Detecting Speed Anomalies in Examinees' Test-Taking Process Based on Response Time Data

Authors: Yunxi Xin, Qin Chunying, Dong Shenghong, Yaoyao Yu, Yu Xiaofeng, Qin Chunying, Yu Xiaofeng

Date: 2024-10-08T00:00:00+00:00

Abstract

Response time data, which contains information about test-takers' response speed and behavior, has garnered increasing attention in educational and psychological measurement. Speed anomalies exhibited by examinees during testing suggest potential aberrant response behaviors, such as prior exposure to specific test items. Based on response time data, this study constructs two statistical measures for detecting speed anomalies, applies these newly developed statistics to empirical data, illustrates the implementation process of the proposed methodology, and provides analysis and discussion of the findings. Simulation experiments are subsequently designed based on the empirical data analysis results and compared against the typical signed likelihood ratio test. The results demonstrate that different statistics exhibit varying sensitivities to speed differences, with the newly constructed statistics showing favorable performance in detecting speed variations among examinees during the testing process.

Full Text

Preamble

Self-Check Report for *Acta Psychologica Sinica* Submission

Please complete the following items and paste them on the first page of your manuscript.

1. List up to three innovative contributions of this study in the form of "Research Highlights," with a total word count not exceeding 200.

Acta Psychologica Sinica aims to publish cutting-edge psychological research that is "both scientifically excellent and of particularly broad interest and significance." If your study only makes minor incremental contributions, does not attempt to open new areas of inquiry, or lacks unique and innovative

perspectives—particularly if it merely studies algorithms or techniques without addressing clear psychological questions—its chance of acceptance by this journal is low. We recommend submitting to other journals instead.

Response: (1) Previous research has shown that the signed likelihood ratio statistic based on response time data is highly effective for analyzing aberrant test-taking behaviors such as item preknowledge. However, it suffers from low statistical power when the degree of aberrance is mild. Building on this, our study innovatively constructs two new statistics to detect item preknowledge during testing: the Bayes factor and posterior probability.

(2) We apply the signed likelihood ratio test and the two newly constructed statistics to a professional qualification exam dataset, yielding many meaningful results and comparing and discussing their detection effectiveness.

(3) Based on the empirical data analysis results, we designed targeted simulation experiments to compare the Bayes factor and posterior probability with the signed likelihood ratio statistic under various conditions. Results show that the newly constructed Bayes factor is most sensitive to examinees' speed differences and has the highest statistical power, while the posterior probability statistic demonstrates high consistency with the signed likelihood ratio test in identifying normal examinees.

2. Have you used the same data as in any previously submitted or published articles? If yes, please attach the article for review. (We do not encourage authors to publish multiple articles using the same data with identical variables, nor do we support splitting a series of related studies into multiple publications.)

Response: No, this study does not use data identical to any previously submitted or published work.

3. For non-experimental, non-intervention studies in management, clinical, personality, and social psychology that rely solely on self-report (questionnaire) methods, you must check for common method bias. What methods did you use to control for or demonstrate that such bias does not affect the validity of your conclusions? (For literature on common method bias, see: http://journal.psych.ac.cn/xlkxjz/CN/abstract/abstract894.shtml) Studies based on cross-sectional data with only self-reports and convenient sampling are easy to conduct but typically lack innovative value and have low acceptance probability.

Response: Not applicable to this article.

4. Did you report and analyze effect sizes (e.g., Cohen's d for t-tests; ^{2} or ^{2} for ANOVA)? (Many studies mechanically report effect sizes without necessary analysis or explanation, such as whether the effect size is small, medium, or large, or what theoretical or practical significance it holds.) (Search "effect size calculator" on Google

for convenient apps. For explanations of effect sizes in Chinese, see: http://journal.psych.ac.cn/xlkxjz/CN/abstract/abstract1150.shtml; in English, see: http://www.uccs.edu/lbecker/effect-size.html) Did you report 95% CIs for statistical analyses? (e.g., 95% CI for differences, 95% CI for correlation/regression coefficients) For calculations and plotting of confidence intervals, see https://thenewstatistics.com/itns/esci/

Response: Not applicable to this article.

5. Please state the planned sample size and actual sample size. If they differ, please explain why. Low statistical power due to insufficient sample size is a widespread problem in psychological research. We recommend explaining the basis for your sample size calculation in the Methods section. Sample size should be determined based on a justified effect size and desired power, and you should report the software or program used for the calculation. For rationale and practices regarding sample size planning, see https://osf.io/5awp4/

Response: Not applicable to this article.

6. To ensure completeness of data reporting, if you excluded any data in statistical analysis, did you report this in the text? What were the reasons? How would the results change if these data were included? How did you handle missing data? Did you delete any individual items when using scales? Why? How would the results change if these items were included? Are there any measured items or variables not reported? Why? Please indicate their location in the paper.

Response: Not applicable to this article.

7. For experimental materials, scales, or questionnaires that have not undergone peer review, are they attached at the end of the file for review? If not, please explain why. If this article is published, are you willing to share these materials with other researchers?

Response: Not applicable to this article.

8. This journal requires authors to provide raw data. Please choose one of the following options:

( ) a) Raw data and programs have been shared on the Psychological Science Data Bank (https://psych.scidb.cn/)
(√) b) Raw data and programs have been shared on the Psychological Science Data Bank (https://psych.scidb.cn/)
( ) c) If unable to provide, please explain the reason or provide relevant proof.

9. Is your study a clinical intervention or laboratory experiment? Yes ( ) No ( )

If yes, please provide pre-registration number: *{{{_}}{{}}}***{**_}___

If no, please explain the reason: *{{{_}}{{}}}***{**_}___

Note: Pre-registration (pre-register) is recommended for clinical interventions or laboratory experiments before data collection. Other experimental studies are also encouraged to pre-register. Pre-registration requires stating all research hypotheses and their support, as well as detailed procedures and steps. This journal's pre-registration website is https://os.psych.ac.cn/preregister (see "Download Center" on the journal website for instructions) or https://osf.io/ or https://aspredicted.org/. If your study is pre-registered, it will significantly increase the chance of acceptance. The importance of pre-registration can be referenced at https://osf.io/5awp4/

Response: Not applicable to this article.

10. If your study used human or animal subjects, was it approved by your institution's ethics committee? If yes, please send a scanned copy to the editorial office email. If no, please explain.

Response: This article does not involve human or animal subjects.

11. Have you written a 400-500 word extended English abstract following the "English Abstract Writing Guidelines" published on the editorial office website? Has the English title and abstract been reviewed by a native English speaker or professionally edited by an SCI/SSCI paper editing company?

Response: [Not provided in original]

12. If the first author is a student, the advisor must send a separate email to the editorial office (xuebao@psych.ac.cn) stating that they have read the paper and carefully reviewed it. Have you reminded your advisor to send this email? (The editorial office will only consider processing the manuscript after receiving the advisor's email.)

Response: [Not provided in original]

13. Please download and complete the "Manuscript Non-Confidentiality Certificate" from the "Download Center" on the right side of the editorial office website homepage, stamp it with the official seal of the corresponding author's institution, and send a scanned copy to the editorial office email (xuebao@psych.ac.cn). If there is no official confidentiality seal, please use the institution's official seal. Have you sent the email?

Response: [Not provided in original]

Detecting Speed Anomalies in Test-Taking Based on Response Time Data

Abstract

Response time data contains information about examinees' response speed and behavior, and is receiving increasing attention in educational and psychological measurement. Abnormal speed during testing indicates that an examinee may have engaged in aberrant test-taking behavior, such as having prior knowledge of some test items. Based on response time data, this paper constructs two statistics that can test for speed differences: the Bayes factor and posterior probability. The newly constructed statistics are applied to empirical data to demonstrate the use of the new methods, and the results are analyzed and discussed. Based on the empirical data analysis results, targeted simulation experiments were designed and compared with the typical signed likelihood ratio test. The results show that the newly constructed statistics have better performance in detecting examinees' speed differences during testing, especially when the degree of aberrance is low, where statistical power shows substantial improvement.

Keywords: response time, speed, posterior probability, difference detection, Bayesian factor

With societal development, the application scenarios of testing have greatly expanded, including entrance examinations, qualification tests, licensing exams, and more (Sinharay, 2021; Belov, 2014, 2016). Compared to lower-stakes tests, high-stakes tests that are closely related to examinees' interests tend to have higher rates of aberrant test-taking behavior (van der Linden, 2009). High-quality measurement data is a prerequisite for accurately assessing examinees' trait levels. However, actual measurement data may contain various types of aberrant "noise," such as data anomalies caused by cheating (Shu et al., 2013; Luo et al., 2020). Item preknowledge (Sinharay, 2017) is listed as one of the three most widespread types of test cheating in educational assessment (item preknowledge, test tampering, and answer copying) (Wollack & Schoenig, 2018). When examinees engage in aberrant test-taking behaviors such as item preknowledge, their response data often exhibits significantly different characteristics from their normal response patterns, and these anomalies reduce the quality of both individual and overall test data (Hong et al., 2021; Liu & Liu, 2022; Zhong et al., 2022), thereby causing a series of adverse effects on subsequent analyses, such as serious negative impacts on ability estimation, model fit, and reliability and validity (Oshima, 1994; Schnipke, 1996; Schnipke & Scrams, 1997; Lu & Sireci, 2007; Guo et al., 2010; Cizek & Wollack, 2017; Hong et al., 2020).

Detecting and handling aberrant test-taking behaviors is of great significance, and many researchers have sought various solutions. These studies can be broadly divided into two categories. The first category involves modeling aberrant test-taking behaviors, which incorporates examinees' aberrant behaviors into the model and evaluates their response patterns through model parame-

ters. For example, Wang et al. (2015) proposed a mixture hierarchical model (MHM) based on van der Linden's (2007) hierarchical response time model. Subsequent researchers have conducted a series of extensions based on MHM, including Lu et al. (2020), Ulitzsch et al. (2020), and Wang (2018). Liu and Liu (2021) summarized research on using mixture model methods to detect aberrant test-taking behaviors. The second category of research tests collected test data based on the assumption that examinees' latent traits (ability and speed) remain constant during testing. This category mainly includes person-fit tests (Sinharay, 2016) and change point analysis (CPA; Page, 1954). There are relatively more studies in this area, such as Bejar (1985), Sinharay (2016, 2021), Shao (2016), Sinharay et al. (2020), Liu et al. (2022), and Yu and Cheng (2022).

Existing research shows that when examinees engage in aberrant test-taking behaviors such as item preknowledge during testing, their response data exhibits changes in measurement characteristics. These changes are reflected not only in examinees' response scores but also in response time data (van der Linden, 2011; Cheng & Shao, 2022). The advantages of response time data over score data are mainly manifested in three aspects: (1) As continuous data, response time allows for more statistical analysis methods (Cheng & Shao, 2022; Wise & Kong, 2005); (2) Response time data reflects not only item measurement characteristics but also examinees' latent speed information (Marianti et al., 2014); (3) Response time data can be collected without examinees' awareness, minimizing impact on their test-taking behavior (Shao, 2016). In fact, with the popularity of computer-based testing, response time data is as easily collected as response score data. Response time data can be used not only to infer examinees' latent speed (van der Linden, 2010) but also to reveal information about test characteristics and examinee behavior that cannot be identified using score information alone (Fox et al., 2020). Additionally, as process data, response time has unique advantages in test data analysis, such as improving the accuracy of trait parameter estimation and facilitating the detection of aberrant response data (van der Linden & van Krimpen-Stoop, 2003; Fox & Marianti, 2016; Sinharay & Johnson, 2020; van der Linden & Guo, 2008; Pan & Wollack, 2021).

Wollack and Schoenig (2018) noted that testing for ability differences between different item sets is one of six statistical methods for detecting test fraud. The corresponding null hypothesis is that examinees perform similarly on two item sets (no difference in latent traits), while the alternative hypothesis is that examinees perform better on one item set. Sinharay and Johnson (2021) tested for examinee ability differences based on score data, where a significant difference between estimated abilities on one item set versus another indicates aberrant test-taking behavior. Previous research has shown that the signed likelihood ratio statistic performs well in detecting item preknowledge but suffers from low statistical power when the degree of aberrance is mild (Sinharay, 2017a, 2017b, 2020). Given the many advantages of response time data and its increasing availability, combined with its demonstrated effectiveness in detecting aberrant test-taking behavior (Sinharay, 2020; Cheng & Shao, 2022), this pa-

per proposes Bayesian-based speed difference detection methods using response time data. This includes two new approaches: a Bayes factor and a posterior probability-based test for response speed differences, which are compared with the typical signed likelihood ratio test.

The remainder of this paper is organized as follows: Section 2 introduces response time differences. Section 3 presents methods for testing response speed differences, first introducing the signed likelihood ratio test, then proposing Bayes factor and posterior probability-based speed difference test statistics using response time data. Section 4 applies the three detection methods to a widely studied empirical dataset, analyzes and discusses the results, and provides references for simulation study design. The simulation study is presented in Section 5, where the performance of the three speed difference test methods is comprehensively compared and evaluated under different test conditions. Finally, we discuss the study and future research directions.

2. Response Time Differences

Both ability measured from response score data and speed measured from response time data share a common assumption in measurement models: examinees' latent traits (ability or speed) remain fixed throughout the testing process (Cizek & Wollack, 2017). Sinharay (2021) tested for examinee ability differences based on response score data, which essentially tests the assumption that examinees' abilities remain constant when responding to two sets of items.

Inspired by this approach, this paper constructs statistics to detect item preknowledge based on response time data. Before formally introducing the item preknowledge detection methods based on response time differences, we first clarify the assumptions and notation involved.

A test taken by an examinee is divided into two parts, denoted as subtest 1 and subtest 2, where the examinee exhibits no aberrant behavior in subtest 1, while subtest 2 is suspected to contain aberrant behavior. Let 1 and 2 represent the examinee's response speed on subtests 1 and 2, respectively. For examinees without item preknowledge, the prior distribution of the speed difference 2 − 1 between the two subtests follows a normal distribution with mean 0 and standard deviation . For examinees with item preknowledge, the prior distribution of their speed difference 2 − 1 follows a normal distribution with mean and standard deviation . A descriptive example illustrates this below.

[Figure 1: see original paper] depicts the distributions of speed differences for two types of examinees when responding to 1 and 2, where the solid and dashed lines represent the density curves of examinees without/with speed differences 2 − 1, respectively. It can be seen that when examinees have no speed difference between 1 and 2, 2 − 1 follows a normal distribution with mean 0. In this example, when examinees have speed differences between 1 and 2, 2 − 1

follows a normal distribution with mean 1, indicating faster speed on subtest 2.

Figure 1. Distribution of Speed Differences

Note: The two green vertical lines in the figure represent the means of the speed difference distributions for two types of examinees (normal examinees and examinees with item preknowledge). It can be seen that normal examinees' speed differences between the two subtests follow a normal distribution with mean 0, indicating no speed difference, while aberrant examinees (those with item preknowledge) show speed differences following a normal distribution with mean 1, indicating the presence of speed differences.

Examinees' response scores on items reflect their mastery of the knowledge being tested, representing their ability. In contrast, examinees' response times (RT) on items reflect their response speed. Speed and ability have a compensatory relationship (van der Linden, 2007). Typically, examinees who respond too quickly show decreased accuracy, meaning their demonstrated ability decreases (Klein & Fox, 2009). Differences in examinee speed can be reflected through their response time data on items. We first present the response time model used in this paper, then introduce the signed likelihood ratio statistic for detecting item preknowledge based on response time data, followed by the construction process for the new indicators.

3. Testing Differences in Response Time Data

Van der Linden's (2006) lognormal response time model is widely popular and used. In this model, response times follow a lognormal distribution, and examinees maintain constant speed throughout the test. Let represent the number of examinees and represent the number of test items. Under this model, the probability density of the time spent by examinee on item can be expressed as:

$$f(t_i;\tau_n,\alpha_i,\beta_i)=\frac{\alpha_i}{t_i\sqrt{2\pi}}\exp\left\{-\frac{1}{2}[\alpha_i(\ln t_i-(\beta_i-\tau_n))]^2\right\},\quad(1)$$

where is the examinee's response time on item ; represents the -th examinee's response speed; β is the item time intensity parameter (larger β means more time spent on item); and α is the item time discrimination parameter (larger α means smaller variance in the response time distribution for item and better discrimination between examinees with different speed levels).

3.1 Signed Likelihood Ratio Statistic

As previously described, let the item sets in subtests 1 and 2 be denoted as and , representing sets of compromised items (items preknown by examinees in this paper) and uncompromised items (normal items), respectively. Let , , and

represent the examinee's response time data on all items ( ), compromised item set , and normal item set , respectively, with , being the corresponding log-transformed time data. Let and represent the examinee's response speed on compromised and normal items, respectively, with ˆ and ˆ being their estimates. ˆ represents the estimated response speed based on response time data from all test items. The likelihood ratio statistic based on response time data can then be given by:

$$\Lambda_T = 2[\ell(y_C|\hat{\tau}_C) + \ell(y_{\bar{C}}|\hat{\tau}_{\bar{C}}) - \ell(y|\hat{\tau})], \quad (2)$$

where $\ell(y_C|\hat{\tau}_C)$, $\ell(y_{\bar{C}}|\hat{\tau}_{\bar{C}})$, and $\ell(y|\hat{\tau})$ represent the log-likelihood functions based on speed $\hat{\tau}_C$ on compromised items, speed $\hat{\tau}_{\bar{C}}$ on uncompromised items, and speed $\hat{\tau}$ on all items, respectively. Under van der Linden's (2006) hierarchical modeling framework, the formula for calculating $\ell(y_C|\hat{\tau}_C)$ is:

$$\ell(y_C|\hat{\tau}_C) = \sum_{i\in C}\left[-\frac{1}{2}\log(2\pi) + \log(\alpha_i)\right] + \hat{\tau}_C\sum_{i\in C}\alpha_i - \frac{1}{2}\sum_{i\in C}\alpha_i(y_i - \beta_i)^2. \quad (3)$$

Therefore, detecting item preknowledge can test the null hypothesis H_0 based on the signed likelihood ratio (Sinharay, 2020). Under H_0, L_T follows an asymptotic standard normal distribution (Cox, 2006; Sinharay, 2017), and larger L_T values lead to rejection of the null hypothesis that examinees do not have item preknowledge.

$$L_T = \begin{cases} \sqrt{\Lambda_T} & \text{if } \hat{\tau}_C \geq \hat{\tau}_{\bar{C}} \\ -\sqrt{\Lambda_T} & \text{if } \hat{\tau}_C < \hat{\tau}_{\bar{C}} \end{cases} \quad (4)$$

Studies by Sinharay (2017a, 2017b, 2020) and Sinharay and Johnson (2020) have shown that compared to other existing methods, the signed likelihood ratio statistic L_s performs excellently in terms of Type I error rate and statistical power. Therefore, this paper only compares the two new response speed difference test indicators with the signed likelihood ratio statistic.

3.2.1 Bayes Factor Based on Response Time Data

On one hand, based on response score data, besides using the signed likelihood ratio test to detect item preknowledge, Sinharay and Johnson (2020) recommended using the Bayes factor (Kass & Raftery, 1995). The Bayes factor is a Bayesian statistical model comparison method that measures the probability that the data of interest fit a target model M_2 better than an alternative model M_1. Considering that this study focuses on response time data, the Bayes factor can be expressed as:

$$BF_{21} = \frac{p(t|M_2)}{p(t|M_1)}, \quad (5)$$

where $p(t|M_2)$ and $p(t|M_1)$ represent the marginal probabilities of response time data t under models M_2 and M_1, respectively, calculated as:

$$p(t|M_1) = \int p(t|\psi, M_1)p(\psi|M_1)d\psi, \quad p(t|M_2) = \int p(t|\psi, M_2)p(\psi|M_2)d\psi,$$

where $p(t|\psi, M_1)$ is the data distribution given parameters ψ under model M_1; $p(\psi|M_1)$ is the prior distribution of parameters under model M_1; and $p(t|\psi, M_2)$ and $p(\psi|M_2)$ have similar meanings. Larger BF_{21} values indicate stronger evidence supporting model M_2 fitting the data better. Researchers have provided guidelines for interpreting Bayes factor values in relation to evidence strength (Kass & Raftery, 1995).

For speed differences, testing can be viewed as a comparison between two alternative models. Model 1 assumes that item response time data are based on examinees having a fixed response speed (τ), while Model 2 assumes that response time data for item sets $\mathbf{C}$ and $\bar{\mathbf{C}}$ are based on two different response speeds (τ_1 and τ_2). Thus, under Models 1 and 2, the likelihood functions for examinee response times are $L(\tau; t)$ and $L(\tau_1; t_1)L(\tau_2; t_2)$, respectively.

Therefore, in the context of detecting speed differences, the Bayes factor can be calculated as:

$$BF_{21} = \frac{P(t|M_2)}{P(t|M_1)} = \frac{\int_{\tau_1=-\infty}^{\infty} \int_{\tau_2=\tau_1}^{\infty} L(\tau_1; t_1)L(\tau_2; t_2)p(\tau_1; \tau_2)d\tau_1 d\tau_2}{\int_{-\infty}^{\infty} L(\tau; t)\phi(\tau)d\tau}, \quad (8)$$

where $p(\tau_1; \tau_2)$ is the joint prior distribution of τ_1 and τ_2. Larger BF_{21} values indicate higher likelihood that the data fit model M_2.

3.2.2 Time Difference Testing Based on Posterior Probability

On the other hand, some researchers (Stern, 2005; Robert, 2007; Gelman, 2014) argue that a direct measure supporting the alternative hypothesis versus the null hypothesis is the posterior probability of the event corresponding to the alternative hypothesis. Following this approach, we consider detecting item preknowledge based on posterior probability using response time data.

Given an examinee's response time data t_1 and t_2 on subtests S1 and S2, the joint posterior distribution of speeds τ_1 and τ_2 is defined as $g(\tau_1, \tau_2|t)$. Under the local independence assumption, $g(\tau_1, \tau_2|t)$ is calculated as:

$$g(\tau_1, \tau_2|t) = \frac{L(\tau_1; t_1)L(\tau_2; t_2)p(\tau_1; \tau_2)}{\int_{\tau_1=-\infty}^{\infty}\int_{\tau_2=-\infty}^{\infty} L(\tau_1; t_1)L(\tau_2; t_2)p(\tau_1; \tau_2)d\tau_1 d\tau_2}. \quad (9)$$

Based on Equation 9, the posterior probability $P(\tau_2 \geq \tau_1|t)$ can be calculated as:

$$p(\tau_2 \geq \tau_1|t) = \int_{\tau_1=-\infty}^{\infty}\int_{\tau_2=\tau_1}^{\infty} g(\tau_1, \tau_2|t)d\tau_1 d\tau_2 = \frac{\int_{\tau_1=-\infty}^{\infty}\int_{\tau_2=\tau_1}^{\infty} L(\tau_1; t_1)L(\tau_2; t_2)p(\tau_1; \tau_2)d\tau_1 d\tau_2}{\int_{\tau_1=-\infty}^{\infty}\int_{\tau_2=-\infty}^{\infty} L(\tau_1; \tau_1)L(\tau_2; t_2)p(\tau_1; \tau_2)d\tau_1 d\tau_2}. \quad (10)$$

It can be seen that the integral formulas in the numerator and denominator of Equation 10 are the same, differing only in the integration limits. The integrals in Equation 10 need to be computed using numerical integration. Here, we use Riemann sum approximation, where the numerator and denominator can be calculated as Equations 11 and 12, respectively:

$$\int_{\tau_1=-\infty}^{\infty}\int_{\tau_2=\tau_1}^{\infty} L(\tau_1; t_1)L(\tau_2; t_2)p(\tau_1; \tau_2)d\tau_1 d\tau_2 \approx \sum_{\tau_{2m}>\tau_{1k}} L(\tau_{1k}; t_1)L(\tau_{2m}; t_2)p(\tau_{1k}; \tau_{2m})\Delta_1\Delta_2, \quad (11)$$

$$\int_{\tau_1=-\infty}^{\infty}\int_{\tau_2=-\infty}^{\infty} L(\tau_1; t_1)L(\tau_2; t_2)p(\tau_1; \tau_2)d\tau_1 d\tau_2 \approx \sum_{k=1}^{K}\sum_{m=1}^{M} L(\tau_{1k}; t_1)L(\tau_{2m}; t_2)p(\tau_{1k}; \tau_{2m})\Delta_1\Delta_2, \quad (12)$$

where $\tau_{11}, ..., \tau_{1k}$ are K equally spaced points, $\tau_{21}, ..., \tau_{2m}$ are M equally spaced points, $\Delta_1 = \tau_{1,k+1} - \tau_{1k}$, and $\Delta_2 = \tau_{2,m+1} - \tau_{2m}$. In this study, we use 101 equally spaced points between (-5, 5) to approximate the numerical integral, with a step size of 0.1. Based on examinees' response time data, larger probability values from Equation 10 indicate a higher likelihood that the examinee's response speed on item set $\mathbf{C}$ is greater than on item set $\bar{\mathbf{C}}$. To more clearly illustrate the role of the three response time-based statistics in detecting item preknowledge, a descriptive example is presented below.

3.3 A Descriptive Example

Consider a test with 20 items. For illustration purposes, the item time intensity parameter α and time discrimination parameter β are fixed at 2 and 1, respectively. The response time data for 8 examinees are designed in the simplest way, as shown in Table 1 . Specifically, examinees have preknowledge of some items among the last 10 items (the number of preknown items increases from 0 to 7 across the 8 examinees). This design allows examination of the methods' performance under both severe and mild item preknowledge scenarios, demonstrating

the robustness of the test methods. The approach involves hypothesis testing of response speed differences between the first 10 and last 10 items, with the null hypothesis being that examinees use the same response speed on both halves, and the alternative hypothesis being that their speed on the last 10 items is faster than on the first 10 items. Varying the number of preknown items from 0 to 7 in the last 10 items makes the simulation scenario more realistic, as in many applications only some items in the suspected set are actually preknown by examinees.

Table 1. Examinee Response Times in the Descriptive Example

Note: Items marked in red indicate items preknown by the examinee, resulting in shorter response times.

Based on the response data in Table 1 and the item parameters, three speed parameters are estimated for each examinee: based on the first 10 items, the last 10 items, and all 20 items. The SLR statistic and its corresponding p-value, Bayes factor BF, and posterior probability PP are calculated for each examinee. For these three statistics, larger values indicate a higher likelihood of response time differences between subtests (item sets), reflecting different response speeds across test sections.

Table 2. Statistic Values Calculated Based on the Descriptive Example

Note: Time difference is calculated as the difference in total response time between the first and last 10 items. In this example, a smaller difference indicates more similar response speeds between the two halves. $\hat{\tau}_1$, $\hat{\tau}_2$, and $\hat{\tau}$ represent speed parameters estimated based on the first 10 items, last 10 items, and all 20 items, respectively. Darker shading indicates more severe aberrance. The last three columns correspond to test statistics based on response time data, where SLR(p) denotes the signed likelihood ratio statistic with p-value in parentheses, and BF and PP correspond to Bayes factor and posterior probability values, respectively.

This descriptive example shows that the three statistics—SLR, BF, and PP—begin to provide evidence supporting response time differences between subtests (item sets) starting from Examinee 5, indicating different response speeds across test sections. According to Kass and Raftery's (1995) guidelines for interpreting BF values, larger BF values indicate stronger evidence of response time differences (i.e., speed differences) during the test. Examinees with larger numbers in the table show greater differences in response time between the first and second halves, and the statistics provide stronger evidence of speed differences. For example, the BF statistic shows an increasing trend from front to back. Using a significance level of 0.05, the SLR statistic indicates that the first 4 examinees show no speed difference, while the last 4 do. Using 0.95 as the cutoff, PP yields the same conclusion as SLR. The main reason why the statistics do not support speed differences for the first half of examinees is their small response time differences. This example simply illustrates the use of each statistic in detecting

response time differences (i.e., speed differences).

To evaluate the performance of the three response time-based statistics in detecting item preknowledge, this paper first analyzes a widely studied dataset (Kasli et al., 2023; Lee, 2018; Sinharay, 2017a, 2017b, 2020). This dataset has been analyzed and discussed in numerous studies on item preknowledge, allowing comparison of our methods' results with existing research. Furthermore, based on the analysis results, targeted simulation studies were designed to further evaluate the proposed methods' performance.

4. Empirical Data Analysis

The empirical data come from a computer-administered professional certification test. Detailed information about this dataset can be found in Cizek and Wollack (2017). We analyzed Form 1 of this dataset, which contains 170 dichotomously scored items with response and response time data from 1,636 examinees. After extensive investigation, the testing organization identified 64 compromised items in Form 1. Using various statistical methods and detection procedures, the organization flagged 46 examinees as suspected cheaters. This dataset has received widespread attention in recent years, with many researchers analyzing and studying it (Kasli et al., 2023; Lee, 2018; Sinharay, 2017a, 2017b, 2020).

Based on the identified compromised items, the entire test was divided into two parts: a set of normal items and a set of compromised items. To more intuitively examine whether the "flagged" aberrant examinees in the original dataset exhibited response speed differences, we conducted the following analyses on Form 1 data: (1) After removing data with missing responses and "flagged" aberrant examinees, we fit the lognormal time model to obtain item parameters. (2) After removing examinees with missing data, we estimated speed parameters for the 41 "flagged" aberrant examinees based on all 170 items, the 106 uncompromised items, and the 64 compromised items, with results shown in Figure 2 [Figure 2: see original paper]. (3) After removing examinees with missing data, we analyzed all 1,624 examinees using SLR, BF, and PP, and estimated speed parameters for the "flagged" aberrant examinees, as shown in Figures 3 [Figure 3: see original paper], 4 [Figure 4: see original paper], and 5 [Figure 5: see original paper]. It should be emphasized that not all "aberrant examinees" flagged in the original data showed greater speed on compromised items than on normal items, possibly because not all "aberrant examinees" in the original data were flagged due to item preknowledge.

Figure 2 [Figure 2: see original paper] shows that 24 flagged aberrant examinees had greater speed on compromised items than on normal items, represented in the figure by red lines being higher than green lines. Meanwhile, 17 examinees did not show this pattern. Since the original dataset did not provide further information about the types of flagged aberrant examinees, we cannot make further judgments about these 17 examinees.

Furthermore, using the item classification in Form 1 (uncompromised vs. com-

promised items) and the corresponding response time data, we applied the signed likelihood ratio SLR and the constructed Bayes factor BF and posterior probability PP to detect each examinee's response time data. The results compared with the original dataset's examinee flags are shown in Table 3 . It can be seen that all three statistical methods are more "conservative" than the original dataset's "aberrant examinee" flags. BF, SLR, and PP detected only 13, 11, and 9 "aberrant examinees," respectively. Moreover, the sets of examinees detected by these three methods have an inclusive relationship: BF's results include those detected by SLR and PP, and SLR's results include those detected by PP. Thus, PP is the strictest method in flagging aberrant examinees, while BF is the most lenient.

Table 3. Analysis Results of Three Statistics on Empirical Data

Method	Examinees Detected
BF	251, 324, 452, 453, 465, 524, 525, 624, 679, 700, 792, 993, 1252
SLR	251, 324, 453, 465, 524, 525, 624, 679, 700, 792, 993
PP	251, 324, 453, 524, 525, 624, 679, 792, 993

Note: Numbers in gray background and bold indicate examinees also flagged as "aberrant" in the original data. Critical values used were: $p < 0.05$ for SLR, $BF > 1$, and $PP > 0.95$.

We also analyzed the speed parameter estimates for these examinees, as shown in Figures 3, 4, and 5. The three methods all detected examinees whose response speed on compromised items was greater than on normal items. The three colored lines (red, green, and blue) in the figures correspond to speed estimates on compromised items, normal items, and the entire test, respectively. In Figures 3, 4, and 5, the three lines maintain the same order: red, blue, and green from top to bottom, with the distance between the red and blue lines greater than that between the blue and green lines, demonstrating the effect of item preknowledge on examinees' response speed.

Figure 2. Speed Parameter Estimates for "Aberrant Examinees" Flagged in Original Dataset

Note: (1) Dataset from Cizek and Wollack (2017); (2) The three lines correspond to speed parameter estimates for "aberrant examinees" flagged in the dataset, where the red line is based on compromised items, the blue line on all items, and the green line on uncompromised items.

Figure 3. Speed Parameter Estimates for "Aberrant Examinees" Flagged by BF

Figure 4. Speed Parameter Estimates for "Aberrant Examinees" Flagged by SLR

Figure 5. Speed Parameter Estimates for "Aberrant Examinees" Flagged by PP

Figures 3, 4, and 5 show that the "aberrant examinees" detected by the three methods all had greater response speed on compromised items than on normal items. The difference lies in that PP detected the fewest examinees, while BF detected the most. However, examinees detected by BF but not by PP still showed speed differences, indicating that the statistical power of BF, SLR, and PP decreases in that order.

Figure 6 [Figure 6: see original paper] depicts the response speed differences and mean speed differences for aberrant examinees detected by the three methods. The 13 green points are examinees flagged by BF, with 11 circled in purple representing those flagged by SLR, and 9 circled in blue representing those flagged by PP. The three horizontal lines correspond to the mean speed differences for examinees flagged by each method. It can be seen that the speed differences detected by BF, SLR, and PP are ordered from smallest to largest: 0.316, 0.326, and 0.340, respectively. This indicates that BF is most sensitive to speed differences, while PP is least sensitive, consistent with results from Table 3 and Figures 3, 4, and 5.

Figure 6. Average Speed Differences on Compromised and Normal Items for "Aberrant Examinees" Detected by Three Methods

Note: Green, purple, and blue points correspond to speed differences (speed on compromised items minus speed on normal items) for examinees detected by SLR, BF, and PP, respectively. The three horizontal lines represent the average speed differences for examinees detected by each method.

Figure 6 further demonstrates that the three methods differ in their sensitivity to detecting speed differences. To further evaluate the performance of SLR, BF, and PP under different test conditions, we conducted the following simulation experiment based on the empirical analysis results.

5. Simulation Study

Referencing the empirical data analysis results, we manipulated three variables: the proportion of preknown items, the proportion of examinees affected by item preknowledge, and the degree of impact from item preknowledge.

5.1 Experimental Design

Test length was set following Sinharay and Johnson (2021), with tests including 100 items. Item parameter distributions came from Cizek and Wollack (2017, p. 14), and examinee speed parameters were drawn from a standard normal distribution. The three manipulated variables were: (1) Proportion of preknown items, with fixed proportions randomly selected (three levels: 10%, 20%, 30%). (2) Proportion of examinees affected by item preknowledge, also with three levels: 5%, 10%, and 20%. The number of examinees was fixed at 1,000, resulting

in 50, 100, and 200 examinees with item preknowledge, respectively. (3) Degree of impact from item preknowledge (the amount by which speed parameters increase for affected examinees, with larger values indicating more severe impact). Referencing the empirical data analysis, we examined three levels: low U(0.20-0.35), medium U(0.35-0.50), and high U(0.50-1), representing that affected examinees' speed differences between compromised and normal items were drawn from the corresponding uniform distributions. This setting references the empirical data analysis results and uses relatively smaller speed difference values compared to previous studies that set speed differences at 1 or 2 (Wang et al., 2018; Zhu et al., 2023), making it more realistic for practical applications.

Response times for unaffected examinees on all items and for affected examinees on uncompromised items were simulated using the LRT model. For affected examinees on compromised items, we added the difference value to their speed parameters (Sinharay & Johnson, 2021). This design primarily examines the performance of the new methods under varying proportions of aberrant items, aberrant examinees, and degrees of aberrance, thereby evaluating the robustness of each method.

Table 4. Experimental Design

Factor	Levels (Values)
Proportion of preknown items	3 (0.1, 0.2, 0.3)
Proportion of examinees affected by preknown items	3 (0.05, 0.1, 0.2)
Degree of speed impact from item preknowledge	3 (U(0.20-0.35), U(0.35-0.50), U(0.50-1))

Note: The degree of speed impact has three levels corresponding to three uniform distributions, indicating that affected examinees' speed differences between compromised and normal items are drawn from the corresponding uniform distribution, i.e., $(\tau_2 - \tau_1) \sim U(a, b)$. The magnitude and range of speed differences can be controlled through a and b.

In total, there are $3 \times 3 \times 3 = 27$ experimental conditions, as shown in Table 4. Different combinations of speed difference levels and preknown item proportions represent varying severities of item preknowledge. The proportion of affected examinees represents the "prevalence" of item preknowledge in the examinee population. For each experimental condition, 20 datasets were generated and analyzed using the three methods, with corresponding evaluation metrics calculated.

5.2 Evaluation Metrics

Evaluation metrics include Type I error rate and statistical power. Type I error rate represents the proportion of normal examinees incorrectly identified as having item preknowledge, while statistical power represents the proportion of examinees with item preknowledge successfully identified. The formulas are:

$$\text{Statistical Power} = \frac{\text{Number of correctly identified preknowledge-affected examinees}}{\text{Total number of preknowledge-affected examinees}}$$

$$\text{Type I Error Rate} = \frac{\text{Number of normal examinees incorrectly flagged as aberrant}}{\text{Total number of normal examinees}}$$

Higher statistical power indicates stronger ability to identify examinees with preknowledge, while Type I error rates closer to the significance level indicate better control of Type I errors.

5.3 Data Generation Process

Response times for examinees on normal items (non-preknown items) were simulated using van der Linden's lognormal model. For examinees on preknown items, we referenced the empirical data analysis and considered three levels of speed differences. This approach simulates different levels of speed differences between normal and preknown items. After generating datasets, we calculated SLR statistics and corresponding p-values, Bayes factors BF, and posterior probabilities PP. SLR judgments were based on p-values, BF critical values followed Kass and Raftery's (1995) recommended standard (BF $>$ 1 indicates item preknowledge), and PP critical value was set at 0.95.

5.4 Research Findings

Both larger SLR and PP values indicate higher likelihood of speed differences between subtests S1 and S2. We illustrate this relationship using a scatter plot with the x-axis representing 1 - p values from SLR statistics and the y-axis representing posterior probabilities PP (Figure 7 [Figure 7: see original paper]). As speed differences increase, both 1 - p and PP values increase correspondingly. Figure 7 depicts the experimental condition with 10% aberrant examinees having preknowledge of 20 items, with speed differences drawn from U(0.20-0.35). Each circle represents one examinee, with green circles for examinees without preknowledge and red circles for those with preknowledge.

Figure 7. Scatter Plot of 1 - p Values vs. Posterior Probability

Note: The two dashed lines correspond to 1 - p = 0.95 and posterior probability PP = 0.95.

At a significance level of 0.05, SLR's 1 - p value > 0.95 and posterior probability > 0.95 indicate that the corresponding method flags the examinee as having item preknowledge. Figure 7 shows good consistency between the two statistics in detecting examinees without item preknowledge. In detecting examinees with preknowledge, SLR performs slightly better than posterior probability, as some preknowledge examinees have 1 - p values > 0.95 but posterior probabilities < 0.95. Consistent with findings by Berger and Sellke (1987), Casella and Berger (1987), and Pratt (1965) regarding the consistency between posterior probability and frequentist p-values in testing one-sided alternatives, the consistency between PP and 1 - p is expected.

Overall, compared to detection based on response score data (see Sinharay and Johnson, 2021, Fig. 2), SLR and PP based on response time data show higher consistency, evidenced by examinee data being more concentrated near the diagonal with a narrower "bandwidth," while the corresponding figure for response score data shows a larger "bandwidth." This suggests that response time data is more advantageous than score data for detecting item preknowledge, primarily for two reasons: (1) As continuous data, response time contains richer information than discrete response score data, facilitating analysis of examinee behavior (Cheng & Shao, 2022; Sinharay, 2017); (2) When examinees have item preknowledge, their direct performance is typically shorter response times and faster response speeds on those items.

Table 5 presents the statistical power and Type I error rates of BF, SLR, and PP methods under different conditions for detecting examinees with item preknowledge. When the speed difference level is low "U(0.20-0.35)" and the proportion of preknown items is also low (0.1), all methods have relatively low power, not exceeding 0.6. However, when the speed difference level reaches high "U(0.5-1)," all methods achieve high power, with maximum exceeding 0.99. Referencing the empirical data results, the medium speed difference level is closest to the empirical data. Under this condition, BF and SLR achieve power above 80% even under the mildest item leakage condition (leakage proportion 0.1). As the proportion of preknown items increases, the power of all three methods increases, because relatively more compromised items facilitate detection of affected examinees. Since this study was conducted under known item parameters, the proportion of affected examinees has minimal impact on power. The results show that all three methods control Type I error rates well across conditions. Overall, BF performs best in statistical power, achieving the highest accuracy across conditions, followed closely by SLR with slightly lower power, while PP has the lowest power among the three methods, directly related to its lower sensitivity to speed differences.

Table 5. Performance of BF, SLR, and PP in Detecting Item Preknowledge Under Different Conditions

Speed Difference Distribution	Proportion of Preknown Items	Proportion of Affected Examinees	U(0.20-0.35)	U(0.35-0.50)	U(0.50-1)
Statistical Power					
Type I Error Rate					

Note: BF, SLR, and PP represent Bayes factor, signed likelihood ratio, and posterior probability statistics based on response time data, respectively.

Figure 8 [Figure 8: see original paper] compares the statistical power of the three methods across three proportions of preknown items (0.1, 0.2, 0.3) when the speed difference level is U(0.20-0.35). It is clear that power increases with the proportion of preknown items. When the proportion increases from 0.1 to 0.2, the increase in power is about 15%, with the maximum increase exceeding 20%. When the proportion increases from 0.2 to 0.3, the increase does not exceed 10%, with the minimum increase less than 3%. This suggests that to achieve accurate detection of examinees with item preknowledge, the proportion of preknown items needs to reach 20% or higher.

Figure 8. Comparison of Three Methods' Performance in Detecting Item Preknowledge Based on Response Time

Note: X-axis labels show two numbers representing the proportion of preknown items and the proportion of affected examinees (e.g., 0.10-0.05 means 0.10 preknown item proportion and 0.05 affected examinee proportion). Y-axis represents statistical power for detecting item preknowledge.

Figure 9 [Figure 9: see original paper] compares statistical power across different speed difference levels ("low U(0.20-0.35), medium U(0.35-0.50), high U(0.50-1)") under two preknown item proportions (0.2, 0.3). With a preknown item proportion of 0.2, BF's mean power across the three speed difference levels is 0.710, 0.962, and 0.965; SLR's is 0.657, 0.948, and 0.944; and PP's is 0.596, 0.814, and 0.822. It is evident that power increases substantially when speed difference levels increase from U(0.20-0.35) to U(0.35-0.50), while remaining relatively stable when increasing from U(0.35-0.50) to U(0.50-1). This indicates that all methods can accurately detect examinees with item preknowledge when speed differences reach medium levels.

Figure 9. Comparison of Three Methods' Performance in Detecting Item Preknowledge Based on Response Time

Note: X-axis labels show the speed difference interval and proportion of affected examinees (e.g., 0.20-0.35 (0.05) means speed differences drawn from U(0.20-0.35) and 0.05 affected examinee proportion). Y-axis represents statistical power for detecting item preknowledge.

6. Discussion and Future Research Directions

This paper extends Bayes factor and posterior probability statistics to construct two statistics for detecting item preknowledge using response time data and compares them with the signed likelihood ratio test statistic to detect differences in examinees' response speeds on compromised versus normal items. Empirical data analysis from a professional qualification test shows that when examinees have item preknowledge, their response speeds differ between compromised and normal items. The three statistics (SLR, BF, and PP) differ in detection sensitivity, with BF being most sensitive to speed differences and PP being least sensitive. Based on the empirical analysis results, we designed targeted simulation experiments to examine the impact of three factors—different degrees of speed differences caused by item preknowledge, the prevalence of such preknowledge among examinees, and the proportion of preknown items in the total test—on their performance under various test conditions.

Simulation results further indicate that: (1) BF is most sensitive to speed differences, with power slightly higher than SLR, while PP is least sensitive; (2) All three methods control Type I error rates well; (3) When speed differences reach medium levels "U(0.35, 0.50)," all three methods achieve high detection accuracy; (4) To accurately detect examinees with item preknowledge, the proportion of preknown items needs to reach 20% or higher; (5) Since this study was conducted under known item parameters, the prevalence of preknowledge in the examinee population has minimal impact on detection results.

Although this study shows good performance in detecting item preknowledge, several limitations remain and warrant future research: (1) The performance of BF and PP based on time data needs further exploration with more empirical data and simulation conditions, such as investigating their performance in computer adaptive testing scenarios. (2) This study only used examinees' time data; incorporating both score and time data simultaneously into BF and PP deserves further exploration. (3) Multidimensional ability tests are common in practical applications, so extending current BF and PP to multidimensional abilities is needed. (4) This study used theoretical critical values for SLR and PP and empirical critical values for BF; whether these critical values are appropriate in practical research and exploring more suitable critical values require further investigation. (5) This study was conducted under known item parameters; future research needs to explore unknown item parameter conditions and how to handle the "masking effect" (Fung, 1993; Yuan & Zhong, 2008) caused by unknown item parameters. (6) This study only examined three speed difference levels; since PP is less sensitive to speed differences, its performance was relatively poor, but how it performs under larger difference ranges and degrees needs further study. (7) In practical applications, preknown item information may not be fully known (e.g., among 10 flagged preknown items, only 8 may actually be preknown by examinees) (Belov, 2016); incorporating uncertainty about preknown item information is a direction for future research. (8) Regarding the integrals in BF and PP formulas, this study used Riemann sum approximation;

future research should examine other numerical calculation methods to improve computational accuracy and efficiency.

References

Allen, J., & Ghattas, A. (2016). Estimating the probability of traditional copying, conditional on answer-copying statistics. *Applied Psychological Measurement, 40*(4), 258–273.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association, 82*(397), 112-122.

Belov, D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing, 2*(3), 37-58.

Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement, 40*(2), 83–97.

Casella, G., & Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association, 82*(397), 106–111.

Cheng, Y., & Shao, C. (2022). Application of change point analysis of response time data to detect test speededness. *Educational and Psychological Measurement, 82*(5), 1031–1062.

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of detecting cheating on tests*. Routledge.

Fox, J.-P., Koops, J., Feskens, R., & Beinhauer, L. (2007). Bayesian covariance structure modelling for measurement invariance testing. *Behaviormetrika, 47*(2), 385-410.

Fox, J.-P., & Marianti, S. (2016). Joint Modeling of Ability and Differential Speed Using Responses and Response Times. *Multivariate Behavioral Research, 51*(4), 540-553.

Fung, W. K. (1993). Unmasking outliers and leverage points: A confirmation. *Journal of the American Statistical Association, 88*(422), 515–519.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Chapman & Hall.

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored internet tests: The Z-test and the likelihood ratio test. *International Journal of Selection and Assessment, 18*(4), 351–364.

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.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795.

Klein, E. R. H., Fox, J. P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika, 74*(1), 21–48.

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

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

Liu, Y., & Liu, H. Y. (2018). A comparison study for the four parameter logistic model and traditional logistic models. *Psychological Exploration, 38*(3), 228−235.

Liu, Y., & Liu, H. Y. (2021). Mixture Model Method: A new method to handle aberrant responses in psychological and educational testing. *Advances in Psychological Science, 29*(9), 1696-1710.

Liu, Y., & Liu, H. Y. (2022). A comparison of standard residual methods and a mixture hierarchical model for detecting non-effortful responses. *Acta Psychologica Sinica, 54*(4), 411-425.

Marianti, S., Fox, J.-P., Maranna, A., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics, 39*(6), 426–451.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*(3), 200–219.

Pan, Y. Q., & Wollack, J. A. (2021). An Unsupervised-Learning-Based Approach to Compromised Items Detection. *Journal of Education Measurement, 58*(3), 413-433.

Pratt, J. W. (1965). Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society: Series B, 27*(2), 169–192.

Robert, C. P. (2007). *The Bayesian choice* (2nd ed.). Springer.

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.

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.

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

Shu, Z., Henson, R., & Luecht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika, 78*(3), 481-497.

Sinharay, S. (2016). Asymptotic corrections of standardized extended caution indices. *Applied Psychological Measurement, 40*(6), 418–433.

Sinharay, S. (2017a). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics, 42*(1), 46-68.

Sinharay, S. (2017b). Which statistic should be used to detect item preknowledge when the set of compromised items is known? *Applied Psychological Measurement, 41*(6), 403-421.

Sinharay, S. (2020). Detection of item preknowledge using response times. *Applied Psychological Measurement, 44*(5), 376-392.

Sinharay, S., & Johnson, M. S. (2020). The use of item scores and response times to detect examinees who may have benefited from item preknowledge. *British Journal of Mathematical and Statistical Psychology, 73*(3), 397-419.

Sinharay, S., & Johnson, M. S. (2021). The use of the posterior probability in score differencing. *Journal of Educational and Behavioral Statistics, 46*(4), 403-429.

Skorupski, W. P., & Wainer, H. (2017). The case for Bayesian methods when investigating test fraud. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 214–231). Routledge.

Stern, H. S. (2005). Model inference or model selection: Discussion of Klugkist, Laudy, and Hoijtink (2005). *Psychological Methods, 10*(4), 494–499.

Ulitzsch E., Von D. M., & Pohl, S. (2020). Using Response Times for Joint Modeling of Response and Omission Behavior. *Multivariate Behavior Research, 55*(3), 425-453.

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

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287–308.

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. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46*(3), 247–272.

van der Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika, 75*(1), 120–139.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*(3), 456-477.

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics, 43*(4), 469–501.

Wang, X., Liu, Y., & Hambleton, R. K. (2017). Detecting item preknowledge using a predictive checking method. *Applied Psychological Measurement, 41*(4), 243-263.

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.

Wollack, J. A., & Schoenig, R. W. (2018). Cheating. In B. B. Frey (Ed.), *The Sage encyclopedia of educational research, measurement, and evaluation* (pp. 260–265). Thousand Oaks, CA: Sage.

Yu, X., & Cheng, Y. (2022). A comprehensive review and comparison of CUSUM and change-point analysis methods to detect test speededness. *Multivariate Behavioral Research, 57*(1), 112-133.

Zhong, X.Y., Yu, X. F., Miao, Y., Qin, C.Y., Peng, Y. F., & Tong, H. (2022). Exploration of change point analysis in detecting speededness based on response time data with known/unknown item parameters. *Acta Psychologica Sinica, 54*(10), 1277-1292.

Zhu, H. Y., Jiao, H., Gao, W., & Meng, X. B. (2023). Bayesian change-point analysis approach to detecting aberrant test-taking behavior using response times. *Journal of Educational and Behavioral Statistics, 48*(4), 490–520.

Yuan, K. H., & Zhong, X. (2008). Outliers, leverage observations, and influential cases in factor analysis: Using robust procedures to minimize their effect. *Sociological Methodology, 38*(1), 329–368.

Speed Difference Detection Based on Response Time Data

Abstract: Response time data contains information on examinees' response speed and behavior, and is gaining increasing attention in educational and psychological measurement. Abnormal speed during testing indicates that an examinee may have engaged in aberrant test-taking behavior, such as having prior knowledge of some test items. Based on response time data, this paper constructs two statistics (referred to as BF and PP) that can test for speed differences, uses the newly constructed statistics and the signed likelihood ratio test statistic (SLR) on empirical data, and demonstrates the process of using the new methods. The results show that when examinees have preknowledge of some items, there is a difference in their response speed on leaked and normal items. The three statistics (SLR, BF, and PP) differ in detection sensitivity, with BF being the most sensitive to speed differences and PP being the least sensitive. Based on the empirical data analysis results, targeted simulation experiments were further designed to examine the impact of different degrees of speed differences brought about by item preknowledge, the prevalence of such preknowledge among examinees, and the proportion of known items in the total test under various test conditions, and a comprehensive comparison of their performance was conducted. The results show that (1) three methods can control the error rate well; (2) When the speed difference reaches medium "U(0.35, 0.50)," three methods can achieve a high detection accuracy rate; (3) To achieve a more accurate test of examinees with item preknowledge, examinees need to know 20% or more of items in advance; (4) Since this study is carried out under the condition of known item parameters, the prevalence of preknowledge in the population is predicted to have less effect on the test results. The newly constructed statistics have good performance in testing the difference in response speed in the examination process.

Keywords: response time, speed, posterior probability, difference detection, Bayesian factor

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