Article

Application of Change Point Analysis of Response Time Data to Detect Test Speededness

Educational and Psychological Measurement
2022, Vol. 82(5) 1031–1062

DOI: 10.1177/00131644211046392
journals.sagepub.com/home/epm
SAGE

Ying Cheng[1] and Can Shao[2]

Abstract

Computer-based and web-based testing have become increasingly popular in recent years. Their popularity has dramatically expanded the availability of response time data. Compared to the conventional item response data that are often dichotomous or polytomous, response time has the advantage of being continuous and can be collected in an unobstrusive manner. It therefore has great potential to improve many measurement activities. In this paper, we propose a change point analysis (CPA) procedure to detect test speededness using response time data. Specifically, two test statistics based on CPA, the likelihood ratio test and Wald test, are proposed to detect test speededness. A simulation study has been conducted to evaluate the performance of the proposed CPA procedure, as well as the use of asymptotic and empirical critical values. Results indicate that the proposed procedure leads to high power in detecting test speededness, while keeping the false positive rate under control, even when simplistic and liberal critical values are used. Accuracy of the estimation of the actual change point, however, is highly dependent on the true change point. A real data example is also provided to illustrate the utility of the proposed procedure and its contrast to the response-only procedure. Implications of the findings are discussed at the end.

Keywords

change point analysis, intra-individual change, response time, test speededness, log-normal model

[1]University of Notre Dame, Notre Dame, IN, USA
[2]Applied Materials Inc, Santa Clara, CA, USA
Ying Cheng and Can Shao made equal contributions and are listed alphabetically.

Corresponding Author:
Ying Cheng, Department of Psychology, University of Notre Dame, 390 Corbett Hall, Notre Dame, IN 46556, USA.
Email: ycheng4@nd.edu

Introduction

In recent years, response time has received rapidly growing amount of attention in psychometric research (Lee & Chen, 2011), likely due to the increasing availability of (item-level) response time data through computer-based testing and online survey data collection. The collection of response time data is unobtrusive as well, meaning that test takers are typically unaware that response time data are being collected. As such, response time on test items has the potential to improve many measurement activities, for example, item parameter estimation of item response theory (IRT) models (Bhola, 1994; Schnipke, 1995, 1996, 1999; van der Linden et al., 2010), test assembly (van der Linden, 2011; van der Linden & Xiong, 2013), and item selection in computerized adaptive testing (CAT; Cheng et al., 2017; Fan et al., 2012; van der Linden, 2008).

One notable application of response time has been detection of aberrant response behavior (van der Linden & van Krimpen-Stoop, 2003). Many researchers have explored the use of response time to detect various types of aberrant response behavior, such as speededness (Schnipke & Scrams, 1997; van der Linden et al., 1999), low motivation or lack of effort (Wise & Kong, 2005), or item pre-knowledge (van der Linden & Guo, 2008). In this paper we focus on test speededness. The test speededness occurs when time limit affects examinees' performance on a test while speed is not part of the construct(s) of the test purports to measure (Evans & Reilly, 1972; Shao et al., 2016). Test speededness may cause lower response accuracy toward the end of a test or missing responses (Goegebeur et al., 2010). Test speededness could also bias the estimates of item and ability estimates (Douglas et al., 1998; Oshima, 1994) and undermine the validity of the test scores, and may distort or weaken the relationship between test scores and other variables. It is therefore critical to detect response patterns that are affected by speededness.

Traditionally, unreached items toward the end of a test are considered an indication of test speededness (Schnipke & Scrams, 1997), in which case response time on these items is 0. A more elusive and more prevalent form of test speededness is ''rapid guessing behavior'' (Schnipke & Scrams, 1997; Wise & Kong, 2005) on multiple choice items, meaning when time is running out, examinees may rapidly respond to remaining items before time expires, and it is expected that ''the correctness of these answers will be at or near chance levels'' (Wise & Kong, 2005, p. 167). Many rapid guessers may appear to finish the test in time, that is, there is no unreached items on their test. Therefore more sophisticated approaches than looking for unreached items are needed to detect speededness.

To date, there have been at least three approaches to detect test speededness using response time data. The first approach is to model the distribution of response time that are unaffected by that speededness and detecting speededness by model-based inferences. Various measurement models have been proposed for response time in psychological and educational testing. These include parametric models such as log-normal model (Thissen, 1983; van der Linden, 2006), Weibull model (Rouder et al., 2003; Wang, 2006), and gamma model (Maris, 1993). In addition to these parametric

models, a number of ''flexible parametric'' and semiparametric models have been proposed, which make weaker distributional assumptions and can subsume existing parametric models (Ranger & Kuhn, 2012). These models include the Box-Cox normal model (Entink et al., 2009), the proportional hazards model (Ranger & Ortner, 2012; Wang, Fan et al., 2013), and the linear transformation model (Ranger & Kuhn, 2013; Wang, Chang et al., 2013). To detect speededness, a chosen model is fit to the item response time data. Someone whose response time pattern does not fit the model (i.e., with large residuals) can be flagged (van der Linden & van Krimpen-Stoop, 2003; van der Linden et al., 2007). This approach is similar to detecting person misfit under the IRT framework. A chosen IRT model, that is, the null model, is fit to the data, and response patterns with large residuals (e.g., large l_z statistics) get flagged.

The second approach is to directly model response time data that are affected by test speededness. One common approach is to use mixture modeling, with one latent class corresponding to one type of test-taking behavior (Meyer, 2010; Molenaar et al., 2016), for instance solution behavior and rapid guessing behavior (Wang & Xu, 2015), respectively. One can fit a mixture model to the response time data and identify respondents with rapid guessing behavior using estimated class membership. Another approach is to model the trajectory of the change within one individual during the test taking process using latent growth model (Fox & Marianti, 2016). By fitting the latent growth response time model, a person going through the test with non-invariant speed, for example, linearly increasing or quadratically increasing speed, can be flagged.

The third approach does not explicitly utilize any measurement model of response time. Instead, one can use visual inspection or arbitrarily chosen thresholds to flag respondents who may have been affected by speededness. For example, Schnipke (1995) used response time (in conjunction with the accuracy rates) to assess item-level speededness in computer-based testing. In that study, she first examined the computer-based Graduate Record Examinations (GRE) in terms of the percentage of examinees who reached 75%, 80% of the test, or the last item. The response time and response accuracy (i.e., the proportion of test takers getting the item right) were then used to identify examinees who had rapid guessing behavior through visual inspection and a series of analyses of variance. Guo et al. (2016) compared five methods to detect rapid guessing behavior. These all methods involve identifying a threshold that indicates rapid guessing for each item. Then test takers whose response time is less than the threshold will be flagged. For each test taker, the proportion of items on the test for which his or her response time exceeded the established threshold values (Swerdzewski et al., 2011; Wise & Kong, 2005) is computed. This proportion is referred to as the overall response time effort (RTE) index. Test takers with low RTE are considered showing rapid guessing behavior. These methods are item-based in the sense that thresholds need to be first established for each item. For example, a commonly used threshold is the 3 second threshold for all items. Demars (2007) discussed visually inspecting the distribution of response time of each item and identify a gap. The gap, if exists, can be a natural threshold for the item. Wise and Ma (2012) proposed to identify rapid guessing by using the normative time threshold, which is

defined as a certain percentage (e.g., 10%) of the mean time the entire sample of test takers spent on the item. Anyone spending less time than the normative time threshold would be considered as a rapid guesser on the item. Such normative thresholds were also adopted in Lee and Jia (2014) in flagging non-effortful test takers in a The National Assessment of Educational Progress (NAEP) study. The establishment of appropriate thresholds is critical to the performance of these methods (Kong et al., 2007).

In this paper we propose a procedure based on change point analysis (CPA) to detect speededness using item response time data. Similar to the first approach described above, the CPA approach assumes a null distribution of response time, that is, distribution of response time when it is not affected by speededness. However, it does not examine how an individual response time pattern deviates from the group behavior (in the sense of Mahalanobis distance) or model-implied behavior (in the sense of residuals such as person-fit statistics). Instead, it concerns itself with intra-individual change during the test taking process. In doing so it challenges the assumption that the working speed of an individual is a constant throughout the test taking process. From this perspective it is conceptually similar to the latent growth model described under the second approach above. In spite of the conceptual resemblance between the two approaches, the CPA approach does not fit an alternative model and hence does not make an assumption of the growth trajectory, linear or non-linear. Instead, all it requires is the null model, which makes it also distinct from the second approach. Compared to the third approach, the CPA approach is model-based. It does not require the establishment of a threshold for each item or visual inspection. Further, CPA allows the estimation of the item position from which a test taker starts to show speededness, or the speeding point. This makes it distinct from all three approaches above. An estimated change point enables partial filtering of a response pattern instead of removing the entire response pattern of a suspected test taker, as typically done when one is flagged.

The rest of the paper is organized as follows. We first discuss the existing research on using CPA to address psychometric issues. Then we introduce the proposed CPA procedure to detect test speededness using item response time data. More specifically, two variations are discussed, one based on the likelihood ratio test and the other based on the Wald test. A simulation study is then described that evaluates the performance of the proposed procedure under various conditions, followed by a real data example. Lastly, we discuss implications of the findings and future directions.

Method

Very recently, CPA has provided testing professionals a new lens to understand test taking behavior at both the examinee and item levels (Shao et al., 2016; Sinharay, 2016; Yu & Cheng, 2019; Zhang, 2014). For example, Zhang (2014) used CPA to detect compromised items using time series of item usage data, for example, proportion of test takers answering each item correctly over a number of days. The most

relevant to the current study is Shao et al. (2016), which proposed a CPA procedure to detect test speededness using item response patterns. Consider a test of J items. Suppose that test taker i operates under time pressure on the last s_i items; in other words, he or she starts to show speededness from the $(J - s_i + 1)$ th item onward. Shao et al. (2016) tried to detect speededness by testing if there exists a drop in ability during the test taking process. This is done by comparing two likelihoods: a null likelihood (l_{i0}) assuming constant ability and an alternative likelihood (l_{ia}) assuming two separate abilities before and after the change point. Because the actual change point is unknown, the alternative likelihood l_{ia} assuming two separate abilities is computed for every possible change point, from after the first item to the $(J - 1)$ th item. If the maximum change in likelihood, that is, $\Delta l_i = \max\{l_{ia} - l_{i0}\}$ is significant, the null hypothesis will be rejected that the ability is non-changing. The point that leads to the largest l_{ia} is the estimated change point.

The significance testing on Δl_i requires the null distribution of Δl_i when there is no change point. Because there is no closed form distribution for Δl_i, in Shao et al. (2016) the null distribution was obtained by random permutation of the item responses. By referencing the sample Δl_i against the null distribution, a decision can be made to reject or retain the null hypothesis, that is, to determine whether a change point occurred or not. Shao et al. (2016) was the earliest study on using CPA to detect person-level aberrant response patterns, but it relied solely on the dichotomous item response data. Reliance on the permutation test to derive the null distribution of the test statistic also makes it computationally cumbersome. Later Sinharay (2016) pointed out that asymptotic critical values previously obtained in Andrews (1993) are applicable even in the context of dichotomous item response data under certain regularity conditions, thus possibly alleviating the need of a computationally intensive permutation test.

In this study, we would like to introduce a new CPA procedure to detect speededness. This study differs from existing studies in several important ways. First, the new procedure is developed using continuous response time data. This study therefore allows us to evaluate the gain from using response time over dichotomous item responses. When item-level response times are available (and their availability has indeed greatly expanded in the past decade), they can be powerful resources to help control the quality of response data. Second, we will build on previous work by Andrews (1993) and Sinharay (2016) to establish generally applicable critical values for easy implementation of the procedure. The permutation-based method to establish cutoffs in Shao et al. (2016) is computationally prohibitive when the sample size and test length increase, and greatly limits its applicability to large datasets. In this paper we validated the use of generally applicable critical values by simulations, and illustrated how they can be used on a large dataset. The gain in computational efficiency is clearly demonstrated in the real data example. Further, this study closely examines the estimation of the actual change point, and identify factors that influence the accuracy of change point estimation. The change point itself can be important to testing

programs, for example, to determine appropriate test length. It is therefore an imperative to gain deeper understanding of what affects change point estimation.

CPA for Item Response Time Data

In this paper we introduce two CPA test statistics to detect test speededness using item response time data instead of item response data. Sinharay (2017) provided a general framework for the applications of CPA to psychometric research, as well as guidelines regarding the choice of test statistics and critical values. In the remainder of the paper we mostly follow his notations.

Let $X_1, X_2, \cdots, X_J$ be independent random variables and suppose $X_j (j = 1, 2, \cdots, J)$ has probability function of $f(X_j; \pi)$, where π is the underlying latent variable governing the distribution of X_j. Testing if a change has occurred amounts to testing the null hypothesis that π has changed significantly at one point t, or in other words, testing whether it is true that the distribution of $X_j, j = 1, 2, \cdots, t$ is governed by $f(X_j; \pi_1)$ whereas the distribution of $X_j, j = t+1, t+2, \cdots, J$ is governed by $f(X_j; \pi_2)$. In the context of educational testing, $X_1, X_2, \cdots, X_J$ can be item-level data, such as item response data or item response time data, where J is the test length. If item response data are used, an item response function (IRF) given by an IRT model such as the two-parameter logistic model (2PLM) could be the $f(X_j; \pi)$. If item response time data are used, there are also many choices of models that could serve as the $f(X_j; \pi)$, including the log-normal model, Weibull model etc. that are discussed in the introduction. For a recent review of some popular models for response time in psychological and educational research, please see Cheng et al. (2017).

In this study we choose the log-normal model as the $f(X_j; \pi)$, which has been shown to fit well empirical response time data from high-stakes educational testing (van der Linden, 2006) and has been used widely to model item response time in achievement testing (Fan et al., 2012; van der Linden & van Krimpen-Stoop, 2003). Following the standard notation of the log-normal model, the density of response time t_{ij}, that is, the response time of test taker i to item j, takes the following form:

$$f(t_{ij}; \tau_i, \alpha_j, \beta_j) = \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} exp\{-\frac{1}{2}[\alpha_j(\ln t_{ij} - (\beta_j - \tau_i))]^2\}, \quad (1)$$

or equivalently

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

where $\beta_j \in (-\infty, \infty)$ is the item time parameter. A larger β_j suggests the item tends to be more time consuming. The $\tau_i \in (-\infty, \infty)$ is the working speed parameter for test taker i. τ is assumed to be normally distributed in the population. In addition, $\alpha_j \in (0, \infty)$ is the inverse scale parameter or time discrimination parameter. The role of α_j is analogous to the discrimination parameter in the 2PLM in IRT. The larger

value α_j takes, the less dispersion for $\ln(t_{ij})|\tau_i$. An item with a larger time discrimination parameter therefore tend to better distinguish respondents with high or low τ. Detection of speededness essentially amounts to detecting an increase in the working speed parameter, τ, for each test taker. Next we introduce how to use the CPA procedure based on the likelihood ratio test and the Wald test to detect the increase in τ.

CPA Based on the Likelihood Ratio Test

Assuming that the log-normal model in Equation (1) fits and local independence holds, the likelihood function of observing an item response time pattern $\mathbf{t_i}$ for person i is:

$$L(\tau_i; \mathbf{t_i}) = \prod_{j=1}^{J} \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} exp\{-\frac{1}{2}[\alpha_j(\ln t_{ij} - (\beta_j - \tau_i))]^2\}, \tag{3}$$

where $\mathbf{t_i}' = (t_{i1}, t_{i2}, \cdots, t_{iJ})$ captures the response time of person i to all items on the test. The log-likelihood function is therefore

$$l(\tau_i; \mathbf{t_i}) = lnL(\tau_i; \mathbf{t_i}) = \sum_{j=1}^{J} \ln \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} - \frac{1}{2}\sum_{j=1}^{J}\{[\alpha_j(\ln t_{ij} - (\beta_j - \tau_i))]^2\}. \tag{4}$$

Similar to Shao et al. (2016), the likelihood ratio test is formulated as follows:

$$\Delta l_i^{(k)} = -2(l_i^{H_0} - l_i^{(k)}), \tag{5}$$

where $l_i^{H_0} = l(\hat{\tau}_{i,0}; \mathbf{t_i})$ is the log-likelihood computed following Equation (4) by plugging in $\hat{\tau}_{i,0}$, the working speed parameter estimate for person i using item response time data from all J items. Assuming that an abrupt change occurs immediately after item k, the alternative log-likelihood given the change point k is computed as

$$l_i^{(k)} = l(\hat{\tau}_{i,k-}; \mathbf{t_{i(k-)}}) + l(\hat{\tau}_{i,k+}; \mathbf{t_{i(k+)}}), \tag{6}$$

where $\hat{\tau}_{i,k-}$ is the working speed parameter estimate using the first k items, $\hat{\tau}_{i,k+}$ is the working speed parameter estimate using items $(k+1)$ to J, the response time vector $\mathbf{t_i}'(\mathbf{k}-) = (t_{i1}, t_{i2}, \cdots, t_{ik})$ captures the response time of person i on item 1 to item k, and $\mathbf{t_i}'(\mathbf{k}+)$ captures the response time of person i on the remaining items.

Because the change point is actually unknown, the maximum of the $\Delta l_i^{(k)}$ overall possible change points is taken as the test statistic:

$$\Delta l_{\max,i} = \max_{k=1,2,\ldots,(J-1)} \Delta l_i^{(k)}. \tag{7}$$

An abrupt change is judged to have occurred in test taker i's item response time if the sample $\Delta l_{\max,i}$ is significantly larger than 0.

Similar to the likelihood ratio test based on item responses, there is no closed form distribution for $\Delta l_{\max,i}$ based on response time. Shao et al. (2016) obtained the null distribution of $\Delta l_{\max,i}$ through random permutation of item response data, which can be computationally intensive when the test is long. This makes it cumbersome to apply the CPA procedure. On the other hand, long tests are particularly susceptible to speededness. So there is great need in simplifying the procedure. Sinharay (2016) pointed out that if the change point does not appear too early or too late on the test (e.g., in the middle 70% of the test), the asymptotic critical values established in previous literature such as Andrews (1993) are directly applicable. In that case, detection of the change point can be done by comparing the sample $\Delta l_{\max,i}$ against the asymptotic critical values. The log-normal model for response time is included in the family of distributions considered by Andrews (1993), with the working speed parameter τ and the density function in Equation (1) playing the role of π and $f(X_j;\pi)$ in Andrews (1993), respectively. When τ is estimated by MLE and plugged into the equations to obtain $\Delta l_{\max,i}$, Theorem 3 of Andrews (1993) should hold as all the conditions for this theorem are satisfied. This implies that the asymptotic critical values should be directly applicable to the current study as well, as long as the change point does not occur too early or too late on the test.

In the context of detecting speededness, however, the change point may indeed be late because typically a test taker feels the pressure of the time limit toward the end of the test. Therefore in this study the critical values are obtained through Monte Carlo simulations, that is, by simulating data with no change point, to derive the null distribution. In the context of aberrant response detection, it is common to obtain critical values through computational approaches such as simulation, bootstrap, or permutation (e.g., Armstrong & Shi, 2009; Meijer, 2002; Shao et al., 2016). For instance, Worsley (1979) simulated 9,999 values under the null condition and the 1,000th, 500th, and 100th largest values were taken as the approximated critical value for of 10%, 5%, and 1% of nominal type-I error level for a one-sided test. Therefore in this study, a similar approach to Worsley (1979) is used to generate the null distribution and obtain the critical values. Bootstrap or permutation can also be used but they tend to be time intensive. Details of how we derive the null distribution using Monte Carlo simulations will be provided in the simulation section below.

In the case that $\Delta l_{\max,i}$ is significantly above the critical value at a certain α level, an abrupt change is said to be detected at that significance level and an estimate of the change point would also be computed. For the specific purpose of detecting certain type of aberrant response behavior, we would expect the change in τ to be in a certain direction. For example, in order to detect speededness specifically, the τ would be expected to increase after the change point, that is, $\tau_{i,k+} > \tau_{i,k-}$. In practice it is always important to check the direction of the change in order to understand if the aberrant behavior is the type of interest. The next item to the point at which $l_i^{(k)}$ is maximized is taken as the estimate of the speeding point, or more specifically, the item from which person i starts to show speededness. In other words, person i was estimated to have been affected by speededness on the last $\hat{s}_i$ items, where

$$\hat{s}_i = J - \arg\max_{k=1,2,\ldots,(J-1)} \{l_i^{(k)}\}. \tag{8}$$

CPA Based on the Wald Test

When k is known, the Wald test tests whether the working speed of test taker i on the first k items is the same as that on the last $(J-k)$ items, that is, $\tau_{i,k-} = \tau_{i,k+}$. The Wald test statistic is formulated as below:

$$W_i^{(k)} = \frac{(\hat{\tau}_{i,k+} - \hat{\tau}_{i,k-})^2}{\frac{1}{I_{k-}(\hat{\tau}_{i,0})} + \frac{1}{I_{k+}(\hat{\tau}_{i,0})}}, \tag{9}$$

where $I_{k-}(\hat{\tau}_{i,0})$ is the observed Fisher information based on response time data from items 1 to k of test taker i, and $I_{k+}(\hat{\tau}_{i,0})$ is the same information computed based on items $(k+1)$ to J.

To test the null hypothesis that $\tau_{i,k+} = \tau_{i,k-}$ versus the one-sided alternative $\tau_{i,k+} > \tau_{i,k-}$ for all $k, k = 1, 2, \cdots, (J-1)$, the test statistic is defined as

$$W_{\max,i} = \max_{k=1,2,\ldots,(J-1)} W_i^{(k)}. \tag{10}$$

When k is known, $W_i^{(k)}$ follows asymptotically the χ^2 distribution with 1 degree of freedom. However, with unknown k and dependent $W_i^{(k)}$ s, there is no closed form distribution for $W_{\max,i}$. According to Sinharay (2016), the asymptotic null distribution of the maximum value of the Wald statistic and the likelihood ratio statistic are identical, as they both can be characterized as the supremum of the square of a standardized tied-down Bessel process. Similar to the likelihood ratio test, we can obtain the critical values for $W_{\max,i}$ through Monte Carlo simulations.

Once a significant change is detected, a change point is estimated accordingly. Similar to the likelihood ratio test, the number of speeded responses is estimated to be the number of items after the change point:

$$\hat{s}_i = J - \arg\max_{k=1,2,\ldots,(J-1)} W_i^{(k)}. \tag{11}$$

For both the likelihood ratio test and the Wald test, we need to obtain the estimate of the working speed parameter. For the Wald test, we also need to compute the Fisher information in Equation (9). Technical details for these computations are provided in the Appendix. In the description above, the item parameters in the log-normal model α_j and β_j are considered known. In reality they will need to be estimated based on the item response time data. After they are estimated they can be treated as known and unchanging. Then the working speed before and after each possible change point can be estimated as described in the Appendix. We followed van der Linden (2006) in estimating the item parameters, that is, using MCMC with Gibbs sampler. These structural parameters could also be estimated in a factor-analytic approach as done in Molenaar et al. (2015).

Simulation Study

A simulation study was conducted to evaluate the performance of the proposed CPA method in detecting speeded test takers using item response time data. The idea is to generate datasets that contain both regular response time patterns as well as response time patterns that are affected by speededness. By applying the CPA method on the generated datasets, we obtain information on the power (flagging the speeded examinees as speeded) and false positive rate/type-I error (flagging non-speeded examinees as speeded) of the detection of speededness.

Speeded Response Time Model. As explained earlier, the log-normal model has been chosen as the generating model for regular response time patterns. To generate the response time affected by speededness, one approach is to simulate the response time as fixed values such as 10, 20, or 30 seconds as done in van der Linden and Guo (2008). An alternative is to add a positive value to τ_i in the log-normal model, which results in

$$\ln(t_{ij}) = \beta_j - \tau_i - L + \varepsilon_{ij},\ \ \varepsilon_{ij} \sim N(0, \alpha_j^{-2}), \tag{12}$$

where L quantifies the increase in working speed caused by speededness. In van der Linden and van Krimpen-Stoop (2003), L was set at 0.375 and 0.750 in the simulation.

In practice, however, such constant response time or constant change of working speed is rarely observed; a gradual change is more likely to occur. In this study, therefore, we consider a generating model that considers gradual change. In modeling the impact of speededness on item responses, Wollack and Cohen (2004) proposed a gradual change model to allow for gradual decline in P_{ij}, the probability of answering item j correctly by test taker i. Goegebeur et al. (2008) further showed how to fit this model and estimate the model parameters. The 2PLM version of the gradual change model takes the following form:

$$P_{ij}^{*} = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]} * min(1, [1 - (\frac{j}{J} - \eta_i)])^{\lambda_i}, \tag{13}$$

where $\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}$ is the ordinary 2PLM, and η_i is the stage of the test at which examinee i starts to speed ($0 \leq \eta_i \leq 1$). For instance, an $\eta_i = .8$ suggests that test taker i has speeded responses on the last 20% of the test. The speededness rate parameter λ_i regulates how fast P_{ij}^{*} drops as the test progresses after the speeding point. This model has been used in many studies to simulate speeded responses (Goegebeur et al., 2008; Shao et al., 2016; Suh et al., 2012).

Similarly, in this study we propose a modified log-normal model of response time to capture response time under speededness:

$$\ln(t_{ij}) = (\beta_j - \tau_i + \varepsilon_{ij}) * min(1, [1 - (\frac{j}{J} - \eta_i)])^{\lambda_i},\ \ \varepsilon_{ij} \sim N(0, \alpha_j^{-2}). \tag{14}$$

The model is structured similarly to Equation (13). The parameters η_i and λ_i are interpreted the same way as in Equation (13) as well. When $\frac{j}{J} > \eta_i$, that is, when the test progresses past a certain stage as defined by η_i, the term $min(1, [1 - (\frac{j}{J} - \eta_i)])^{\eta_i}$ will be smaller than 1, meaning that the test taker will spend less time on this item than he/she would have if unaffected by speededness.

We choose this model as the generating model of response time affected by speededness for several reasons. First, it allows us to generate gradual change instead of abrupt change in working speed. As noted earlier, the former may be a more realistic scenario. On the other hand, the CPA procedure is known to be sensitive to abrupt shift in a random process. If we generate abrupt change and apply the CPA, it will certainly highlight the strength of the CPA procedure. By generating gradual change, we evaluate the robustness of our proposed procedure in situations that it is challenged. Luecht and Ackerman (2018) recently criticized common practices in simulation studies that employs a chosen IRT model and then evaluate parameter recovery or model fit by fitting the same model that is used to generate the data. Such setup favors the chosen model by design and does not show how robust the performance of a parameter estimation method or a model fit statistic can be. Instead, they suggest challenging modeling alternatives/choices by generating data from complex models that might better ''represent plausible and important features of real data.'' Our choice of a gradual change model over an abrupt change model to generate data follows the advice of Luecht and Ackerman (2018) in spirit. Second, it allows us to regulate the change point in a very straightforward manner through the parameter η_i. There exist other models that allow for graduate change, for example, the latent growth model proposed by Fox and Marianti (2016). However, it is not straightforward how the change point can be explicitly modeled in that framework. Hence, the evaluation of the estimation of the change point is unclear in that context. Last, we intentionally keep the current simulation study in every way possible parallel to Shao et al. (2016), which adopted the gradual change model of Wollack and Cohen (2004). The purpose of keeping these two simulation studies parallel in key aspects is to isolate the effect of the type of data used, that is, item responses versus item response time.

Simulation Design. Tests of 40, 60, and 80 items were simulated, and their time limit was set at 60, 90, and 120 minutes, respectively. The sample size was $N = 1,000$. The percentage of speeded test takers was set at 10% or 30%. The gradual change log-normal model in Equation (14) was used to generate the response time pattern for test takers who speeded. The response time data for the other test takers were generated following the regular log-normal model. If a test taker ran out of time, the test would terminate automatically. The response time of remaining items would be recorded as 0, meaning those items were unreached. In that case the test taker would be labeled as ''speeded'' without applying any statistical detection techniques.

For test takers who speeded, we followed Suh et al. (2012) and Shao et al. (2016) to generate $\lambda \sim logN(3.912, 1)$, the parameter that governed the rate of the drop in

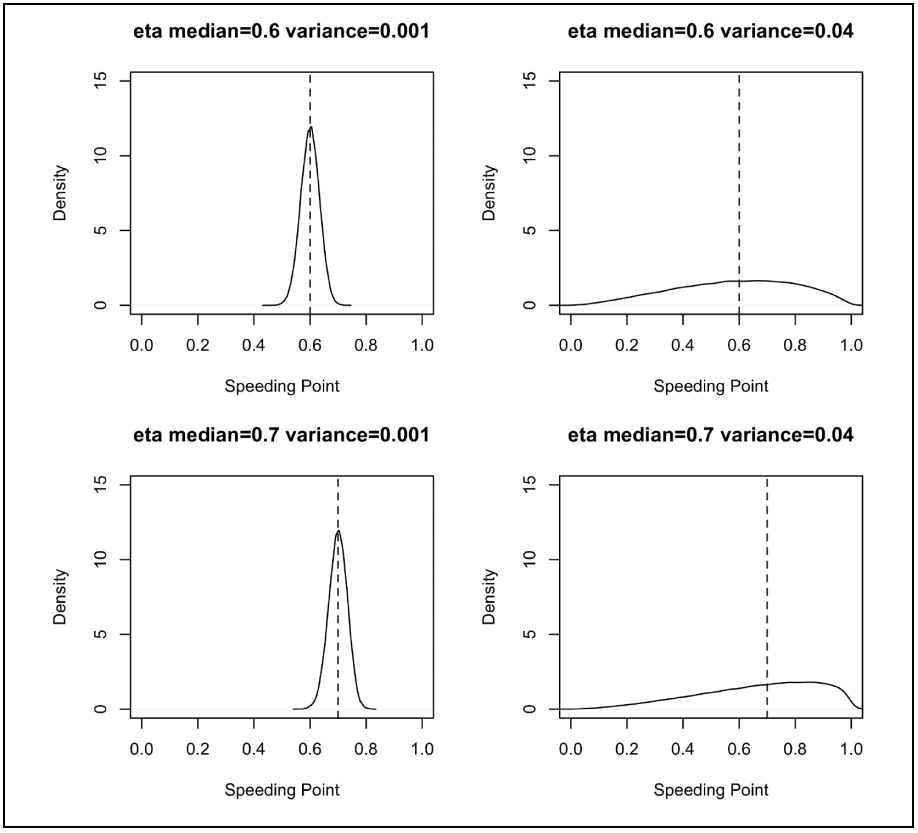

Figure 1. Distribution of the speeding point η.

response time. Also following Shao et al. (2016), the change point η was simulated from the beta distribution, more specifically with the median of η = .6 or .7, and η_{var} = .001 or 40×.001 = .04. The distribution of the change point is shown in Figure 1. Note that η reflects the change point as a percentage. In our simulation speededness starts from the next closest integer of $\eta \times J$. For a 40-item test, η of .6 indicates that a test taker shows speededness from item 25 onward. For a 60-item test an η of .7 means that a test taker starts to speed from item 43 onward. Figure 1 illustrates that with η_{var} of .001, the generated starting points were very close to the median; whereas when η_{var}, the possible change point were much more spread out and could occur anywhere on the test, including very late on the test. For our study, it is important to include cases in which the change point is close to the end of the test for two reasons. First, we are interested in the detection of speededness, which likely occurs late in the test. Second, it will allow us to evaluate if the asymptotic critical

Table 1. Mean (and *SD*) of the Critical Values for Wald and Likelihood Ratio Test.

J	$c_{.05}$	$c_{.01}$	$c_{.001}$
40	8.148 (0.09)	11.345 (0.19)	15.772 (0.59)
60	8.293 (0.09)	11.483 (0.20)	15.883 (0.64)
80	8.765 (0.09)	12.024 (0.20)	16.533 (0.64)

values can be directly applied as suggested in Sinharay (2016) when the change point is not in the midmost of the test.

In generating response time patterns for both speeded and non-speeded test takers, the inverse scale parameter (α) in the response time model was generated using a uniform distribution $U(1.75, 3.25)$. The time intensity parameter (β) was generated following Patton (2015) so that β has a mean of 4 and a *SD* of 1/3, and had a correlation of .3 with a and a correlation of .5 with b, where a and b are the discrimination and difficulty parameters in the 2PL IRT model, respectively. This is because in real data the time intensity parameter has often been found to correlate positively with those parameters. In Patton (2015), random normal deviates were added to a linear combination of the discrimination (a) and difficulty (b) parameters in order to achieve a correlation of .3 for $\rho_{\beta a}$ and a correlation of .5 for $\rho_{\beta b}$. The IRT item parameters were generated by $a\sim$ lnN(0, 0.5), and $b\sim$ N(0, 1) as done in Patton (2015). The working speed parameter for the $N = 1,000$ test takers was simulated by $\tau\sim N(0, .25)$, the same as Patton (2015). Overall there are 3 (test lengths) $\times$ 2 (percentage of speeded test takers) $\times$ 2 (η_{mean}) $\times$ 2 ($\eta_{variance}$) = 24 conditions. Each condition was replicated 50 times. The simulation was performed in **R** (R Development Core Team, 2014).

The goal of the simulation study is twofold. First, we would like to generate the null distribution of the likelihood ratio test and Wald test statistics and obtain critical values. These simulation-based critical values will be compared against the asymptotic critical values given in Andrews (1993) and Sinharay (2016). Second, based on the chosen critical values, we would like to examine the performance of the likelihood ratio test and Wald test in detecting speeded responses, specifically in terms of power and empirical type-I error.

To find the critical values, the null distribution was generated by simulating no speeded responses in the response time data, following the log-normal model in Equation (1) or (2). Similar to Worsley (1979), 10,000 response time patterns were generated under each condition, resulting in 10,000 test statistics, in our case $\Delta l_{\max,i}$ and $W_{max,i}$'s, from which the 500th, 100th, and 10th largest values were chosen as the approximates of $c_{.05}$, $c_{.01}$, and $c_{.001}$, the critical values corresponding to a nominal α level of .05, .01, and .001, respectively. Each of the null condition was replicated 1,000 times and the average of the $c_{.05}$, $c_{.01}$, and $c_{.001}$'s were taken as the empirical critical values. Table 1 provided the mean and SD of the critical values for each condition. The critical values were almost identical for Wald and likelihood ratio test, as expected following Sinharay (2016). Thus only results based on the Wald test were

shown in Table 1. The critical values, though varying substantially at different nominal α levels, did not differ much across different test lengths. As test length increased, there was a slight increase in the average critical values of 1,000 replications. The variance of the critical values across replications at each test length was small for $\alpha = .05$ or .01, suggesting that the critical values were rather stable. For $\alpha_{.001}$, the variance appeared much larger. This was not surprising since only a sample size of 10,000 examinees were simulated to obtain the null distribution for the test statistic. At $\alpha_{.001}$, it means we were picking the 10th largest value in the simulated values. Larger fluctuation was duly expected at that extreme end of the distribution.

Compared to the Table 1 in Sinharay (2016, p. 531), which listed the range of critical values to be between 8.45 and 9.84 for $\alpha = .05$, and 11.69 to 13.01 for $\alpha = .01$, the values in Table 1 here were not too far away from them albeit different. As explained earlier, the asymptotic values would be applicable when the test is long and when the change point occurs in the midmost of the test but not too early or too late. In our simulation the test length was finite and the change point could occur anywhere during the test. That might explain the small but visible differences. That said, the stability of the empirical critical values across test lengths and replications suggested the feasibility of using simple, fixed critical values, rather than nuanced test specific critical values. Hence, based on Tables 1, 8 and 11 were selected as the critical values at $\alpha_{.05}$, and $\alpha_{.01}$, respectively. These two integers were chosen to facilitate easy and straightforward application of the CPA procedure. Note that these two values were all slightly smaller than the average critical values in Table 1, so the resulting empirical type-I error rates were expected to be somewhat inflated. It would make sense to use these simple critical values if the inflation in type-I error is minimal, which will be examined next. To compare with the asymptotic critical values reported in Andrews' (1993), we also included the critical values of 8.85 and 12.35 at $\alpha = .05$ and .01 respectively in our simulations. Please note that the choice of critical values does not affect the estimation of the change location.

Results. There are two points of interest for us when we evaluate the performance of the proposed approach. The first is to examine the empirical power and false positive rate of detecting test speededness. Empirical power is defined as the percentage of test takers in our generated sample who have had speeded responses that are flagged by the proposed procedure. Empirical false positive rate is the percentage of test takers unaffected by speededness that are flagged by the proposed procedure.

The second point of interest is the performance of the proposed approach in estimating the actual change point. The actual change point, that is, the item right after which speededness exerts itself, is $J - s_i$, where s_i is the number of responses affected by speededness. The estimated change point is $J - \hat{s}_i$, where $\hat{s}_i$ is estimated by Equation (8) or (11). In the literature of CPA, the measure often used to capture the accuracy of change point estimation is the lag, that is, the estimated change point minus the actual change point (Shao et al., 2016; Sinharay, 2016). Numerically it is equivalent to $s_i - \hat{s}_i$. A positive lag suggests a delay in detecting the change point.

Statistically the lag is the bias of the change point estimate. The standard error of the lag is also computed. Note that across replications, a positive lag and a negative lag may cancel out. Therefore we also take the average of the absolute value of the lag, or AL_{mean}. To make the results across different test lengths comparable, $AL_{mean\%}$ is also calculated as the relative lag, that is, the absolute value of the lag divided by the test length.

Tables 2 to 4 provided the power and type-I error/false positive rate for each test length. Within each condition, the average of power and false positive rate of the 50 replications were taken. The first row for each of the three tables showed the type-I error rate when no test takers in the sample is affected by speededness. Across all conditions, except for test length of 80, the empirical type-I errors were only slightly higher than the nominal values. This suggests that using these simple, fixed critical values for medium to medium-long tests is indeed feasible. Meanwhile, the power was very high for every condition, regardless of the length of the test, the distribution of the change point, and the percentage of test takers affected by speededness. In contrast, Shao et al. (2016) reported power ranging from 0.60 to 0.90 at α of .05. Given that it was a study with an almost identical setup of simulations except for using item response data, the gain in power here seems to be largely attributable to the use of item response time data in this study. Aside from these general patterns, there were also fine and nuanced patterns. For example, power increased when the test was longer, or when more responses within a person were affected (i.e., $\eta_{median} = .6$ vs.$\eta_{median} = .7$). Using Andrews' (1993) critical values yields similar power across all conditions (in some conditions, slightly smaller than the proposed critical value results), but the type-I errors are consistently lower than the nominal α levels.

The column of % NF in Tables 2 to 4 showed the percentage of examinees who did not finish the test within the time limit. It can be found that under all conditions, the percentage of examinees who did not finish the test remained around 4% to 5% for a test length of 40, and around 5%–7% for a test length of 60 or 80, even when the percentage of test takers affected by speededness can be as high as 30%. This suggested that the majority of the simulees who were affected by speededness still finished their tests in time. Those who did not finish the test within the time limit would automatically be labeled as speeded. The fact that the power was 0.9 or above in Tables 2–4 suggested that the CPA was able to pick up the more subtle and evasive speededness, that is, being speeded but still finishing the test in time.

The last four columns of the tables presented the average absolute lag (*AL*) between the detected change point and the true change point, the absolute value of the lag divided by the test length, and the average lag (or bias) and RMSE of the change point estimate across replications in each condition, respectively. Theoretically the (absolute) lag can be computed for every examinee who was detected as speeded. Given that the power in Tables 2 to 4 was close to 1 for any nominal α level investigated here, it makes little difference which nominal α level we look at with respect to the computation of absolute lag. For every examinee

Table 2. Power, False Positive Rate, %*NF*, (Absolute) Lag, Bias, and RMSE for $J = 40$.

			Power				False positive rate								
%	η_{median}	η_{var}	$C_{.05}$	$A_{.05}$	$C_{.01}$	$A_{.01}$	$C_{.05}$	$A_{.05}$	$C_{.01}$	$A_{.01}$	% *NF*	AL_{mean}	$AL_{mean\%}$	Bias	RMSE
0	—	—	—	—	—	—	0.053	0.035	0.012	0.006	—	—	—	—	—
10	.6	.001	1.00	1.00	1.00	1.00	0.054	0.035	0.012	0.006	4.79	1.17	0.03	1.08	1.77
10	.7	.001	1.00	1.00	1.00	1.00	0.055	0.037	0.012	0.006	4.75	1.17	0.03	1.01	1.58
10	.6	.04	1.00	1.00	1.00	1.00	0.056	0.036	0.012	0.006	4.91	2.25	0.06	1.05	3.22
10	.7	.04	0.98	0.98	0.98	0.97	0.053	0.035	0.011	0.006	5.04	3.76	0.09	0.40	6.49
30	.6	.001	1.00	1.00	1.00	1.00	0.053	0.034	0.012	0.006	3.92	1.16	0.03	1.07	1.68
30	.7	.001	1.00	1.00	1.00	1.00	0.051	0.032	0.011	0.006	3.86	1.32	0.03	1.01	1.82
30	.6	.04	0.99	0.99	0.98	0.98	0.056	0.036	0.012	0.005	3.99	6.10	0.15	0.81	8.63
30	.7	.04	0.94	0.94	0.93	0.92	0.056	0.037	0.013	0.007	3.98	6.61	0.17	−0.14	9.69

$C_{.05}$ = applying the proposed simple critical value at $\alpha_{.05}$; $A_{.05}$ = applying Andrews' critical value at $\alpha_{.05}$; $C_{.01}$ and $A_{.01}$ = similar to $C_{.05}$ and $A_{.05}$ but at $\alpha_{.01}$; % *NF* = percentage of test takers that did not finish the test in time; AL_{mean} = the mean absolute lag between the actual and estimated change point across replications; $AL_{mean\%}$ = the mean absolute lag between the actual and estimated change point divided by the test length across replications; Bias = the bias of change point estimate; RMSE = the root mean square error of the change point estimate.

Table 3. Power, False Positive Rate, % *NF*, (Absolute) Lag, Bias, and RMSE for $J = 60$.

			Power				False positive rate								
%	η_{median}	η_{var}	$C_{.05}$	$A_{.05}$	$C_{.01}$	$A_{.01}$	$C_{.05}$	$A_{.05}$	$C_{.01}$	$A_{.01}$	% *NF*	AL_{mean}	$AL_{mean\%}$	Bias	RMSE
0	—	—	—	—	—	—	0.057	0.037	0.013	0.007	—	—	—	—	—
10	.6	.001	1.00	1.00	1.00	1.00	0.059	0.038	0.014	0.007	7.00	1.65	0.03	1.49	2.47
10	.7	.001	1.00	1.00	1.00	1.00	0.057	0.036	0.013	0.006	6.96	1.66	0.03	1.47	2.37
10	.6	.04	0.98	0.98	0.98	0.98	0.057	0.038	0.013	0.007	7.28	6.25	0.10	1.20	9.37
10	.7	.04	0.98	0.98	0.98	0.98	0.056	0.037	0.012	0.006	7.21	6.81	0.11	1.28	10.16
30	.6	.001	1.00	1.00	1.00	1.00	0.056	0.037	0.012	0.006	5.37	1.57	0.03	1.48	2.28
30	.7	.001	1.00	1.00	1.00	1.00	0.059	0.038	0.012	0.006	5.50	1.97	0.03	1.55	2.70
30	.6	.04	0.99	0.99	0.99	0.99	0.058	0.038	0.013	0.006	5.53	8.83	0.15	1.17	12.74
30	.7	.04	0.97	0.96	0.96	0.96	0.059	0.038	0.013	0.007	5.99	11.63	0.19	0.27	16.01

Table 4. Power, False Positive Rate, % *NF*, (Absolute) Lag, Bias and RMSE for $J = 80$.

			Power				False positive rate								
%	η_{median}	η_{var}	$C_{.05}$	$A_{.05}$	$C_{.01}$	$A_{.01}$	$C_{.05}$	$A_{.05}$	$C_{.01}$	$A_{.01}$	% *NF*	AL_{mean}	$AL_{mean\%}$	Bias	RMSE
0	—	—	—	—	—	—	0.073	0.049	0.017	0.008	—	—	—	—	—
10	.6	.001	1.00	1.00	1.00	1.00	0.071	0.046	0.016	0.008	6.43	2.31	0.03	2.26	3.32
10	.7	.001	1.00	1.00	1.00	1.00	0.073	0.049	0.017	0.008	6.45	2.34	0.03	1.99	3.12
10	.6	.04	0.99	0.99	0.99	0.99	0.072	0.048	0.017	0.009	6.53	6.53	0.08	1.85	9.54
10	.7	.04	0.98	0.98	0.97	0.96	0.071	0.047	0.017	0.009	6.67	10.71	0.13	0.88	16.89
30	.6	.001	1.00	1.00	1.00	1.00	0.072	0.047	0.015	0.008	5.02	2.12	0.03	1.93	2.98
30	.7	.001	1.00	1.00	1.00	1.00	0.073	0.048	0.017	0.009	5.22	2.71	0.03	2.07	3.53
30	.6	.04	0.99	0.99	0.99	0.99	0.070	0.046	0.016	0.009	5.51	15.17	0.19	1.66	20.82
30	.7	.04	0.97	0.97	0.96	0.96	0.073	0.048	0.018	0.009	5.52	16.91	0.21	0.65	22.72

detected as speeded, the CPA yields an estimate of the same change point that corresponds to the maximum of the test statistics, regardless of the nominal α level. Thus in computing the AL_{mean}, $AL_{mean\%}$, bias, and RMSE, only values under $\alpha = .05$ were reported. The mean bias and mean absolute bias were small in all conditions where $\eta_{var} = .001$. When η_{var} increases to .04, the mean bias remains small, but the absolute bias and RMSE increases dramatically. This trend is more pronounced when the test is long. It suggests that in some instances the lags were positive and in some cases negative. Hence they cancel out in the computation of bias. The absolute bias and the RMSE better capture the variability in the estimate of the change point when that happens.

Tables 2 to 4 indicate that the change points were better estimated when η_{var} is small. This is because when the η_{var} is large, the change point can appear anywhere on the test, including very early or late on the test (see Figure 1). It would be very challenging to precisely locate the change point in those situations. Previous research such as Andrews (1993) and Hawkins et al. (2003) suggested the search of the change point be limited to the middle of the test. For example, Andrews (1993) suggested limiting the search to $j = n_1, n_1 + 1, ..., N - n_1$, where n_1 was set to be the closest integer of $.15N$. Equivalently, the change point would be restricted to roughly the middle 70% of a test. Note that the average (absolute) lags reported in Tables 2 to 4 were aggregated across different levels of true change point.

Figures 2 to 4, on the other hand, showed how well the change points were estimated at each true change point value at each test length. They presented the estimated versus the true change point for each condition when there were 10% of speeded test takers. The plots for 30% of speeded test generally showed the same pattern. Due to space limit, they are omitted from the manuscript and are available upon request. Each figure of Figures 2 to 4 included four panels. The two panels on the left showed the estimated against the true change point when $\eta_{median} = .6$ and the two right panels showed that when $\eta_{median} = .7$. The two top panels were created for small η_{var} whereas the bottom panels were for much larger η_{var}. Within each panel, the dark line showed the average of the estimated change point at each true change point over 50 replications, and the shaded area showed the 95% bound of the estimated change point over 50 replications. When the dark curve is close to the diagonal $Y = X$ (the dotted line), the bias is small. When the dark curve falls above $Y = X$, there exists positive bias; otherwise there exists negative bias. The shaded area between the two dashed curves indicates the variability in the change point estimate. Bigger distance between the two dashed curves at a given true change point suggests larger variability in the change point estimate at this location. Take Figure 2 as an example, which represented the conditions of test length of 40. In most cases the dark line hovered above $Y = X$ in the four panels, except when the change point occurred very late on the test. This suggests that on average there's delay in the detection of the change point at most change point locations. The fact that the dark line was close to $Y = X$ showed that overall at each true change point the positive bias in some replications and the negative bias in others mostly canceled each other

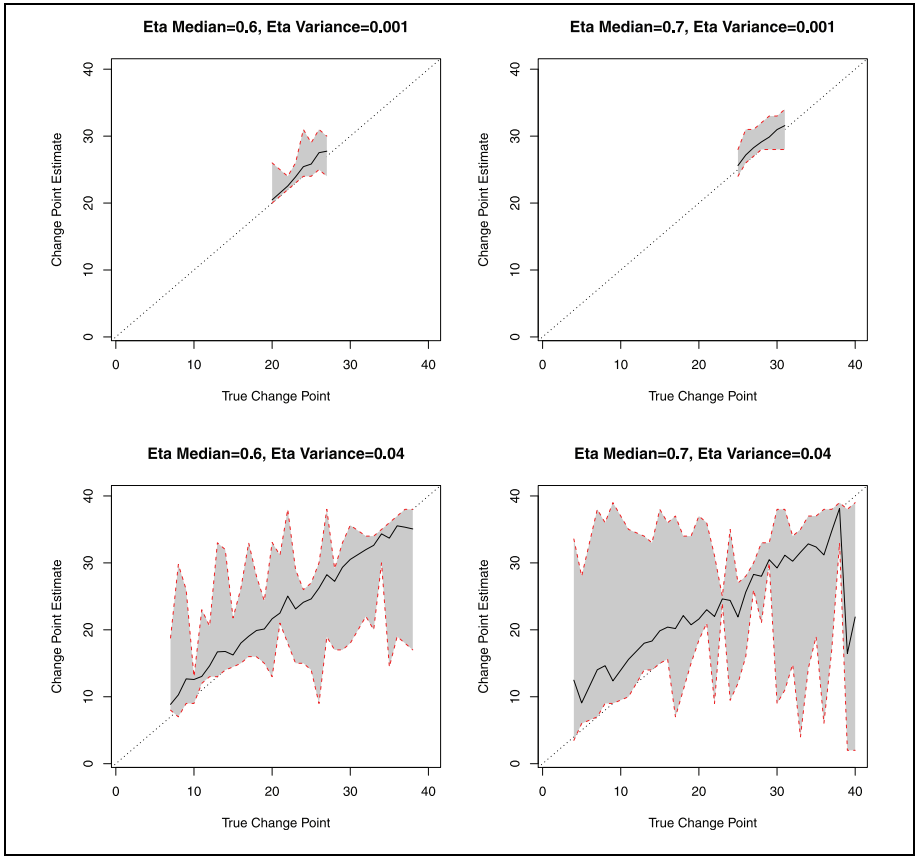

Figure 2. Estimated versus true change point with 10% of speeded test takers at test length of 40.

out. Meanwhile, larger η_{var} led to a wider range of true change point and a thicker shaded area, indicative of less stability in the change point estimate, consistent with the literature. The other figures showed a similar pattern: when the change point can occur very early or late on the test, the change point estimate was rather unstable. It warrants further investigation how to improve the estimation of the actual change point.

Real Data Analysis

To illustrate the application of the proposed CPA method, we performed CPA on the response time data of 50,000 test takers on a 30-item multiple-choice computer-based state assessment on mathematics. We cleansed the dataset by removing test

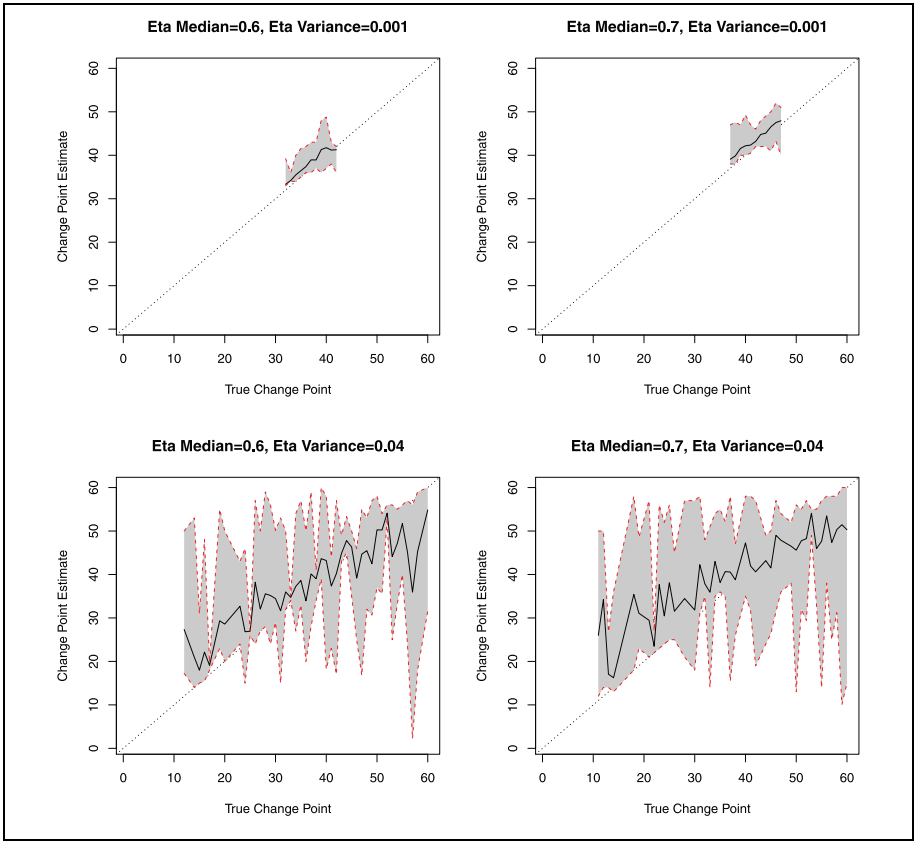

Figure 3. Estimated versus true change point with 10% of speeded test takers at test length of 60.

retakers and those who finished the entire test within 5 minutes. We also removed cases with response time of 0 on late items, that is, conspicuous cases of speededness and focused purely on detecting subtle cases of speededness. This resulted in a sample of around 46,000 test takers, which we randomly split into five samples each containing data from 9,200 students. This enables us to cross-validate our findings while having a large sample size for each sample.

We fitted the log-normal model to each sample of 9,200 response time patterns, and obtained the parameter estimates of α_j and β_j's. These structural estimates were treated as known and unchanging in estimating the person or incidental parameters, including $\tau_{i,0}$, $\tau_{i,k-}$ and $\tau_{i,k+}$'s. Then all these parameter estimates were used in the computation of likelihoods and the Wald test statistics. Eventually we obtained 9,200 $\Delta l_{\max,i}$ and $W_{max,i}$'s. Results were very similar using either statistic so only those based on $W_{max,i}$ were reported next. Again both the currently proposed critical values

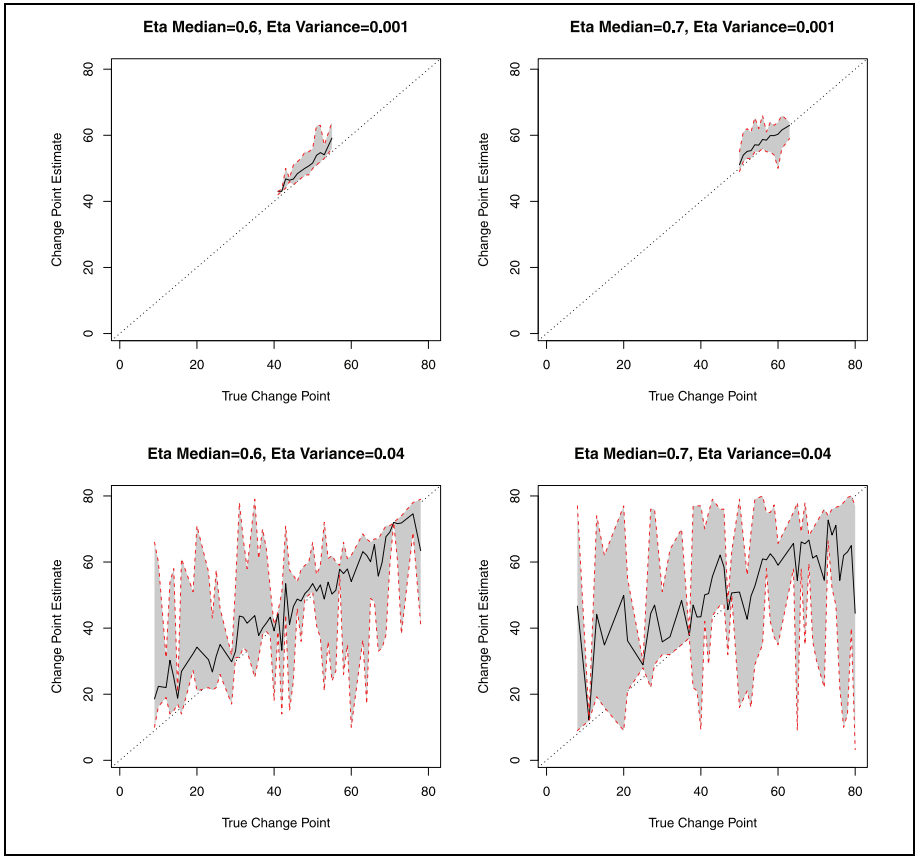

Figure 4. Estimated versus true change point with 10% of speeded test takers at test length of 80.

and Andrews (1993) values at $\alpha_{.05}$, and $\alpha_{.01}$ were applied in the detection. In addition, CPA was also carried out using the same 9,200 students with responses only data following Shao et al. (2016). To allow for more direct comparison, critical values were determined based on 100 random permutation of the item responses (which can be treated as null distribution) at $\alpha_{.05}$, and $\alpha_{.01}$ respectively. The process was replicated for all five samples.

First, we would like to highlight the computational gains by using the simple cutoffs. The analysis was run on a desktop with hardware specification as follows: 6 core, 2.90 GHz Intel Core i5-9400 processor, and 16.0 GB RAM. For the five samples, on one core, it took about 5 hours to run the CPA analysis with 100 permutations using item responses for each sample, and less than 1 minute using response times with simple or asymptotic cutoffs, with maximum memory used around 400M.

Table 5. Number of Test Takers Flagged and the Percent of Flagged Using Response Time and Item Response.

	$\alpha_{.05}$			$\alpha_{.01}$		
	Response time			Response time		
Sample	Current	Andrews	Response	Current	Andrews	Response
1	1384 (15.0%)	1152 (12.5%)	276 (3.0%)	812 (8.8%)	687 (7.5%)	41 (0.4%)
2	1442 (15.7%)	1205 (13.1%)	348 (3.8%)	835 (9.1%)	683 (7.4%)	71 (0.8%)
3	1450 (15.8%)	1214 (13.2%)	296 (3.2%)	837 (9.1%)	715 (7.8%)	58 (0.6%)
4	1427 (15.5%)	1203 (13.1%)	300 (3.3%)	831 (9.0 %)	702 (7.6%)	57 (0.6%)
5	1385 (15.1%)	1182 (12.8%)	270 (2.9%)	828 (9.0%)	688 (7.5%)	53 (0.6%)

This suggests that there is substantial gain in computational efficiency by using the simple or asymptotic cutoffs.

Second, Table 5 shows the number of test takes flagged as speeded by each type of critical values at $\alpha = .05$ (left) or $\alpha = .01$ (right) in five samples. Under $\alpha = .05$, 1,384 respondents were flagged in the first sample when CPA was applied to their response times using our proposed simple cutoff. This cutoff, as expected, was more liberal than Andrews' asymptotic cutoff, which led to 1,152 respondents being flagged in the same sample. If CPA was applied to responses instead of response times, only 276 respondents were flagged. The same trend was observed in the other four samples, that is, Andrews' asymptotic critical values tend to result in a bit fewer flagged test takers than the proposed simple critical values, and much fewer cases were flagged using only response data. The latter is consistent with what was reported by Shao et al. (2016) where their power in the simulation was lower than those shown in this study. Across five samples, the percentages of participants being flagged by each method remained largely stable, suggesting that the patterns we observe are unlikely due to chance, but rather robust.

Meanwhile, there are a few caveats to the results to highlight. First, using either the simple or asymptotic cutoff, response times led to over 80% of agreement with item responses, due largely to the agreement on the vast majority of test takers judged as non-speeded by both methods. Second, at $\alpha = .05$, roughly 15% to 16% or 12% to 13% of participants were flagged by response times, depending on the cutoff used. On the other hand, only about 3% of test takers were flagged by responses, which is even lower than the nominal α level. At $\alpha = .01$, using response times leads to 10 to 20 times more cases to be flagged than using responses. The latter led to again a smaller-than-nominal-level proportion of participants flagged. Third, only 20% to 30% of the cases flagged by the responses were also flagged by response times. This is consistent across all five samples, both α levels and both cutoffs. All these raise questions whether many cases flagged by responses were cases of type-I errors. This is a question to be best answered by a simulation study.

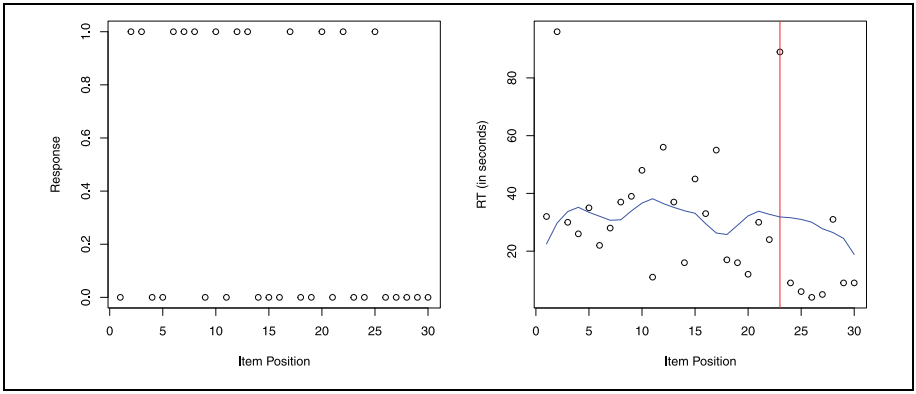

Figure 5. The response and response time pattern of test taker (first example).

Next, we further illustrate the different detection methods using two test takers' data. Figure 5 showed the response and response time (in seconds) pattern to 30 items of the first example. He or she was not flagged using response data, but was flagged by the CPA response time procedure at both $\alpha = .05$ and $\alpha = .01$ using either the simple or asymptotic critical values. The detected change point was shown as a vertical line in the right panel of the figure, which is around item 23. This indicates that his or her responses to the last 7 items were deemed affected by speededness using response time data, but no change was detected using response data. The blue curve in the right panel represents the median response time at each item position for all sampled students. It appeared that the response time on the last 7 items mostly hovered around 10 seconds, which were substantially lower than those on the first 23 items, as well as the medians at these item positions. This seems to lend support to the CPA response time procedure in flagging this test taker.

Figure 6 showed the response and response time (in seconds) pattern to 30 items of the second example, one of the few flagged by both response and response time using both types of critical values. Again, the detected change point was shown as a vertical line in both panels. Both indicate that there is a change point in the middle of the testing process. In the left panel, it shows that the test taker has a mixture of correct and incorrect responses prior to item 18 but nothing correct afterward. In the right panel, the response time was shown to be consistently much shorter than the median in the second half of the test. Though agreeing with each other in flagging this test taker, CPA-response procedure estimated the change point to be item 18, while CPA-response time procedure led to an estimated change point of 15. Given that there are too few test takers flagged by both data sources, we would not like to over-generalize on the comparison of change point estimation, but this is definitely something worth more attention in a future study.

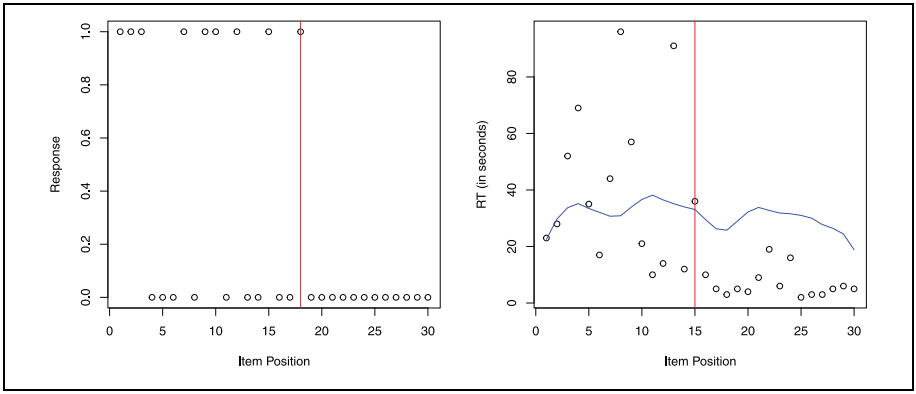

Figure 6. The response and response time pattern of test taker (second example).

Conclusions and Discussion

This study proposes a CPA method to detect test speededness using response time data. Its performance of detecting speededness is demonstrated and evaluated in a simulation study and a real-data example. In the simulation study, the proposed method shows high power in detecting speeded examinees while keeping the false positive rate well controlled, even when simple and fixed critical values are used. The power in this study is substantially higher than in Shao et al. (2016), a study that's parallel in simulation design but used item response data. As explained earlier, because response time data are continuous, we would expect the power to improve compared against dichotomous item response data but the extent to which power can be improved is unknown. This study showed that the improvement in power can be substantial without any inflation of the false positive rate.

In addition to the success in detecting speeded responses, the proposed method is also very flexible. In this study we assumed that the log-normal model fits the response time data, but the method is not bounded by that assumption. The CPA method can also be applied to other types of response time models such as the four-parameter response time model (Wang & Hanson, 2005) where a slowness parameter is incorporated. The slowness parameter can also be used for detecting test speededness. Similar to detecting an increase in working speed, we can use CPA to detect a decrease in the slowness parameter. Second, through the simulation study we were able to demonstrate that it is possible to use simple and fixed critical values, which makes the application of the CPA method straightforward. The critical values are similar to the asymptotic values reported in Andrews (1993) and used in Sinharay (2016), suggesting that they are rather independent of item parameters. It also means that it is unnecessary to re-conduct the simulation to update critical values when small changes occur to the test, for example when one item for some reason has to be removed. Third, the general principle of the CPA method and the test statistics

discussed in this paper can also be applied to detect aberrant responses on tests with polytomous items, as well as mixed-format tests which contains both dichotomous items and polytomous items. In addition, the simulation showed how the CPA can be applied to response time data in a linear test. Given the stability of the critical values across different tests (indicated by the very small standard deviation in the critical values in Table 1) and test lengths, we expect the method to be applicable to CAT or a multi-stage testing. Last but not least, the CPA method using response time data can be applied to detect other types of aberrant responses. For instance, fatigue and inattentiveness can manifest themselves in similar fashions to speededness, that is, reduced response time. The CPA can be a very promising approach to detect inattentiveness on a low-stakes survey if item response time is recorded, particularly when inattentiveness starts in the middle of the test. In survey research this is often referred to as back random responding or BRR, that is, respondents provide inattentive responses on the later portion of the assessment (Clark et al., 2003; Meade & Craig, 2012; Yu & Cheng, 2019).

In the meantime, this study has several limitations. First, the CPA procedure based on response times versus based on responses make very different assumptions. For this reason we would like to suggest caution against over-generalizing the findings from this study. The assumption of the CPA-response procedure is that speededness manifests itself in performance decline. The assumption of the CPA-response time procedure is that speededness will manifest in faster responding, irrespective of performance. In the literature, different definitions of speededness exist. For example, some defines speededness as ''the situation where the time limits on a standardized test do not allow substantial numbers of examinees to fully consider all test items'' (Lu & Sireci, 2007). One can argue whether the consequence of not having enough time to fully consider all test items is having unreached items, or rapid responses or guessing on some items, and/or performance decline on them. In fact some seem to suggest that both should be considered. Schnipke (1997) stated that test takers affected by speededness ''will have very fast response times and the responses will be at or near chance levels of accuracy.'' The first half of the statement is consistent with the assumption of CPA-response time procedure, while the second half is more in line with the assumption of CPA-response procedure. There are reasons to question the assumption associated with responses. For one, time pressure and associated anxiety does not necessarily lead to performance decline. A small amount of test anxiety could act as motivation and can improve performance (Akanbi, 2013). In addition, for test takers with very low ability, their performance may seem unaffected or even boosted by rapid guessing. Meanwhile, literature has shown how heightened anxiety can be associated with decreased processing efficiency and slow reaction time (Eysenck et al., 2007; Nishisato, 1966) in certain scenarios. That means the assumption associated with response time can also be challenged. Such complex effect of time pressure and the associated anxiety on item responses and response times are not considered in this study. Therefore the comparison between the current study and Shao et al. (2016) is based on simplified assumptions and should only be

interpreted as such. Given the high power of the CPA procedure based on response times to detect test speededness when its assumption holds, the next urgent task is to evaluate the validity of the assumption: Does test speededness manifest in faster response times? If so, in what context (e.g., high-stakes testing) and to what extent? Future research is certainly warranted in this area.

If one accepts the stringent assumption of Schnipke (1997), that is, speededness is characterized by faster response times *and* reduced response accuracy, one could potentially leverage information from both item responses and response time to detect test speededness. It is possible to develop a CPA procedure that directly utilizes both data sources based on a model that models item responses and response time simultaneously, for example, the hierarchical model by van der Linden (2007). It models at the first level the item responses by a three-parameter normal ogive model and the item response time by the same log-normal model used in this study. At the second level, the examinee ability θ and working speed τ are assumed to follow bivariate normal distribution. Instead of testing the change in θ and τ separately, one can test the change in the vector of $(\theta, \tau)'$ using the Wald test (authors, 2019).

Second, applying the CPA procedure introduced in this paper requires knowledge of the pre-change and post-change distribution structure, with the unknowns being only the parameters in the distribution. In reality even the probability structure may be unknown, in which case data-driven quickest change detection methods as proposed in Li (2016) may be helpful. In the future, we would like to explore the application of model-free quickest change detection methods to psychometric research. Second, when a change point is detected, one can only infer what is the underlying cause of the change. As suggested earlier, low motivation and speededness could both lead to rapid guessing late in the testing process. We may be able to statistically detect an increase in the working speed and locate the point when that increase starts to occur, but the CPA will not be able to pinpoint the cause. It will take other sources of information, for example, expert domain knowledge, to identify the cause. Wang et al. (2018) proposed a two-stage approach through which normal and aberrant behaviors are distinguished in the first stage, and in the second stage different types of aberrant behavior such as rapid guessing versus cheating are separated. We could pursue a similar approach down the road.

Third, it is assumed in this study that there exist only one change point. In practice, we might see multiple change points, for example, when an examinee is affected by a warm-up effect in the beginning of the test and the speededness effect toward the end. In theory, one can search for the first change point, and then search another possible change point given the first change point. This is certainly a topic that is worth further exploration.

Yet another limitation shown in the current study is that estimation of the actual change point is unsatisfactory when the change point varies widely across test takers, particularly for long tests. This is a well known issue in CPA. As Andrews (1993) pointed out, when the change occurs very early or late, the detection of the change and the estimation of the change point can be very challenging. In the future, it may

be possible to leverage the information in both response time and item responses to better estimate the actual change point. This is certainly an area that we will pursue to improve.

In spite of these limitations, this study has strong practical implications to psychometric researchers and testing professionals. Developing methods to check examinees' behavior for possible aberrant responses is one of the most important quality control components in testing industry. Failing to address this issue will not only result in inaccurate item and ability parameter estimations and biased scores, but also poses a threat to the public due to misleading interpretations of examinees' performance. Some comprehensive tests can be very long as they need to cover broad content. It needs to be thoroughly investigated the proper test length for these tests so that the majority of examinees will have enough time to finish (van der Linden, 2011). For a high-stakes test, examinees are motivated to give answers to all questions even when they are running out of time, which can result in rapid guessing. In that case, the proportion of examinees who have unreached items may be small, but this does not necessarily mean there is only a small proportion of speeded examinees. Thus a rigorous procedure such as the CPA method can be very helpful to understand the prevalence of speededness. Based on the findings of this study, we recommend testing programs record the item-level response time data in addition to item responses, and use CPA method to closely monitor aberrant responses during and after test administration. That said, one should exercise extreme caution when it comes to removing any response or test taker data, and one should refrain from relying solely on statistical results to make such decisions. As indicated by Allalouf et al. (2017), typically human review should follow statistical quality control procedures, and it should be no different when they are applied to testing.

Appendix

To calculate the Fisher information used in Wald test, we need to first get the second derivative of the log-likelihood function. According to Equation (4), the first derivative of the log-likelihood function is:

$$l'(\tau_i; \mathbf{t_i}) = -\sum_{j=1}^{J} \alpha_j^2 (\ln t_{ij} - (\beta_j - \tau_i)). \tag{A1}$$

Thus the second derivative of the log-likelihood function takes the following form:

$$l''(\tau_i; \mathbf{t_i}) = -\sum_{j=1}^{J} \alpha_j^2. \tag{A2}$$

By definition, the Fisher information of τ_i is given by:

$$I(\tau_i) = \sum_{j=1}^{J} \alpha_j^2. \tag{A3}$$

The MLE estimate of τ_i using all the response time ($\hat{\tau}_{i,0}$) can be obtained by setting $l'(\tau_i; \mathbf{t_i})$ to be 0 as shown below:

$$-\sum_{j=1}^{J} \alpha_j^2 (\ln t_{ij} - (\beta_j - \tau_i)) = 0, \tag{A4a}$$

$$\sum_{j=1}^{J} \alpha_j^2 \ln t_{ij} - \sum_{j=1}^{J} \alpha_j^2 \beta_j + \sum_{j=1}^{J} \alpha_j^2 \tau_i = 0, \tag{A4b}$$

$$\sum_{j=1}^{J} \alpha_j^2 \tau_i = \sum_{j=1}^{J} \alpha_j^2 \beta_j - \sum_{j=1}^{J} \alpha_j^2 \ln t_{ij}, \tag{A4c}$$

$$\hat{\tau}_i = \frac{\sum_{j=1}^{J} \alpha_j^2 \beta_j - \sum_{j=1}^{J} \alpha_j^2 \ln t_{ij}}{\sum_{j=1}^{J} \alpha_j^2}, \tag{A4d}$$

which is also given in van der Linden (2008). $\hat{\tau}_{i,k-}$ and $\hat{\tau}_{i,k+}$ can be calculated in a similar fashion.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work is partially supported by NSF grant SES-1853166 awarded to the corresponding author.

ORCID iD

Can Shao https://orcid.org/0000-0002-4528-6714

References

Akanbi, S. T. (2013). Comparisons of test anxiety level of senior secondary school students across gender, year of study, school type and parental educational background. *IFE PsychologIA*, *21*, 40–54.

Allalouf, A., Gutentag, T., & Baumer, M. (2017). Quality control for scoring tests administered in continuous mode: An NCME instructional module. *Educational Measurement Issues and Practice*, *36*, 58–68.

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

Armstrong, R. D., & Shi, M. (2009). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement*, *33*(5), 391–410.

Authors. (2019). *Bivariate change-point analysis for response accuracy and response time data*. Paper presented at the International Meeting of the Psychometric Society, Santiago, Chile.

Bhola, D. S. (1994). *An investigation to determine whether an algorithm based on response latencies and number of words can be used in a prescribed manner to reduce measurement error* (doctoral dissertation). Retrieved from ProQuest Dissertations Publishing (Accession No. 9519527)

Cheng, Y., Diao, Q., & Behrens, J. T. (2017). A simplified version of the maximum information per time unit method in computerized adaptive testing. *Behavior Research Methods*, *49*(2), 502–512.

Clark, M. E., Gironda, R. J., & Young, R. W. (2003). Detection of back random responding: Effectiveness of MMPI-2 and personality assessment inventory validity indices. *Psychological Assessment*, *15*, 223–234.

Demars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment*, *12*, 23–45.

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

Entink, R. H. K., van der Linden, W. J., & Fox, J.-P. (2009). A Box-Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, *62*, 621–640.

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

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, *7*(2), 336–353.

Fan, Z., Wang, C., Chang, H. H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37*, 655–670.

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

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

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

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students rapid guessing responses using response time. *Applied Measurement in Education*, *29*(3), 173–183.

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

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*, 606–619.

Lee, Y.-H., & Chen, H. (2011). Using response time to investigate students test-taking behaviors in a NAEP computer-based study. *Psychological Test and Assessment Modeling*, *53*(3), 359–379.

Lee, Y.-H., & Jia, Y. (2014). A review of recent response-time analyses in educational testing. *Large-scale Assessments in Education*. https://largescaleassessmentsineducation.springeropen.com/articles/10.1186/s40536-014-0008-1

Li, H. (2016). Data driven quickest change detection: An algorithmic complexity approach. In 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain.

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

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement Issues and Practice*, *26*(4), 29–37.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika*, *58*, 445–469.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455.

Meijer, R. R. (2002). Outlier dDetection in hHigh-sStakes-Stakes cCertification tTesting. *Journal of Educational Measurement*, *39*(3), 219–233.

Meyer, J. P. (2010). A mixture rasch model with item response time components. *Applied Psychological Measurement*, *34*, 521–538.

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. *Multivariate Behavioral Research*, *51*, 606–626.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*, 197–219.

Nishisato, S. (1966). Reaction time as a function of arousal and anxiety. *Psychological Science*, *6*(4), 157–158.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*, 200–219.

Patton, J. M. (2015). *Some consequences of response time model misspecification in educational measurement* (doctoral dissertation). Retrieved from ProQuest Dissertations and Theses (Accession No. 3648257)

Ranger, J., & Kuhn, J.-T. (2012). A flexible latent trait model for response times in tests. *Psychometrika*, *77*(1), 31–47.

Ranger, J., & Kuhn, J.-T. (2013). Analyzing response times in tests with rank correlation approaches. *Journal of Educational and Behavioral Statistics*, *38*, 61–80.

Ranger, J., & Ortner, T. (2012). A latent trait model for response times on tests employing the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, *65*, 334–349.

R Development Core Team. (2014). *R: A language and environment for statistical computing [Computer software manual]*. Vienna, Austria: Author. http://www.R-project.org

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, *68*, 589–606.

Schnipke, D. L. (1995). Assessing speededness in computer–based tests using item response times. *Dissertation Abstracts International*, *57*(1), 759B. (University Microfilms No. 9617600).

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. (1997). Assessing speededness in computer-based tests using item response times. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Schnipke, D. L. (1999). *The influence of speededness on item-parameter estimation* (*Vol. 96*, No. 7) Law School Admission Council.

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

Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika*, *81*(4), 1118–1141.

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

Sinharay, S. (2017). Some remarks on applications of tests for detecting a change point to psychometric problems. *Psychometrika*, *82*, 1149–1161.

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*, 285–311.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*(2), 162–188.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179–203). Academic Press.

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

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*(1), 5–20.

van der Linden, W. J. (2011). Setting time limits on tests. *Applied Psychological Measurement*, *35*(3), 183–199.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*(2), 117–130.

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

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*, 327–347.

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

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

van der Linden, W. J., & Xiong, X. (2013). Speededness and adaptive testing. *Journal of Educational and Behavioral Statistics*, *38*, 418–438.

Wang, C., Chang, H.-H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 144–168.

Wang, C., Fan, Z., Chang, H.-H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*, 381–417.

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

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

Wang, T. (2006). *A model for the joint distribution of item response and response time using one-parameter Weibull distribution (CASMA research report 20*. Center for Advanced Studies in Measurement and Assessment.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323–339.

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

Wise, S. L., & Ma, L. (2012). Setting response time thresholds for a CAT item pool: the normative threshold method. Paper presented at the 2012 Meeting of the National Council on Measurement in Education, Vancouver, Canada.

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

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

Yu, X., & Cheng, Y. (2019). A change-point analysis procedure based on weighted residuals to detect back random responding. *Psychological Methods*, *24*, 658–674.

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