding: This work is partially supported by NSF grant SES-1853166 awarded to the corresponding author.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like to thank the reviewers and AE for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions is not intended and should not be inferred.

References

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

Armstrong, R. D., & Kung, M. T. (2011). CUSUM hypothesis tests and alternative response probabilities for finite poisson trials. *Communications in Statistics - Simulation and Computation*, *40*(7), 1057–1073. https://doi.org/10.1080/03610918.2011.563003

Armstrong, R. D., & Shi, M. (2009a). Model-free CUSUM methods for person fit. *Journal of Educational Measurement*, *46*(4), 408–428. https://doi.org/10.1111/j.1745-3984.2009.00090.x

Armstrong, R. D., & Shi, M. (2009b). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement*, *33*(5), 391–410. https://doi.org/10.1177/0146621609331961

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Marcel Dekker. https://doi.org/10.1201/9781482276725

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

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under Conditions of test speededness: Application of a Mixture rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331–348. 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/10.1080/01621459.1998.10473747

Cheng, Y., & Yuan, K. H. (2010). The impact of fallible item parameter estimates on latent trait recovery. *Psychometrika*, *75*(2), 280–291. https://doi.org/10.1007/s11336-009-9144-x

Cheng, Y., Patton, J. M., & Shao, C. (2015). a-Stratified computerized adaptive testing in the presence of calibration error. *Educational and Psychological Measurement*, *75*(2), 260–283. https://doi.org/10.1177/0013164414530719

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Egberink, I. J. L., Meijer, R. R., Veldkamp, B. P., Schakel, L., & Smid, N. G. (2010). Detection of aberrant item score patterns in computerized adaptive testing: A empirical example using the CUSUM. *Personality and Individual Differences*, *48*(8), 921–925. https://doi.org/10.1016/j.paid.2010.02.023

Glliksen, H. (1950). *Theory of mental tests*. Wiley. https://doi.org/10.1037/13240-000

Goegebeur, Y., De Boeck, P., & Molenberghs, G. (2010). Person fit for test speededness: Normal curvatures, likelihood ratio tests and empirical Bayes estimates. *Methodology*, *6*(1), 3–16. https://doi.org/10.1027/1614-2241/a000002

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory. Principles and applications* (pp. 15–31). Norwell, MA: Kluwer Academic Publishers.

Han, K. T. (2013). Item pocket method to allow response review and change in computerized adaptive testing. *Applied Psychological Measurement*, *37*(4), 259–275. https://doi.org/10.1177/0146621612473638

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

Hong, M., Steedle, J. T., & Cheng, Y. (2020). Methods of detecting insufficient effort responding comparisons and practical recommendations. *Educational and Psychological Measurement*, *80*(2), 312–345. https://doi.org/10.1177/0013164419865316

Kolen, M. J., & Brennan, R. L. (2014). Item response theory methods. In *Test equating: Methods and practices* (pp. 171–245). Springer Science & Business Media. https://doi.org/10.1007/978-1-4939-0317-7 https://doi.org/10.1007/978-1-4757-4310-4

Lathrop, Q. N., & Cheng, Y. (2013). Two approaches to estimation of classification accuracy rate under item response theory. *Applied Psychological Measurement*, *37*(3), 226–241. https://doi.org/10.1177/0146621612471888

Lee, Y. -H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika*, *78*(3), 557–575. https://doi.org/10.1007/s11336-013-9317-5L

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

Meijer, R. R. (2002). Outlier detection in high-stakes certification testing. *Journal of Educational Measurement*, *39*(3), 219–233. https://doi.org/10.1111/j.1745-3984.2002.tb01175.x

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, *23*(3), 187–194. https://doi.org/10.1177/01466219922031310

Michaelides, M. P. (2010). A review of the effects on IRT item parameter estimates with a focus on misbehaving common items in test equating. *Frontiers in Psychology*, *1*, 167–167. https://doi.org/10.3389/fpsyg.2010.00167

Montgomery, D. C. (2013). *Introduction to statistical quality control* (7th ed.). Wiley.

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

Ozturk, N. K., & Karabatsos, G. (2017). A bayesian robust IRT outlier-detection model. *Applied Psychological Measurement*, *41*(3), 195–208. https://doi.org/10.1177/0146621616679394

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *41*(1–2), 100–115. https://doi.org/10.1093/biomet/41.1-2.100

Patton, J. M., 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

Rost, J. (1990). Rasch models in latent classes: Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282. https://doi.org/10.1177/014662169001400305

Schlemer, L. T. (2007). *Test speededness and cognitive styles: A study using a CAT version of the SAT* [Doctoral dissertation]. University of California.

Schnipke, D. L. (1996). *How contaminated by guessing are item-parameter estimates and what can be done about it? Paper presented at the annual meeting of the National Council on Measurement in Education*, New York, NY.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling *item* response times with a two-state mixture model: A new

method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

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

Shao, C., & Cheng, Y. (2017). *Detection of test speededness using change-point analysis with response time data* [Paper presentation]. Annual Meeting of National Council for Measurement in Education, April), San Antonio, TX.

Shao, C., Li, J., & Cheng, Y. (2014). *Test speededness detection based on the detection of change point* [Paper presentation]. Paper Presented at the 79th Annual Meeting of the Psychometric Society, Madison, WI.

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

Shi, M. (2007). *Detection of aberrant response patterns in testing using cumulative sum control schemes* [Doctoral dissertation]. Rutgers University.

Siegmund, D. (1985). *Sequential analysis: Tests and confidence intervals*. Springer.

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

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research*, *56*(4), 495–529. https://doi.org/10.3102/00346543056004495

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*(3), 331–342. https://doi.org/10.1007/bf02294437

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

Tendeiro, J. N., Meijer, R. R., Schakel, L.,& Maij-de Meij, A. M. (2013). Using cumulative sum statistics to detect inconsistencies in unproctored internet testing. *Educational and Psychological Measurement*, *73*(1), 143–161. https://doi.org/10.1177/0013164412444787

Tendeiro, J. N., & Meijer, R. R. (2012). A CUSUM to detect person misfit: A discussion and some alternatives for existing procedures. *Applied Psychological Measurement*, *36*(5), 420–442. https://doi.org/10.1177/0146621612446305

van der Linden, W. J. (2009). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement*, *33*(1), 25–42. https://doi.org/10.1177/0146621607314042

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., & Glas, G. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Springer. https://doi.org/10.1007/0-306-47531-6

van der Linden, W. J., & Xiong, X. H. (2013). Speededness and adaptive testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 418–438. https://doi.org/10.1177/0146621607314042

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 195–210. https://doi.org/10.1177/01466219922031329

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detecting person-misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 210–219). Kluwer. https://doi.org/10.1007/0-306-47531-6_11

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, *26*(2), 199–217. https://doi.org/10.3102/10769986026002199

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement*, *26*(2), 164–180. https://doi.org/10.1177/01421602026002004

Wang, C., & Xu, G. J. (2015). A mixture hierarchical model for response times and response accuracy. *The British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. https://doi.org/10.1111/bmsp.12054

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

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement*, *40*(4), 307–330. https://doi.org/10.1111/j.1745-3984.2003.tb01149.x

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

Yamamoto, K. (1989). *HYBRID model of IRT and latent class models (RR-89-41)*. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1982.tb01326.x

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Application of latent trait and latent class models in the social sciences* (pp. 89–98). Waxmann. https://doi.org/10.1002/j.2333-8504.1995.tb01637.x

Yan, D. L., von Davier, A. A., & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*. CRC Press. https://doi.org/10.1201/b16858

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of*

Educational Measurement, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Yu, X., & 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

Zhang, J. M. (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

Zwick, R., Thayer, D. T., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement*, *32*(4), 341–363. https://doi.org/10.1111/j.1745-3984.1995.tb00471.x

Appendix

Study 3: Performance of under HM and GCM based on asymptotic critical values

In order to further evaluate the performance of the three CPA statistics in speededness detection based on asymptotic critical values (see Table 2 of Sinharay, 2016) under HM and GCM, a simulation study was conducted. Everything else was kept the same as the previous studies except the critical values. It should be noted that there is some difference in calculating every statistic for each examinee in this Study, as these asymptotic critical values are only applicable when the speededness point is within the middle 70% of the test (Sinharay, 2016). For example, we only survey from the 6th to the 34th item for a 40-item test. Tables A1 and A2 show the results of W_{max}, L_{max}, and S_{max} based on the asymptotic critical values. Only the results for the 80-item test were reported due to the space limit. Compared to the results in the first two studies, this study shows a somewhat different trend because W_{max} has a comparable detection rate with L_{max} and S_{max} based on the asymptotic critical values. Regardless of HM or GCM, the asymptotic critical value seems too liberal for W_{max}, which leads the detection rate and the type I Error rate to increase considerably and simultaneously. Under the HM, S_{max} always has a higher detection rate than L_{max}. A similar conclusion can be drawn under the GCM. The inflated type I Error rate of S_{max} indicates that its asymptotic critical value is also too liberal for the 80-item test. The same as Study 1 and Study 2, the three CPA-based statistics performed better under the GCM than under HM as expected. L_{max} showed particularly strong performance with a high classification accuracy rate and a relatively small type I Error rate.

What leads to the increased power and in some cases, inflated Type-I error? The reason could lie in the critical values and the search range. The critical values for W_{max} obtained by Monte Carlo simulations are much larger than the asymptotic critical values, because we search for the change point from the beginning to the end of the test in Study 1 and Study 2, but only for the middle range of the test in Study 3. This may lead to large values for W_{max} in some cases in the simulations, and in turn a larger critical value, especially under the MLE ability estimator in which the boundary values (set to −3 or 3 in the study) of the ability estimate were assigned to those test takers with all wrong or all correct answers. Therefore, the simulated critical value are more conservative than the asymptotic critical value for W_{max}.

Table A1. Results based on HM, J = 80, speededness prevalence: 10% & 30%, speededness severity: 30%, asymptotic critical value.

		10%			30%		
100*n1/n	Statistics	CCR	Power	Type I Error	CCR	Power	Type I Error
5	W	0.781	0.765	0.218	0.774	0.779	0.228
	L	0.944	0.712	0.03	0.894	0.721	0.032
	S	0.9	**0.792**	0.088	0.879	**0.806**	0.09
10	W	0.851	0.772	0.14	0.834	0.775	0.141
	L	0.943	0.726	0.033	0.898	0.739	0.034
	S	0.909	**0.811**	0.08	0.886	**0.815**	0.083
15	W	0.88	0.779	0.109	0.855	0.778	0.111
	L	0.941	0.736	0.037	0.897	0.744	0.037
	S	0.922	**0.807**	0.065	0.894	**0.805**	0.069
20	W	0.894	0.759	0.091	0.864	0.769	0.095
	L	0.939	0.727	0.037	0.894	0.736	0.038
	S	0.927	**0.794**	0.058	0.902	**0.808**	0.058

Note. W, L, and S refer to the statistics of Wald test, likelihood ratio test, and score test, respectively.

Table A2. Results based on GCM, J = 80, C7, speededness prevalence: 10% & 30%, asymptotic critical value.

		10%			30%		
100*n1/n	Statistics	CCR	Power	Type I Error	CCR	Power	Type I Error
5	W	0.671	0.887	0.353	0.719	0.891	0.355
	L	0.965	0.916	0.029	0.956	0.925	0.03
	S	0.916	**0.996**	0.092	0.936	**0.995**	0.089
10	W	0.816	0.891	0.193	0.836	0.896	0.190
	L	0.962	0.923	0.033	0.954	0.927	0.035
	S	0.930	**0.997**	0.078	0.942	**0.996**	0.081
15	W	0.892	0.901	0.109	0.894	0.897	0.108
	L	0.96	0.932	0.037	0.954	0.931	0.036
	S	0.937	**0.997**	0.069	0.951	**0.997**	0.069
20	W	0.908	0.901	0.091	0.907	0.899	0.09
	L	0.961	0.929	0.036	0.953	0.932	0.038
	S	0.946	**0.999**	0.06	0.959	**0.999**	0.058

Note. W, L, and S refer to the statistics of Wald test, likelihood ratio test, and score test, respectively.