The Use of Item Scores and Response Times to Detect Examinees Who May Have Benefited from Item Preknowledge

Sandip Sinharay and Matthew S. Johnson,
Educational Testing Service

An Updated Version of this document appeared on 08/16/2019 in the British Journal of Mathematical and Statistical Psychology. The website for the article is https://onlinelibrary.wiley.com/doi/abs/10.1111/bmsp.12187

The citation for the article is: Sinharay, S. & Johnson, M. S. (2019). The use of item scores and response times To detect the examinees who may have benefitted from item preknowledge. British Journal of Mathematical and Statistical Psychology. Advance Online Publication. DOI: bmsp.12187

Note: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D170026. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education or of Educational Testing Service.

The Use of Item Scores and Response Times To Detect the Examinees Who May Have Benefitted from Item Preknowledge

Sandip Sinharay and Matthew S. Johnson, Educational Testing Service

July 11, 2019

Note: Any opinions expressed in this publication are those of the authors and not necessarily of Educational Testing Service or Institute of Education Sciences.

The Use of Item Scores and Response Times To Detect the Examinees Who May Have Benefitted from Item Preknowledge

Abstract

According to Wollack and Schoenig (2018), benefitting from item preknowledge is one of the three broad types of test fraud that occur in educational assessments. We use tools from constrained statistical inference to suggest a new statistic that is based on item scores and response times and can be used to detect the examinees who may have benefitted from item preknowledge for the case when the set of compromised items is known. The asymptotic distribution of the new statistic under no preknowledge is proved to be a simple mixture of two χ^2 distributions. We perform a detailed simulation study to show that the Type I error rate of the new statistic is very close to the nominal level and that the power of the new statistic is satisfactory in comparison to that of the existing statistics for detecting item preknowledge based on both item scores and response times. We also include a real data example to demonstrate the usefulness of the suggested statistic.

Key words: chi-bar-square distribution, likelihood ratio statistic, Wald statistic.

Standard 6.6 of the Standards for educational and psychological testing (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014) includes the recommendation that testing programs with high-stakes consequences should have defined procedures for detecting potential testing irregularities. One type of testing irregularity is the sharing of assessment questions and/or their answers by a source (such as a website) followed by several examinees memorizing the questions and/or answers. The examinees who are involved in such a phenomenon are referred to have benefitted from *item preknowledge* and the shared items are referred to as *compromised* items. Wollack and Schoenig (2018) listed item preknowledge as one of three broad types of test fraud that occur in educational assessments (the other two types being test tampering and answer-copying). In this paper, we consider the detection of item preknowledge for the case when the investigator knows which items are compromised. Cizek and Wollack (2017, p. 14) and Eckerly, Smith, and Lee (2018) provided examples of real data sets for which several items were known to have been compromised. In cases where the set of compromised items is unknown, it is possible to apply a method to detect compromised items (e.g., Veerkamp & Glas, 2000) before applying the methods discussed in this paper.

While most research on detecting item preknowledge is based only on item scores (e.g., Drasgow, Levine, & Zickar, 1996; McLeod, Lewis, & Thissen, 2003; Sinharay, 2017a; Wang, Liu, & Hambleton, 2017), the increasing popularity of computerized assessments has allowed the recording of response times, and, subsequently, detection of item preknowledge using both item scores and response times. Researchers such as Fox and Marianti (2017), Lee and Wollack (2017), van der Linden and Guo (2008), and Wang, Xu, Shang, and Kuncel (2018) suggested a variety of methods that can be used to detect item preknowledge using both item scores and response times. However, all of these existing approaches are designed to detect response patterns that are in general aberrant (aberrant response patterns are those that do not fit the joint model for item scores and response times) and are not specifically designed to detect item preknowledge. It is expected that a statistic that is based on both item scores and response times and specifically targets item preknowledge will be more

powerful than the existing approaches. Therefore, the goal of this paper is to suggest a statistic that is (a) based on both item scores and response times and (b) specifically designed to detect item preknowledge. The statistic works with the hierarchical/joint model of van der Linden (2007) for item scores and response times and is predicated on the idea that the performance of those with item preknowledge is likely to differ over the compromised items and non-compromised items.

The next section includes reviews of the hierarchical model of van der Linden (2007) for item scores and response times, the existing approaches for estimation of the parameters of the model, and the existing approaches for detection of item preknowledge using item scores and/or response times. In the Methods section, we describe a new statistic for detection of item preknowledge based on item scores and response times and prove that the asymptotic null distribution of the new statistic is a simple mixture of two χ^2 distributions. A study of the Type I error rate and power of the new statistic is included in the Simulation section. The Real Data section includes an application of the new statistic to an operational data set that involves actual item preknowledge. The last section includes some conclusions and recommendations.

Literature Review: A Model and Some Existing Methods

The Hierarchical Model for Item Scores and Response Times

The hierarchical/joint modeling approach of van der Linden (2007) involves the application of a model for response times in combination with an item response theory (IRT) model for the item scores. Let us consider an assessment that consists of I items. Let t_i and x_i respectively denote the response time and item score of a randomly chosen examinee on item i. Let y_i denote the logarithm of t_i. In the first stage of the hierarchical modeling approach of van der Linden (2007), one assumes that

- the response time follows the lognormal model (LNMRT; van der Linden, 2006), that

is, y_i follows a normal distribution with mean $\beta_i - \tau$ and variance $\frac{1}{\alpha_i^2}$, or,

$$f(y_i|\tau, \alpha_i, \beta_i) = \frac{\alpha_i}{\sqrt{2\pi}} e^{-\frac{1}{2}\alpha_i^2(y_i - \beta_i + \tau)^2}.$$

The parameters τ, β_i and α_i respectively are the examinee's speed parameter, the time-intensity parameter for item i, and the time-discrimination parameter for item i.

- the item scores follow an IRT model; for example, if the two-parameter logistic model (2PLM) is used and a_i and b_i respectively denote the item slope and item difficulty parameter of item i, then

$$P(x_i = 1|\theta, a_i, b_i) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}},$$

where θ is the examinee ability.

Though van der Linden (2007) used the three-parameter normal ogive model for the item scores in his hierarchical approach, the flexibility of the approach allows the use of any other IRT model such as the 2PLM or the three-parameter logistic model. In the second stage of the hierarchical modeling approach of van der Linden (2007), one assumes a suitable prior distribution, typically a bivariate normal distribution with means equal to 0, for the vector of examinee parameters, $(\tau, \theta)'$.

Klein Entink, Fox, and van der Linden (2009) and van der Linden (2007) suggested Bayesian approaches to estimate the item parameters of the hierarchical model involving the LNMRT and an IRT model (such as the 2PLM or 3PLM or their corresponding normal-ogive versions) using the Markov chain Monte Carlo algorithm. The Bayesian approach is implemented in the R package LNIRT (Fox, Klein Entink, & Klotzke, 2017). Glas and van der Linden (2010) suggested an expectation-maximization (EM) algorithm to compute the marginal maximum likelihood estimates (MMLEs) of the item parameters of the hierarchical model of van der Linden (2007). van Rijn and Ali (2017) provided further details (such as expressions of the first and second derivatives of the marginal likelihood function) on the EM algorithm for the hierarchical model involving the LNMRT and the 2PLM. Molenaar, Tuerlinckx, and van der Maas (2015) provided Mplus codes for computing the MMLEs of the item parameters of the hierarchical model.

The joint distribution of the item scores and the response times of a person can be expressed as the product of the distribution of the item scores and that of the response times (e.g., van der Linden, 2007). Therefore, to compute the joint maximum likelihood estimates (MLEs) of τ and θ for a person after the estimation of the item parameters, one can compute the MLEs separately—the one for τ only using the response times and the one for θ only using the item scores. Standard textbooks such as Baker and Kim (2004) describe approaches for the estimation of θ given item parameters. The computation of the estimates of the person speed parameter τ given α_i^2's and β_i's for the LNMRT is discussed in van der Linden (2006).

Detecting Item Preknowledge Using Item Scores and/or Response Times

Let $\mathcal{C}$ and $\overline{\mathcal{C}}$ respectively denote the set of compromised items and non-compromised items that were administered to the abovementioned randomly chosen examinee. Let $\boldsymbol{x}$, $\boldsymbol{x}_{\mathcal{C}}$, and $\boldsymbol{x}_{\overline{\mathcal{C}}}$ respectively denote the collection of the scores of the examinee on all items, on the items in $\mathcal{C}$ and on the items in $\overline{\mathcal{C}}$, respectively. Thus, for example, $\boldsymbol{x}_{\mathcal{C}} = \{x_i, i \in \mathcal{C}\}$. Similarly, let $\boldsymbol{y}$, $\boldsymbol{y}_{\mathcal{C}}$, and $\boldsymbol{y}_{\overline{\mathcal{C}}}$ respectively denote the collection of logarithm of response times of the examinee on all items, the items in $\mathcal{C}$ and the items in $\overline{\mathcal{C}}$. The existing approaches for detecting item preknowledge using item scores and/or response times are briefly described below.

The Signed Likelihood Ratio Test Based on the Item Scores

For the examinee, let us denote the true ability based on the whole test, $\mathcal{C}$, and $\overline{\mathcal{C}}$ as θ, $\theta_{\mathcal{C}}$, and $\theta_{\overline{\mathcal{C}}}$, respectively.[1] Let us denote the MLEs of these parameters as $\hat{\theta}$, $\hat{\theta}_{\mathcal{C}}$, and $\hat{\theta}_{\overline{\mathcal{C}}}$, respectively. Note that $\hat{\theta}$ is computed from all the items on the test while $\hat{\theta}_{\mathcal{C}}$ and $\hat{\theta}_{\overline{\mathcal{C}}}$ are computed from the subsets $\mathcal{C}$ and $\overline{\mathcal{C}}$, respectively. The likelihood ratio test (LRT) statistic

[1] The true or estimated ability based on a part of the test is rarely of interest in operational score reporting. However, the introduction of $\theta_{\mathcal{C}}$, and $\theta_{\overline{\mathcal{C}}}$ here facilitates the derivation of the new statistic, as will be clear later.

for testing the null hypothesis $H_0 : \theta_{\mathcal{C}} = \theta_{\overline{\mathcal{C}}}$ versus the alternative hypothesis $H_1' : \theta_{\mathcal{C}} \neq \theta_{\overline{\mathcal{C}}}$ is given by

$$\Lambda_S = 2[\ell(\boldsymbol{x}_{\mathcal{C}}|\hat{\theta}_{\mathcal{C}}) + \ell(\boldsymbol{x}_{\overline{\mathcal{C}}}|\hat{\theta}_{\overline{\mathcal{C}}}) - \ell(\boldsymbol{x}|\hat{\theta})], \tag{1}$$

where

$$\ell(\boldsymbol{x}_{\mathcal{C}}|\hat{\theta}_{\mathcal{C}}) = \text{log-likelihood of the scores on } \mathcal{C} \text{ at } \hat{\theta}_{\mathcal{C}},$$

$$\ell(\boldsymbol{x}_{\overline{\mathcal{C}}}|\hat{\theta}_{\overline{\mathcal{C}}}) = \text{log-likelihood of the scores on } \overline{\mathcal{C}} \text{ at } \hat{\theta}_{\overline{\mathcal{C}}},$$

$$\text{and } \ell(\boldsymbol{x}|\hat{\theta}) = \text{log-likelihood of the scores on all the items at } \hat{\theta}.$$

The local independence assumption underlying the standard IRT models leads to equalities such as

$$\ell(\boldsymbol{x}_{\mathcal{C}}|\hat{\theta}_{\mathcal{C}}) = \sum_{i \in \mathcal{C}} \log P_i(x_i|\hat{\theta}_{\mathcal{C}}),$$

where $P_i(x_i|\hat{\theta}_{\mathcal{C}})$ is the probability of a score x_i on item i at $\theta = \hat{\theta}_{\mathcal{C}}$. In the context of the hierarchical modeling approach of van der Linden (2007), terms such as $P_i(x_i|\hat{\theta}_{\mathcal{C}})$ depend on the IRT model used under. For example, if the 2PLM is used as the IRT model, then

$$P_i(x_i|\hat{\theta}_{\mathcal{C}}) = \left(\frac{e^{a_i(\hat{\theta}_{\mathcal{C}} - b_i)}}{1 + e^{a_i(\hat{\theta}_{\mathcal{C}} - b_i)}}\right)^{x_i} \left(\frac{1}{1 + e^{a_i(\hat{\theta}_{\mathcal{C}} - b_i)}}\right)^{1 - x_i}.$$

Then one may express the LRT statistic given in Equation 1 as

$$\Lambda_S = 2\left\{\sum_{i \in \mathcal{C}} \log P_i(x_i|\hat{\theta}_{\mathcal{C}}) + \sum_{i \in \overline{\mathcal{C}}} \log P_i(x_i|\hat{\theta}_{\overline{\mathcal{C}}}) - \sum_{i=1}^{I} \log P_i(x_i|\hat{\theta})\right\}.$$

Sinharay (2017a) suggested that to detect item preknowledge, one can test $H_0 : \theta_{\mathcal{C}} = \theta_{\overline{\mathcal{C}}}$ versus $H_1 : \theta_{\mathcal{C}} \geq \theta_{\overline{\mathcal{C}}}$, and the hypothesis can be tested using the signed likelihood ratio test statistic given by

$$L_S = \begin{cases} \sqrt{\Lambda_S} \text{ if } \hat{\theta}_{\mathcal{C}} \geq \hat{\theta}_{\overline{\mathcal{C}}}, \\ -\sqrt{\Lambda_S} \text{ if } \hat{\theta}_{\mathcal{C}} < \hat{\theta}_{\overline{\mathcal{C}}}. \end{cases}$$

A large positive value of L_S leads to the rejection of the null hypothesis of no item preknowledge. The statistic L_S follows the standard normal distribution for large $\mathcal{C}$ and

$\overline{\mathcal{C}}$ under the null hypothesis of no item preknowledge (e.g., Sinharay, 2017a; Cox, 2006, p. 104). Sinharay (2017a) and Sinharay (2017b) found the Type I error rate and power of L_S to be quite satisfactory in comparison to the existing statistics for detecting item preknowledge based on item scores.

The Signed Likelihood Ratio Test Based on Response Times

Let $\tau_{\mathcal{C}}$ and $\tau_{\overline{\mathcal{C}}}$ respectively denote the examinee's true speed parameters on the compromised and non-compromised items, respectively, and let $\hat{\tau}_{\mathcal{C}}$ and $\hat{\tau}_{\overline{\mathcal{C}}}$ denote their MLEs. Let $\hat{\tau}$ denote the MLE of the examinee's true speed parameter based on all the I items on the test. The LRT statistic for testing $H_0 : \tau_{\mathcal{C}} = \tau_{\overline{\mathcal{C}}}$ versus $H_1' : \tau_{\mathcal{C}} \neq \tau_{\overline{\mathcal{C}}}$ is given by

$$\Lambda_T = 2[\ell(\boldsymbol{y}_{\mathcal{C}}|\hat{\tau}_{\mathcal{C}}) + \ell(\boldsymbol{y}_{\overline{\mathcal{C}}}|\hat{\tau}_{\overline{\mathcal{C}}}) - \ell(\boldsymbol{y}|\hat{\tau})], \tag{2}$$

where, for example,

$$\ell(\boldsymbol{y}_{\mathcal{C}}|\hat{\tau}_{\mathcal{C}}) = \text{log-likelihood of the log-response times of the items in } \mathcal{C}, \text{ computed at } \hat{\tau}_{\mathcal{C}}.$$

Sinharay (2019) showed that when one uses the LNMRT (van der Linden, 2006) for the response times under the hierarchical modeling framework of van der Linden (2007), $\ell(\boldsymbol{y}_{\mathcal{C}}|\hat{\tau}_{\mathcal{C}})$ can be expressed as

$$\ell(\boldsymbol{y}_{\mathcal{C}}|\hat{\tau}_{\mathcal{C}}) = \sum_{i\in\mathcal{C}} \left[-\frac{1}{2}\log[2\pi] + \log(\alpha_i)\right] + \hat{\tau}_{\mathcal{C}}^2 \sum_{i\in\mathcal{C}} \frac{\alpha_i^2}{2} - \sum_{i\in\mathcal{C}} \frac{\alpha_i^2}{2}(y_i - \beta_i)^2, \tag{3}$$

and that

$$\Lambda_T \;=\; \hat{\tau}_{\mathcal{C}}^2 \sum_{i\in\mathcal{C}} \alpha_i^2 + \hat{\tau}_{\overline{\mathcal{C}}}^2 \sum_{i\in\overline{\mathcal{C}}} \alpha_i^2 - \hat{\tau}^2 \sum_{i=1}^{I} \alpha_i^2. \tag{4}$$

Consequently, to detect item preknowledge based on response times, one can test $H_0 : \tau_{\mathcal{C}} = \tau_{\overline{\mathcal{C}}}$ versus $H_1 : \tau_{\mathcal{C}} > \tau_{\overline{\mathcal{C}}}$ using the signed likelihood ratio statistic given by

$$L_T = \begin{cases} \sqrt{\Lambda_T} \text{ if } \hat{\tau}_{\mathcal{C}} \geq \hat{\tau}_{\overline{\mathcal{C}}}, \\ -\sqrt{\Lambda_T} \text{ if } \hat{\tau}_{\mathcal{C}} < \hat{\tau}_{\overline{\mathcal{C}}} \end{cases} \tag{5}$$

(Sinharay, 2019).

It can be shown that for this hypothesis-testing problem, the L_T statistic is identical to the Wald test statistic given by

$$\frac{\hat{\tau}_{\mathcal{C}} - \hat{\tau}_{\overline{\mathcal{C}}}}{\sqrt{\text{Var}(\hat{\tau}_{\mathcal{C}}) + \text{Var}(\hat{\tau}_{\overline{\mathcal{C}}})}} = \frac{\hat{\tau}_{\mathcal{C}} - \hat{\tau}_{\overline{\mathcal{C}}}}{\sqrt{[\sum_{i \in \mathcal{C}} \alpha_i^2]^{-1} + [\sum_{i \in \overline{\mathcal{C}}} \alpha_i^2]^{-1}}}.$$

Sinharay (2019) also showed that under the LNMRT, L_T follows the standard normal distribution under the null hypothesis of no item preknowledge irrespective of the sizes of $\mathcal{C}$ and $\overline{\mathcal{C}}$.

Using A Bayesian Person-fit Approach to Detect Item Preknowledge

Marianti, Fox, Avetisyan, Veldkamp, and Tijmstra (2014) suggested a person-fit statistic based on response times that is given by

$$l^t = \sum_i \alpha_i^2 (y_i - \beta_i + \tau)^2, \tag{6}$$

and described a Bayesian approach to estimate the posterior probability of an aberrant response-time pattern using l^t. To assess person fit using item scores, Fox and Marianti (2017) suggested a Bayesian approach to estimate the posterior probability of an aberrant item-score pattern using the l_z statistic (Drasgow, Levine, & Williams, 1985) that is given by

$$l_z = \frac{\ell(\boldsymbol{x}|\theta) - \text{E}(\ell(\boldsymbol{x}|\theta))}{\sqrt{\text{Var}(\ell(\boldsymbol{x}|\theta))}}.$$

Fox and Marianti (2017) also suggested a Bayesian approach using both l^t and l_z to estimate the posterior probability of both an aberrant item-score pattern and aberrant response-time pattern. The posterior probability is expected to be large (for example, larger than 0.99) for aberrant response patterns. While the Bayesian approach of Fox and Marianti (2017) is designed to detect a variety of aberrant responses, the approach can be used to detect item preknowledge.

Using Standardized Residuals to Detect Item Preknowledge

van der Linden and Guo (2008) suggested a Bayesian approach for detecting aberrant response times in the context of the hierarchical model of van der Linden

(2007). They showed that the posterior distribution of the predicted value of the log-response time on item i conditional on $\boldsymbol{y}_{-i} = (y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_I)$ and also on $\boldsymbol{x}_{-i} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_i)$ is approximately normal. Then, they defined the standardized residual, e_i, as

$$e_i = \frac{y_i - E(y_i|\boldsymbol{y}_{-i}, \boldsymbol{x}_{-i})}{\sqrt{\text{Var}(y_i|\boldsymbol{y}_{-i}, \boldsymbol{x}_{-i})}}. \tag{7}$$

The response time for the examinee on item i is concluded as aberrant at 1% level if the absolute value of e_i is larger than the 99th percentile of the standard normal distribution. While this approach is designed to detect a variety of aberrant responses, the approach can be used to flag for possible item preknowledge the examinees with several statistically significant and negative e_i's, as was performed in Boughton, Smith, and Ren (2017, p. 181) and Qian, Staniewska, Reckase, and Woo (2016). We follow the strategy of van der Linden and Guo (2008, p. 382) of flagging for possible item preknowledge the examinees with one or more statistically significant e_i's that are negative and associated with a correct answer.

Using Mixture Models to Detect Item Preknowledge

Lee and Wollack (2017) and Wang et al. (2018) suggested using mixture hierarchical IRT models, which are fitted using Bayesian estimation approaches, to detect aberrant item scores and response times. These models include an aberrance indicator Δ_{ij} for each examinee-item combination. The estimated posterior probability of Δ_{ij} being equal to 1 can be used to determine whether the response time and item score for an item-examinee combination are aberrant. An examinee with too many aberrant item-examinee combinations may be identified for possible item preknowledge. These approaches do not require the assumption of known compromised items.

The use of the approaches of Lee and Wollack (2017) or Wang et al. (2018) are most appropriate for the cases when one does not know the set of compromised items; when the investigator knows the set of compromised items, these approaches are not capable to take that information into account and are expected to lead to low power. In some limited simulations, when the set of compromised items is known, these approaches were

found to have much smaller power to detect examinees with preknowledge than the new statistic suggested later. In addition, the approaches of Lee and Wollack (2017) and Wang et al. (2018) are time-consuming; for example, Lee and Wollack (2017) stated that the estimation of their mixture model for a real data set with less than 2,000 examinees (the data set is the same as the one analyzed later in the current paper) took more than three days. Therefore, the approaches of Lee and Wollack (2017) or Wang et al. (2018) are not considered henceforth in this paper.

Motivation of This Paper

The above review shows that all of the existing approaches that can be used to detect item preknowledge based on item scores and response times are designed to detect response patterns that are in general aberrant and are not specifically designed to detect item preknowledge. Given that statistics based only on item scores and specifically designed to detect item preknowledge have been found more powerful than statistics based only on item scores and designed to detect aberrant response patterns in general (e.g., Sinharay, 2017a, p. 59), it is expected that a statistic that is based on both item scores and response times and specifically targets item preknowledge will be more powerful than the existing approaches. The major goal of this paper is to suggest one such statistic. As is demonstrated below, it is possible to construct a new statistic that combines information from the L_S and L_T statistics to detect item preknowledge based on both item scores and response times. Ideas from constrained statistical inference (e.g., Silvapulle & Sen, 2001) are used in the construction of the new statistic and the derivation of its asymptotic distribution under the null hypothesis. Given the satisfactory performances of both L_S and L_T (e.g., Sinharay, 2017a, 2019), the new statistic is expected to have satisfactory Type I error rate and power.

Method: A New Statistic Based on Item Scores and Response Times

The examinees who benefited from item preknowledge are likely to answer the compromised items faster than the non-compromised items, as was found for real data

sets by Kasli and Zopluoglu (2018) and Smith and Davis-Becker (2011). They are also likely to perform better on the compromised items in comparison to the non-compromised items, as was found for real data sets by researchers such as Sinharay (2017a) and Smith and Davis-Becker (2011). Consequently, using notation introduced above, one can detect item preknowledge by testing the null hypothesis $H_0 : \tau_{\mathcal{C}} = \tau_{\overline{\mathcal{C}}}$ and $\theta_{\mathcal{C}} = \theta_{\overline{\mathcal{C}}}$ versus the alternative hypothesis $H_1 : \tau_{\mathcal{C}} > \tau_{\overline{\mathcal{C}}}$ or $\theta_{\mathcal{C}} > \theta_{\overline{\mathcal{C}}}$. The rejection of the null hypothesis may indicate possible item preknowledge. The LRT statistic for testing the null hypothesis $H_0 : \tau_{\mathcal{C}} = \tau_{\overline{\mathcal{C}}}$ and $\theta_{\mathcal{C}} = \theta_{\overline{\mathcal{C}}}$ versus the alternative hypothesis $H_1' : \tau_{\mathcal{C}} \neq \tau_{\overline{\mathcal{C}}}$ or $\theta_{\mathcal{C}} \neq \theta_{\overline{\mathcal{C}}}$ can be obtained, in a manner similar to Equation 1, as

$$\Lambda_{ST} = 2 \max_{\theta_{\mathcal{C}},\tau_{\mathcal{C}},\theta_{\overline{\mathcal{C}}},\tau_{\overline{\mathcal{C}}}} [\ell(\boldsymbol{x}_{\mathcal{C}}, \boldsymbol{y}_{\mathcal{C}}|\theta_{\mathcal{C}}, \tau_{\mathcal{C}}) + \ell(\boldsymbol{x}_{\overline{\mathcal{C}}}, \boldsymbol{y}_{\overline{\mathcal{C}}}|\theta_{\overline{\mathcal{C}}}, \tau_{\overline{\mathcal{C}}})] - 2 \max_{\theta,\tau} \ell(\boldsymbol{x}, \boldsymbol{y}|\theta, \tau), \tag{8}$$

where, for example, $\ell(\boldsymbol{x}_{\mathcal{C}}, \boldsymbol{y}_{\mathcal{C}}|\theta_{\mathcal{C}}, \tau_{\mathcal{C}})$ denotes the joint log-likelihood of $\theta_{\mathcal{C}}$ and $\tau_{\mathcal{C}}$ for the examinee.

Because the joint likelihood of the item scores and response times of an examinee is equal to the product of the likelihood of the item scores and the likelihood of the response times under the abovementioned hierarchical model (e.g., van der Linden, 2007), one can express the joint log-likelihood of the ability parameter and the speed parameter for the compromised items, non-compromised items, and all items as

$$\ell(\boldsymbol{x}_{\mathcal{C}}, \boldsymbol{y}_{\mathcal{C}}|\theta_{\mathcal{C}}, \tau_{\mathcal{C}}) = \ell(\boldsymbol{x}_{\mathcal{C}}|\theta_{\mathcal{C}}) + \ell(\boldsymbol{y}_{\mathcal{C}}|\tau_{\mathcal{C}}),$$

$$\ell(\boldsymbol{x}_{\overline{\mathcal{C}}}, \boldsymbol{y}_{\overline{\mathcal{C}}}|\theta_{\overline{\mathcal{C}}}, \tau_{\overline{\mathcal{C}}}) = \ell(\boldsymbol{x}_{\overline{\mathcal{C}}}|\theta_{\overline{\mathcal{C}}}) + \ell(\boldsymbol{y}_{\overline{\mathcal{C}}}|\tau_{\overline{\mathcal{C}}}),$$

$$\text{and } \ell(\boldsymbol{x}, \boldsymbol{y}|\theta, \tau) = \ell(\boldsymbol{x}|\theta) + \ell(\boldsymbol{y}|\tau)\cdot$$

As discussed earlier, for example, the computation of the joint MLE of $\theta_{\mathcal{C}}$ and $\tau_{\mathcal{C}}$ based on $\boldsymbol{x}_{\mathcal{C}}$ and $\boldsymbol{y}_{\mathcal{C}}$ for an examinee is equivalent to the computation of two separate MLEs—one of $\theta_{\mathcal{C}}$ based on $\boldsymbol{x}_{\mathcal{C}}$ and the other of $\tau_{\mathcal{C}}$ based on $\boldsymbol{y}_{\mathcal{C}}$. As a consequence, Λ_{ST} given in Equation 8 can be expressed as

$$\begin{aligned} \Lambda_{ST} &= 2\left[\ell(\boldsymbol{x}_{\mathcal{C}}|\hat{\theta}_{\mathcal{C}}) + \ell(\boldsymbol{y}_{\mathcal{C}}|\hat{\tau}_{\mathcal{C}}) + \ell(\boldsymbol{x}_{\overline{\mathcal{C}}}|\hat{\theta}_{\overline{\mathcal{C}}}) + \ell(\boldsymbol{y}_{\overline{\mathcal{C}}}|\hat{\tau}_{\overline{\mathcal{C}}}) - \ell(\boldsymbol{x}|\hat{\theta}) - \ell(\boldsymbol{y}|\hat{\tau})\right] \\ &= \Lambda_S + \Lambda_T, \end{aligned} \tag{9}$$

because of Equations 1 and 2. If both $\mathcal{C}$ and $\overline{\mathcal{C}}$ are large and the null hypothesis is true, then both L_S and L_T follow the standard normal distribution (Sinharay, 2017a, 2019) and are independent because of the local independence assumption underlying the model of van der Linden (2007), and hence Λ_{ST} is the sum of the squares of two independent standard normal variables and follows the χ^2 distribution with two degrees of freedom. However, the use of Λ_{ST} and, for example, the 95th percentile of the χ^2_2 distribution as a cutoff to perform a test at significance level of 0.05 for detecting preknowledge is inappropriate because the alternative hypothesis of our interest consists of the union of two one-sided hypotheses (rather than the union of two two-sided hypotheses). If one uses Λ_{ST} along with the percentiles of the χ^2_2 distribution, then an examinee who performs considerably worse or slower on the compromised items would be incorrectly identified as having preknowledge. The use of Λ_{ST} was found, in our simulations, to lead to (a) low power and (b) inadvertent flagging of those who performed worse or slower on the compromised items, and is not considered henceforth. Also, Sinharay, Duong, and Wood (2017) emphasized the need of using one-sided hypothesis testing in detection of test fraud.

Therefore, to test against the alternative hypothesis $H_1 : \tau_{\mathcal{C}} > \tau_{\overline{\mathcal{C}}}$ or $\theta_{\mathcal{C}} > \theta_{\overline{\mathcal{C}}}$, instead of using Λ_{ST}, we resorted to tools from constrained statistical inference (e.g., Silvapulle & Sen, 2001) and recommend the constrained likelihood ratio test statistic

$$\Lambda^{\star}_{ST} = L^2_{S+} + L^2_{T+},$$

where $L_{S+} = \max\{L_S, 0\}$ and $L_{T+} = \max\{L_T, 0\}$. The statistic $\Lambda^{\star}_{ST}$ is similar to Λ_{ST} and is identical to Λ_{ST} when both L_S and L_T are positive, but, unlike Λ_{ST}, protects the examinees who perform slower or worse on the compromised items from being incorrectly identified. To derive the null distribution of $\Lambda^{\star}_{ST}$, it is useful to consider the four possible scenarios described in Table 1, where each scenario corresponds to a possible combination of values of L_S and L_T.

In the first scenario, both L_S and L_T are larger than zero, or, equivalently, the estimates of the examinee speed and ability parameters based on the compromised items are larger than the estimates based on the non-compromised items; in this case, $\Lambda^{\star}_{ST}$ is composed of

Table 1: The four possible scenarios and their relationships to the constrained likelihood ratio statistic. The third column gives the probability of the scenario under the null hypothesis

Scenario	$\Lambda^\star_{ST}$	Probability under H_0
$L_S > 0,\ L_T > 0$	$L_S^2 + L_T^2$	$\Pr\{L_S > 0, L_T > 0\} = 0.25$
$L_S \leq 0,\ L_T > 0$	L_T^2	$\Pr\{L_S \leq 0, L_T > 0\} = 0.25$
$L_S > 0,\ L_T \leq 0$	L_S^2	$\Pr\{L_S > 0, L_T \leq 0\} = 0.25$
$L_S \leq 0,\ L_T \leq 0$	0	$\Pr\{L_S \leq 0, L_T \leq 0\} = 0.25$

both L_S^2 and L_T^2. In the second and third scenarios, only one of L_S and L_T is positive, and therefore, $\Lambda^\star_{ST}$ is based only on the one positive statistic. In the final scenario, neither is positive, which results in $\Lambda^\star_{ST}$ being equal to 0.

The four scenarios described in Table 1 simply define the four quadrants in the real plane. If both $\mathcal{C}$ and $\overline{\mathcal{C}}$ include a large number of items and the null hypothesis is true, L_S and L_T are standard normal (e.g., Sinharay, 2017a, 2019) and uncorrelated (because of the local independence assumption) and hence the probability of each scenario is $\frac{1}{4}$. Then, the cumulative distribution function (CDF) of $\Lambda^\star_{ST}$ for large $\mathcal{C}$ and $\overline{\mathcal{C}}$ and under the null hypothesis can be obtained as

$$
\begin{aligned}
P(\Lambda^\star_{ST} \leq \lambda) &= P(\Lambda^\star_{ST} \leq \lambda | L_S > 0, L_T > 0) P(L_S > 0, L_T > 0) \\
&\quad + P(\Lambda^\star_{ST} \leq \lambda | L_S \leq 0, L_T > 0) P(L_S \leq 0, L_T > 0) \\
&\quad + P(\Lambda^\star_{ST} \leq \lambda | L_S > 0, L_T \leq 0) P(L_S > 0, L_T \leq 0) \\
&\quad + P(\Lambda^\star_{ST} \leq \lambda | L_S \leq 0, L_T \leq 0) P(L_S \leq 0, L_T \leq 0) \\
&= \frac{1}{4}[P(L_S^2 + L_T^2 \leq \lambda | L_S > 0, L_T > 0) + P(L_T^2 \leq \lambda | L_S \leq 0, L_T > 0) \\
&\quad + P(L_S^2 \leq \lambda | L_S > 0, L_T \leq 0) + P(0 \leq \lambda | L_S \leq 0, L_T \leq 0)] \\
&= \frac{1}{4}[P(L_S^2 + L_T^2 \leq \lambda | L_S > 0, L_T > 0) + P(L_T^2 \leq \lambda | L_T > 0) \\
&\quad + P(L_S^2 \leq \lambda | L_S > 0) + P(0 \leq \lambda | L_S \leq 0, L_T \leq 0)] \cdot \qquad (10)
\end{aligned}
$$

The last equality holds because L_S is independent with L_T (because of the local independence assumption made in the hierarchical modeling approach of van der Linden, 2007). The CDF provided in Equation 10 corresponds to a mixture distribution, with a weight of 0.25 on the four components of the mixture, of the conditional distributions of

- $L_S^2 + L_T^2$ given $L_T > 0$ and $L_S > 0$,
- L_T^2 given $L_T > 0$,
- L_S^2 given $L_S > 0$, and
- a point mass at zero.

As is proved in the appendix, under the null hypothesis, all four of these conditional distributions are χ^2 for large $\mathcal{C}$ and $\overline{\mathcal{C}}$; the first conditional distribution is χ^2_2, the second and third are both χ^2_1, and the fourth is χ^2_0. So, the CDF of $\Lambda^\star_{ST}$ for large $\mathcal{C}$ and $\overline{\mathcal{C}}$ and under the null hypothesis is given by

$$P(\Lambda^\star_{ST} \leq \lambda) = \frac{1}{4}P(\chi^2_2 \leq \lambda) + \frac{1}{2}P(\chi^2_1 \leq \lambda) + \frac{1}{4}I\{\lambda \geq 0\}. \tag{11}$$

The CDF shown in Equation 11 corresponds to a distribution that is referred to as the chi-bar-square ($\bar{\chi}^2$) distribution (e.g., Dykstra, 1991; Silvapulle & Sen, 2001) that is popular in constrained statistical inference. Consequently, for $\lambda > 0$, the p-value for the test statistic $\Lambda^\star_{ST}$ is calculated as[2]

$$\begin{aligned} P(\Lambda^\star_{ST} > \lambda) = 1 - P(\Lambda^\star_{ST} \leq \lambda) &= \frac{3}{4} - \frac{1}{4}P(\chi^2_2 \leq \lambda) - \frac{1}{2}P(\chi^2_1 \leq \lambda) \\ &= \frac{1}{4}P(\chi^2_2 > \lambda) + \frac{1}{2}P(\chi^2_1 > \lambda)\cdot \end{aligned} \tag{12}$$

One can calculate the critical value of the distribution at significance level of α by solving the equation

$$\frac{1}{4}P(\chi^2_2 > \lambda) + \frac{1}{2}P(\chi^2_1 > \lambda) = \alpha\cdot$$

This equation can be solved by using the R (R Core Team, 2019) function "uniroot". The critical values of the distribution for significance levels of 0.001, 0.01, 0.05, and 0.10 are 11.762, 7.289, 4.231, and 2.952 respectively; that is, the right-hand side of Equation 12 is equal to 0.001, 0.01, 0.05, and 0.10 for λ=11.762, 7.289, 4.231, and 2.952 respectively.

If the alternative hypothesis is true, which typically would occur when an examinee has item preknowledge, one or both of L_S and L_T will have a large positive value and, consequently, the value of $\Lambda^\star_{ST}$ will be large and positive.

[2]Note that negative values of λ are not of interest because $\Lambda^\star_{ST}$ cannot be negative by definition.

Simulation Study

We used simulations based on real data rather than simulations based on data generated from any response-time model and/or IRT model to examine the properties of $\Lambda^{\star}_{ST}$ and to compare the properties of $\Lambda^{\star}_{ST}$ to those of L_S, L_T, the residual-based approach of van der Linden and Guo (2008), and the Bayesian person-fit approach of Fox and Marianti (2017).

Design of the Simulation Study

The simulations were based on the item scores and response times of 18,353 test takers on one form of an English proficiency test that is administered on computers. The test consists of 34 multiple-choice items. The average response times on the items ranged between 21 and 52 seconds and the average per-item response times of the examinees ranged between 9 and 53 seconds. There was no knowledge of examinees benefitting from item preknowledge on the test.

The data set was used to artificially create several simulated data sets that involve different extents of item preknowledge. The following three factors were varied in the simulations:

- an indicator I_{SA} of whether the item scores were affected or not by item preknowledge (the values of I_{SA} were 0 or 1),
- the size of the set of compromised items (4, 7, 10, or 17 items)[3],
- a quantity δ (with values 0, 1, 2, or 3) that determines the speed of those with preknowledge on the compromised items;

To simulate the data and compare the approaches, we repeated the following steps 100 times for each combination of values of the three abovementioned factors:

1. Randomly select 12,235 examinees (who comprise about two-thirds of all the examinees

[3]The case of 17 compromised items out of 34 items was considered because the proportion of compromised items has been found to be quite large in practice (e.g., Cizek & Wollack, 2017; Eckerly et al., 2018).

in the original data set) from the original data set.[4]

2. From the 12,235 examinees, randomly select 1,000 examinees[5] who would play the role of the cheaters, that is, those who benefitted from item preknowledge.

3. From the 34 items in the data set, randomly choose the 4, 7, 10, or 17 items that would play the role of the compromised items.

4. For each combination of a compromised item and a cheater, artifically create item preknowledge by replacing the actual logarithm of response time by the same minus $s\delta$, where s is the standard deviation of the logarithm of response times for the item.

5. When I_{SA} is equal to 0, the item scores were not changed for any examinee—these cases represented the scenario that item preknowledge affects only response times and not the item scores. When I_{SA} is equal to 1, the item scores of the cheaters on the compromised items were replaced by numbers randomly drawn from a Bernoulli distribution with success probablity of 0.9—these cases represented the scenario that item preknowledge affects both response times and item scores.

6. Compute the estimated item parameters for the hierarchical model of van der Linden (2007) that comprises the LNMRT and the 2PLM from the (changed) data set using (a) the R package LNIRT (Fox et al., 2017) and (b) a Fortran program to compute the MMLEs of the item parameters for the hierarchical model (Glas & van der Linden, 2010).

7. Compute the MLEs of the examinee ability and speed parameters from the data set.

8. Compute the statistics—$\Lambda^{\star}_{ST}$, L_S, L_T, and the standardized residuals of van der Linden and Guo (2008)—and the posterior probabilities for the Bayesian person-fit approach of Fox and Marianti (2017) for all the examinees in the (changed) data set. The

[4]Changing this number to other large values such as 5,000, 10,000 or 15,000 did not change the comparative performance of the methods.

[5]Changing 1,000 to 500 and 2,000 did not change the comparative performance of the methods.

MMLEs of the item parameters (produced from the Fortran program) were used in the computation of $\Lambda^{\star}_{ST}$, L_S, and L_T while the Bayesian estimates of the item parameters (produced by the LNIRT package[6]) were used in the computation of the standardized residuals (van der Linden & Guo, 2008) and the posterior probabilities for the Bayesian person-fit approach (Fox & Marianti, 2017).[7]

Steps 4 and 5 indicate that no change to the score or time was made for the examinee-item combinations in which the examinee was a non-cheater or the item was a non-compromised item. The response times are actually not changed in the fourth step for the simulation conditions in which δ is equal to 0. Thus, the null hypothesis is true (that is, there is no preknowledge) for each simulation condition in which $\delta = 0$ and $I_{\text{SA}} = 0$; the Type I error rate of each approach was estimated from these conditions as the proportion of all examinees with a statistically significant value under the approach. For each simulation condition with $\delta > 0$ or $I_{\text{SA}} = 1$ or both, which corresponded to the alternative hypothesis being true, the power of each approach was approximated as the proportion of examinees with item preknowledge that had a significant value under the approach. The conditions with $\delta = 0$ and $I_{\text{SA}} = 1$ represent the cases when only the item scores are affected and response times are not affected by preknowledge. The conditions with $\delta > 0$ and $I_{\text{SA}} = 0$ represent the cases when only the item scores are not affected and response times are affected by preknowledge. The conditions with $\delta > 0$ and $I_{\text{SA}} = 1$ represent the cases when both the item scores and response times are affected by preknowledge.

The Distribution of $\Lambda^{\star}_{ST}$ Under the Null and Alternative Hypotheses

The dashed line in the left panel of Figure 1 shows the kernel-density estimate[8] of the distribution of the values of $\Lambda^{\star}_{ST}$ for the simulation case of no item preknowledge

[6]The LNIRT package can only fit probit models—so the estimates for the probit model were transformed to those for the logistic model using the conversion procedure implied in, for example, Birnbaum (1968, p. 399). The conversion involves the multiplication factor of 1.7 to convert the slope parameter estimates.

[7]The MMLEs and Bayesian estimates of the item parameters were very close—so the approach to estimate item parameters did not have any effect on the comparative performance of the approaches.

[8]The estimate was computed using the function "density" in the R software (R Core Team, 2019).

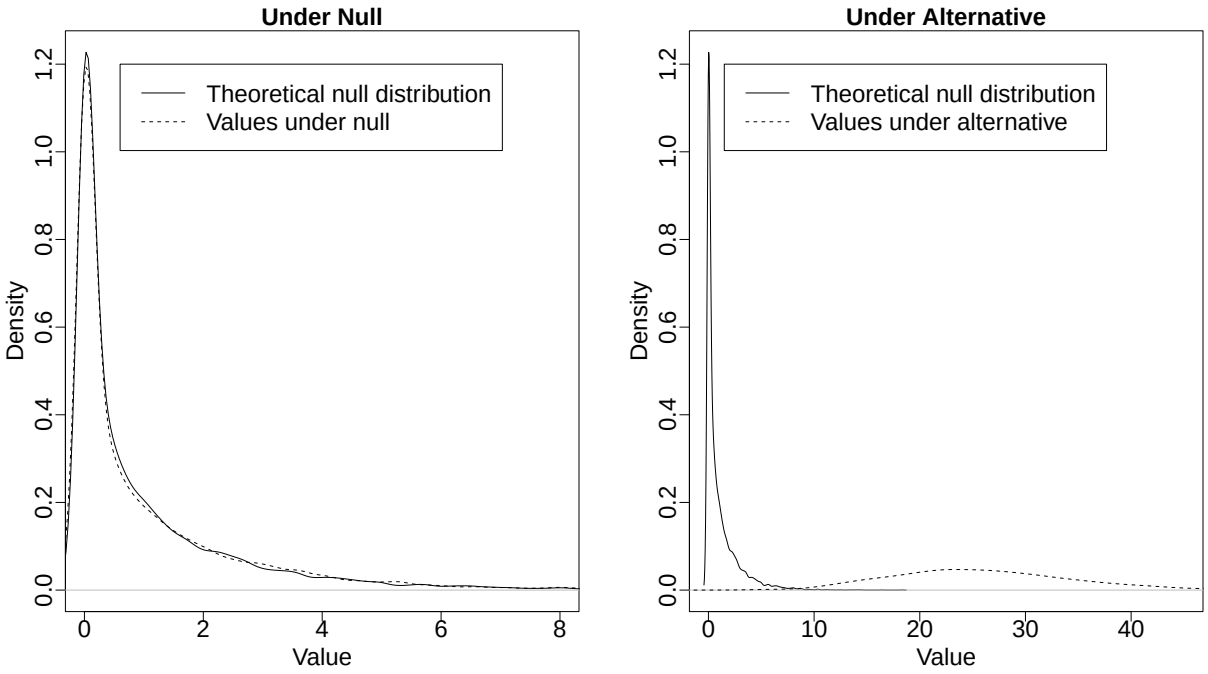

Figure 1: The kernel-density estimate of the distribution of $\Lambda^{\star}_{ST}$ under the null (left panel) and alternative (right panel) hypothesis for the case of 7 compromised items.

when $\mathcal{C}$ includes 7 items. The probability density function (pdf) of the theorized $\bar{\chi}^2$ null distribution given by Equation 11 is also shown in the figure using a solid line. The two lines are very close to each other, indicating that the distribution of the values of $\Lambda^{\star}_{ST}$ under no item preknowledge is very close to the corresponding theorized null distribution and that the Type I error rate of the statistic will be close to the nominal level. Thus, the $\bar{\chi}^2$ null distribution of $\Lambda^{\star}_{ST}$ seems to hold for real data that involve no item preknowledge. The right panel of Figure 1 shows the pdf of the $\bar{\chi}^2$ null distribution and the kernel-density estimate of the values of $\Lambda^{\star}_{ST}$ for the case of preknowledge on 7 items—the panel shows that the distribution of values of $\Lambda^{\star}_{ST}$ under preknowledge is far (towards right) from the distribution under no preknowledge—so $\Lambda^{\star}_{ST}$ is expected to have large power to detect item preknowledge.

The Performance of the Statistics When the Null Hypothesis is True

The Type I error rates of L_S, L_T, $\Lambda^{\star}_{ST}$, the residual-based approach of van der Linden and Guo (2008), and the Bayesian approach (Fox & Marianti, 2017) at the significance

Table 2: The Type I Error Rates at the Level of 0.01

Approach/Statistic	4 items	7 items	10 items	17 items
Bayesian residuals	0.027	0.047	0.053	0.092
Bayesian person-fit	0.052	0.072	0.048	0.063
L_S	0.008	0.010	0.009	0.011
L_T	0.006	0.009	0.008	0.009
$\Lambda^{\star}_{ST}$	0.007	0.010	0.011	0.011

level of 0.01 are shown in Table 2. Columns 2-5 of the table show the rates for 4, 7, 10, and 17 compromised items, respectively.[9] Table 2 indicates that the Type I error rates of the Bayesian person-fit approach and the approach based on residuals (van der Linden & Guo, 2008) are considerably larger than the nominal level. The Type I error rates of $\Lambda^{\star}_{ST}$ are close to the nominal level, which provides favorable evidence for $\Lambda^{\star}_{ST}$ given that the data that were used to compute these rates are not simulated, but real data. Especially, the satisfactory Type I error rates of $\Lambda^{\star}_{ST}$ for 4 and 7 items (Columns 2 and 3 of Table 2) indicate that even though the theoretical result on the null distribution of $\Lambda^{\star}_{ST}$ holds for large $\mathcal{C}$ and $\overline{\mathcal{C}}$, the result seems to hold in simulations even for rather small $\mathcal{C}$. The Type I error rates of L_S and L_T are also close to the nominal level, a finding that agrees with similar findings in Sinharay (2017a) and Sinharay (2019).

The Performance of the Statistics When the Null Hypothesis is False

The power (at level 0.01) of L_S, L_T, $\Lambda^{\star}_{ST}$, the residual-based approach of van der Linden and Guo (2008), and the Bayesian approach (Fox & Marianti, 2017) are shown in Figures 2 and 3 for different numbers of compromised items and δ. Figures 2 and 3 respectively correspond to the cases with $I_{\text{SA}} = 0$ and $I_{\text{SA}} = 1$. The four panels of each figure show the power of the approaches when the number of compromised items was 4, 7, 10, and 17, respectively. In each panel, the values of power for $\Lambda^{\star}_{ST}$, L_T, L_S, the residual-based

[9]The use of any "number of compromised items" may seem at odds with the computation of the Type I error rate given that the latter corresponds to no item preknowledge. However, to compute statistics such as $\Lambda^{\star}_{ST}$, one needs to a assign a set of items as compromised (even though they are actually not compromised) and Table 2 shows the Type I error rates for various sizes of this set.

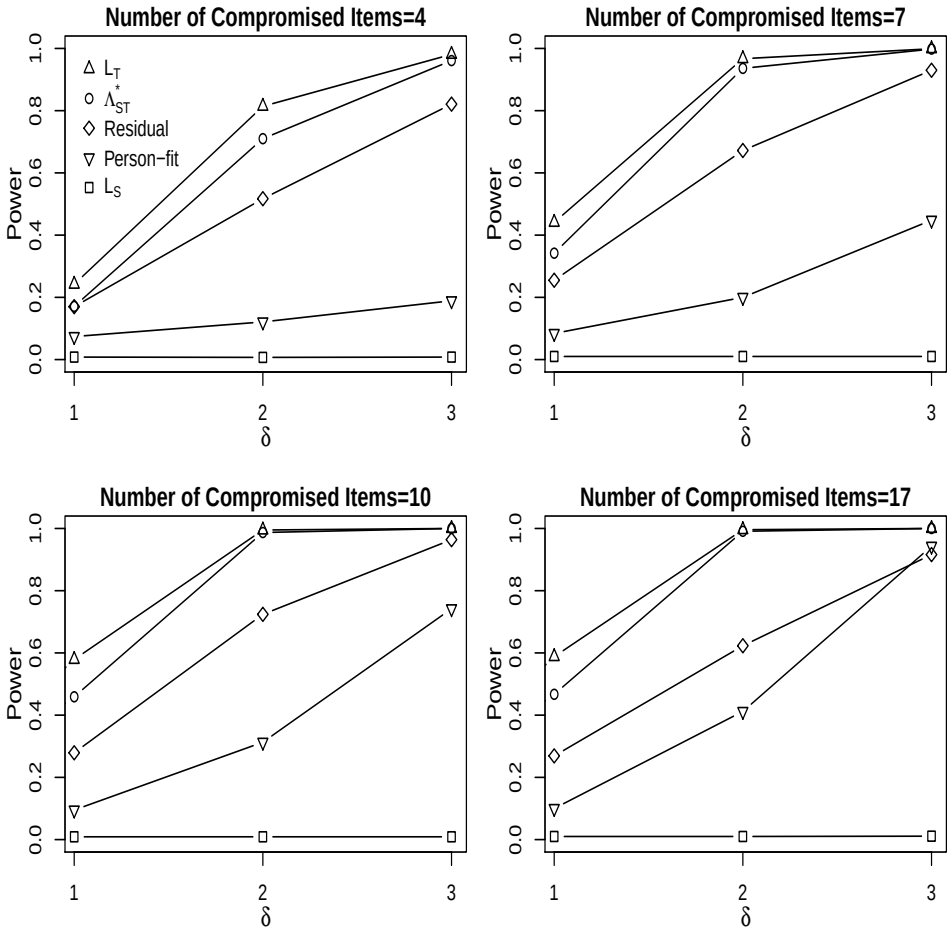

Figure 2: The power of $\Lambda^{\star}_{ST}$, L_T, L_S, the residual-based approach and the person-fit approach at significance level of 0.01 when only response times are affected by preknowledge.

approach (van der Linden & Guo, 2008), and the person-fit approach (Fox & Marianti, 2017) are shown using hollow circles, hollow triangles, hollow squares, hollow diamonds, and hollow inverted-triangles, respectively, joined by a solid line. Given the simulation design, all simulation conditions in Figure 2 represent the conditions when only the response times were affected by item preknowledge, the simulation conditions with δ=0 in Figure 3 represent the conditions when only the item scores were affected by preknowledge, and the simulation conditions with δ >0 in Figure 3 represent the conditions when both the item scores and response times were affected by item preknowledge.

Figures 2 and 3 indicate that in general, the power of each approach increases as the number of compromised items increases, which implies that the chance of detecting item

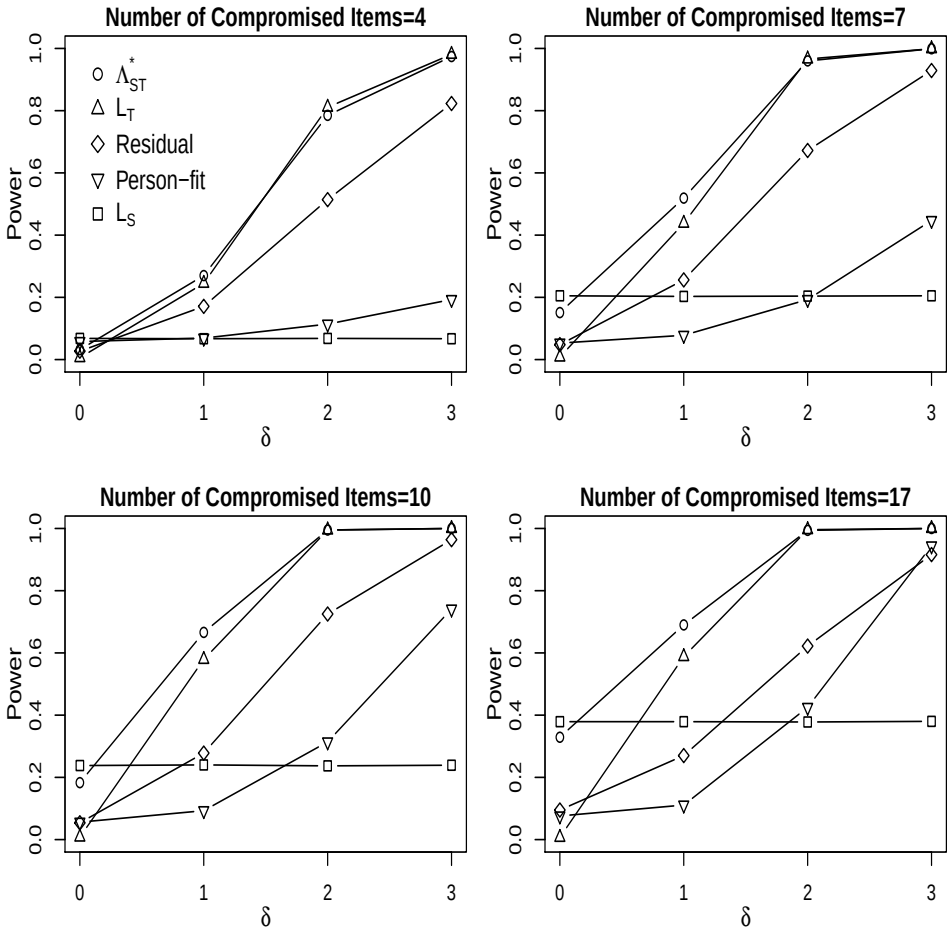

Figure 3: The power of $\Lambda^{\star}_{ST}$, L_T, L_S, the residual-based approach and the person-fit approach at level of 0.01 when both item scores and response times are affected by preknowledge.

preknowledge increases as the extent of preknowledge increases. The figures also indicate that the power of all approaches except L_S increases as δ increases, which is expected because increasing δ means faster responding to the compromised items and all the approaches except L_S incorporate information on speed. The power of $\Lambda^{\star}_{ST}$ is larger than 0.7 when δ is 2 or 3. Figures 2 and 3 indicate that the residual-based approach (van der Linden & Guo, 2008) and the Bayesian person-fit approach (Fox & Marianti, 2017) have smaller power than $\Lambda^{\star}_{ST}$ in all simulation cases. Figure 2 shows that when only response times are affected by preknowledge and item scores are not, L_T will be more powerful, but only by a small margin, compared to $\Lambda^{\star}_{ST}$. Similarly, the values for δ=0 in Figure 3 indicate that when only item scores are affected by preknowledge and response times are not, L_S

will be more powerful, but only by a small margin, compared to $\Lambda^{\star}_{ST}$. Figure 3 shows that when both item scores and response times are affected by preknowledge, $\Lambda^{\star}_{ST}$ is slightly more powerful than L_T and much more powerful than L_S. Given that both item scores and response times are likely to be affected by preknowledge for operational tests (some evidence favoring this assertion was provided by Kasli & Zopluoglu, 2018), Figures 2 and 3 indicate that $\Lambda^{\star}_{ST}$ is the most appropriate candidate (among those considered in this paper) for detecing item preknowledge when both item scores and response times of the examinees are available.

The simulation cases with $\delta > 0$ or $I_{\text{SA}} = 1$ or both also allow one to estimate the false alarm (FA) rates of each approach as the proportion of examinees with no item preknowledge that had a significant value under the approach.[10] The FA rate of $\Lambda^{\star}_{ST}$ was always smaller than the nominal level—this is favorable evidence for the statistic given that the item parameters in these simulation cases were estimated from the contaminated data sets (contaminated in the sense that they included some examinees with preknowledge). Given that an investigator would typically have to work with contaminated data sets in practice, these FA rates imply that $\Lambda^{\star}_{ST}$ will not falsely identify examinees too often in practice.

Real Data Example

Item scores and response times on two forms of a non-adaptive computerized licensure test were available for 1,624 and 1,629 examinees, respectively. The licensure test comprises 170 dichotomous items. The two forms included 63 and 61 items, respectively, which were known to have been compromised. Further, a rigorous investigative process identified as possible cheaters 41 and 42 examinees, respectively, on the two forms. It is not known what methods were used or exactly what types of test fraud were found in the investigative process. For further details about the data set, see Cizek and Wollack (2017, p. 14) who

[10]Note that the data sets from which the FA rates are computed include some examinees with preknowledge, whereas the data sets from which the Type I error rates are computed include no examinees with preknowledge.

stated that while some of the test fraud involved examinees having item preknowledge, other types of fraud were possible as well. Also, while all the examinees identified as cheaters were believed to have engaged in test fraud, it is certainly possible that other examinees should have been identified as cheaters, but were not. Researchers such as Boughton et al. (2017), Eckerly (2017), Kasli and Zopluoglu (2018), and Sinharay (2017a) used these data sets to detect item preknowledge, Fox and Marianti (2017) used these data sets to detect person misfit, and Zopluoglu (2017) used these data sets to detect answer-copying.

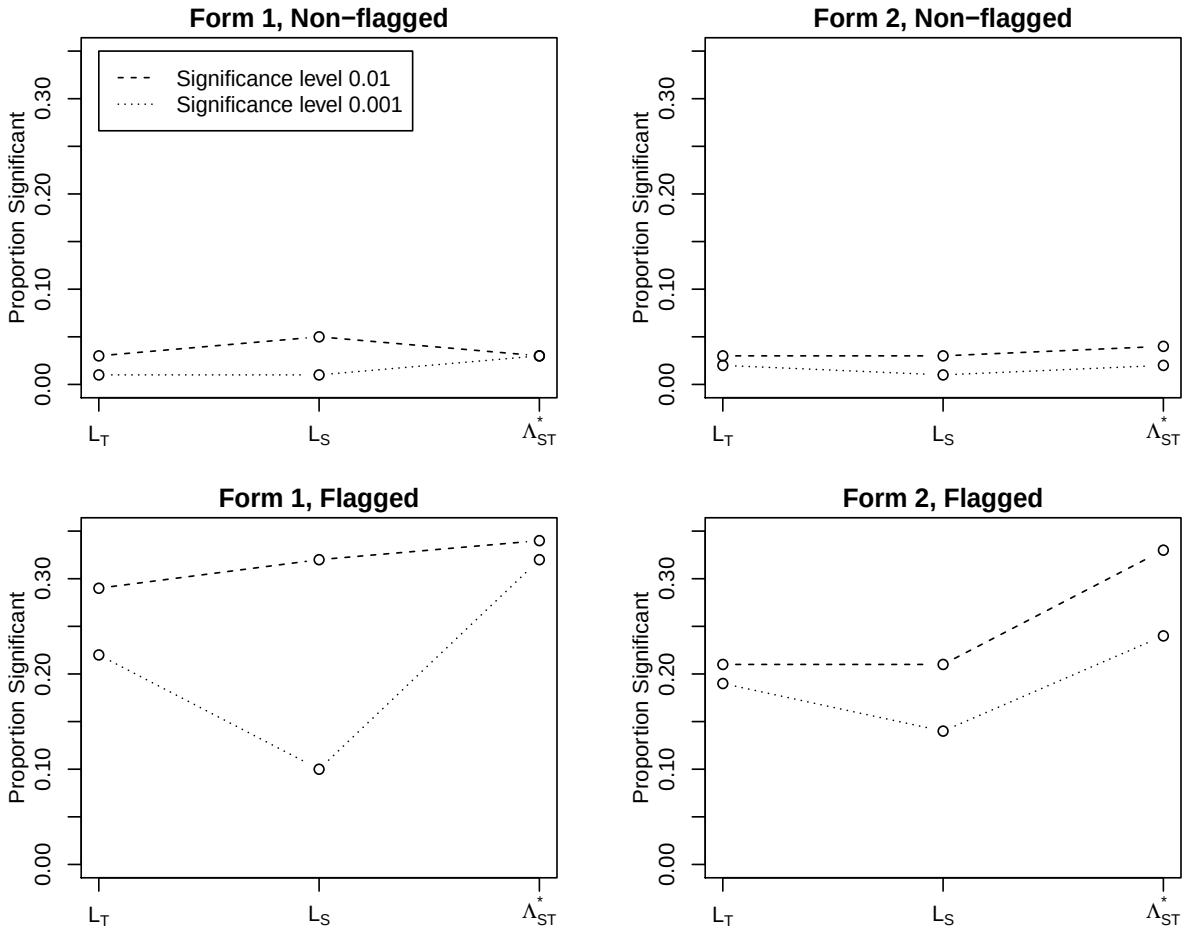

Figure 4: The Proportion of examinees for whom L_T, L_S, and $\Lambda^{\star}_{ST}$ were significant for the real data.

The item parameters for the hierarchical model of van der Linden (2007) were estimated from data for each form using a Fortran program to compute the MMLEs of

item parameters under the assumption that the 2PLM is used for the item scores and the LNMRT is used for the response times. The values of L_T, L_S and $\Lambda^{\star}_{ST}$ were computed from the two data sets using the estimated item parameters. The residual-based approach (van der Linden & Guo, 2008) and the person-fit approach (Fox & Marianti, 2017) were not considered because of their inflated Type I error rate and smaller power in the simulations.

Figure 4 shows the proportions of examinees (along the vertical axis) for whom the three statistics were significant at significance levels of 0.001 (dotted line) and 0.01 (dashed line) for the two forms. The top two panels of the figure show the proportions significant among the examinees who were not flagged as possible cheaters and the bottom two panels show the proportions significant among the examinees who were flagged as possible cheaters by the licensure organization. The title of each panel indicates the form and flag status of the examinees. The range of the vertical axis is the same in all the panels. Note that the proportions are computed from much smaller number of examinees (41 and 42 respectively) in the bottom two panels compared to the top two panels (1,583 and 1,587, respectively).

The top two panels of Figure 4 indicate that the proportions of significant values for $\Lambda^{\star}_{ST}$ are close to those for L_S and L_T among non-flagged examinees. The bottom two panels of Figure 4 indicate that the proportions of significant values for $\Lambda^{\star}_{ST}$ are substantially larger than those for L_S and L_T among flagged examinees for significance level of 0.001 for Form 1 and for both significance levels for Form 2; these proportions indicate that the use of $\Lambda^{\star}_{ST}$ will often lead to the detection of a larger number of examinees compared to the use of only one among L_S and L_T.

Further insight on the relationship between L_S, L_T, and $\Lambda^{\star}_{ST}$ is provided by Figure 5 that shows the values of the $\Lambda^{\star}_{ST}$ statistic (along Y-axis) versus those of the L_T statistic (left panel) and the L_S statistic (right panel) for the 41 examinees who were flagged by the licensure organization for Form 1. Each circle (hollow or solid gray) corresponds to a flagged examinee. Horizontal and vertical dashed lines are shown at the 99.9th percentile of the respective null distribution (any value larger than this quantile is statistically significant at level 0.001). The range of the Y-axis is the same in the two panels of the figure. The figure shows that $\Lambda^{\star}_{ST}$ increases with an increase in either of L_T and L_S, but this relationship is

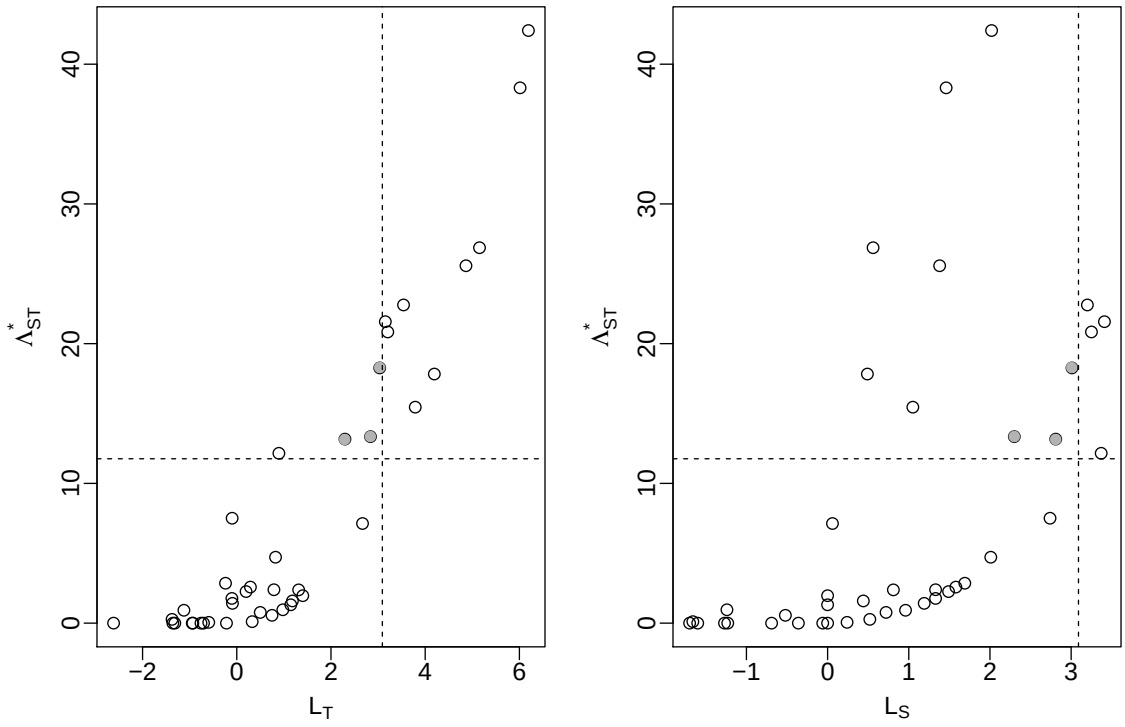

Figure 5: A scatter-plot of $\Lambda^{\star}_{ST}$ versus L_S and L_T for the 41 flagged examinees for Form 1.

more pronounced for L_T. The figure also shows that among the 13 examinees for whom $\Lambda^{\star}_{ST}$ is significant at level of 0.001, L_T was significant for nine examinees, L_S was significant for four, and both L_T and L_S were significant for three. The three solid gray circles correspond to examinees for whom neither of L_T and L_S was significant, but $\Lambda^{\star}_{ST}$ was significant. Thus, the use of $\Lambda^{\star}_{ST}$ would lead to the the detection of examinees who are detected by neither of L_T and L_S. This result is in agreement with the finding in the simulations that when an aberrant examinee answers faster and performs better on the compromised items, the power of $\Lambda^{\star}_{ST}$ is larger than that of either of L_T and L_S. It is also interesting to note that $\Lambda^{\star}_{ST}$ is significant for all the examinees for whom either L_T or L_S was significant.

Figure 6 includes a comparison of the response times versus average response times for the whole sample, of the three examinees for whom $\Lambda^{\star}_{ST}$ were the largest for Form 1 and one additional randomly chosen examinee for whom $\Lambda^{\star}_{ST}$ was not significant. In each panel, the logarithm of the response times of an examinee on the individual items is shown along the Y-axis and the logarithm of the average response time of the items over the whole

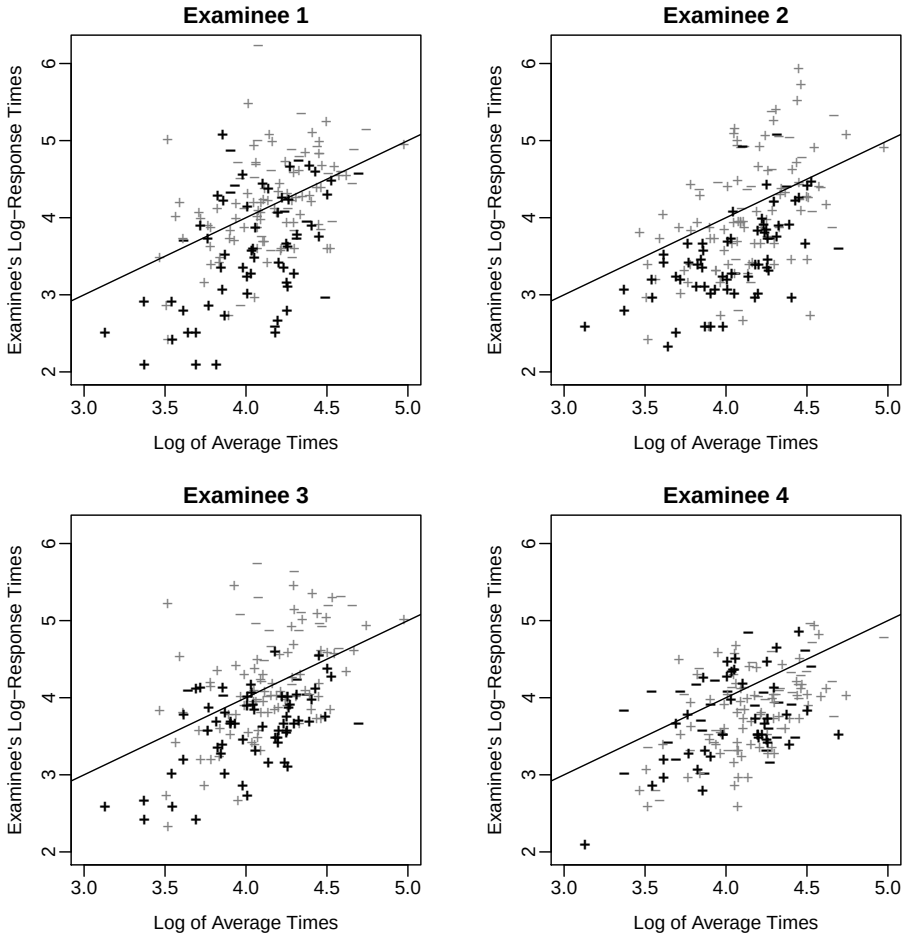

Figure 6: Response times of four examinees.

sample is shown along the X-axis. A plus sign and a minus sign respectively correspond to items that the examinee answered correctly and incorrectly. The black plus or minus signs correspond to the compromised items while the gray plus or minus signs correspond to the non-compromised items. A diagonal line is provided in each panel for convenience. Each of Examinees 1-3 answered several items correctly and faster than the average and more of these items are compromised items, which causes several bold plus signs to appear towards the bottom and far from the diagonal line of Panels 1-3; the items on which Examinee 3 (and Examinees 1 and 2, to a certain extent) spent most time are mostly non-compromised items (that is clear from the abundance of gray symbols towards the top of the bottom left panel); in contrast, the plot for Examinee 4 does not reveal any pattern.

Conclusions and Recommendations

Examinees benefitting from item preknowledge is a serious problem in educational assessments (e.g., Wollack & Schoenig, 2018). In this paper, we suggest a new approach to detect item preknowledge based on item scores and response times. The approach is based on ideas from constrained statistical inference (Silvapulle & Sen, 2001). Wollack and Schoenig (2018) divided the statistical methods used to detect cheating into six categories and the suggested approach combines two of those categories (response-time methods and score differencing). The suggested approach seems promising. The asymptotic distribution of the suggested statistic under the null hypothesis of no item preknowledge is a mixture of χ^2 distributions, the estimated Type I error rate of the new statistic was found to be close to the nominal level, and the power of the statistic was found to be larger in comparison to the existing statistics. The computations required to implement the new statistic are not intensive.

The new statistic should not be used by itself to detect item preknowledge in operational testing. Instead, as recommended by, for example, van der Linden and Guo (2008), the new statistic should be employed as a part of quality control and/or as secondary evidence, along with other statistics and non-statistical evidence (e.g., Hanson, Harris, & Brennan, 1987), in investigations of test fraud. Also, if the goal is to detect aberrant responding in general using response times, then the approaches of, for example, Fox and Marianti (2017), Lee and Wollack (2017), van der Linden and Guo (2008), and Wang et al. (2018) should be used instead of the new statistic.

The statistic $\Lambda^{\star}_{ST}$ applies only to the case where a subset of all the items is compromised. Thus, the statistic cannot be applied when all or almost all items are compromised—the only (suboptimal) solution in such a case is to compare the performance of the examinees to the performance predicted from covariates such as scores on other tests. Also, $\Lambda^{\star}_{ST}$ will have low power if only a few items are compromised, as clear from Figures 2 and 3. In addition, Λ_s can only be applied when the set of compromised items is known. Typically, such a case arises when the test administrators become aware after an administration about some items possibly being compromised (one example of this is that the test administrators

come across a website where some test items have been posted). In cases when the set of compromised items is not precisely known, $\Lambda^{\star}_{ST}$ can be applied if the examinees were also administered a set of items that are new (that is, they were not administered in the past), as was the case in the study of item compromise by Smith and Davis-Becker (2011)—the old and new items would respectively play the roles of the compromised and non-compromised items in such an analysis.

When the proportion of examinees with item preknowledge is large, item-parameter estimates (that are typically estimated from the available examinee sample that will include those with preknowledge) will be biased and the $\Lambda^{\star}_{ST}$ statistic may not perform well in detecting item preknowledge. For example, for a non-adaptive test for which the item parameters are estimated from the examinee sample, the time-intensity parameters of the compromised items would be substantially underestimated and the difficulty parameters will be underestimated if a large number of examinees have preknowledge of those items because they would answer those items faster and more correctly. As a consequence, the speed-parameter estimate and ability estimate based on the compromised items ($\hat{\tau}_c$ and $\hat{\theta}_c$) would be substantially underestimated for those with preknowledge and without preknowledge—this underestimation would make L_{T+}'s and L_{S+}'s smaller than what they actually are and would in turn lead to smaller power and a false alarm rate that is smaller than the nominal level of $\Lambda^{\star}_{ST}$. This phenomenon was verified from a comparison of the results reported in this paper to those from an additional set of simulations[11] in which item parameters were not estimated in sixth step of the simulations and the true item parameters were used instead. One possible solution in the face of item preknowledge for a large proportion of examinees involves the four-step purification process of (a) estimating item parameters from the full sample, (b) computing $\Lambda^{\star}_{ST}$ for the full sample using item-parameter estimates computed in the previous step, (c) reestimating the item parameters from the subset of the sample that does not have significant values of $\Lambda^{\star}_{ST}$, and (d) computing $\Lambda^{\star}_{ST}$ for the full sample using the item-parameter estimates computed in the

[11]The results from these additional simulations are not reported in this paper and can be obtained from the authors upon request.

previous step. Such purification procedures have been successfully applied in other types of person-level analysis such as person-fit analysis (e.g., Patton, Cheng, Hong, & Diao, 2019). However, when the proportion of examinees benefitting from item preknowledge is very large (say, larger than 0.5), then even a purification would not work well and retesting all examinees would probably be the only reasonable choice. However, tests for which a large proportion of examinees benefitted from item preknowledge are very rare, if not unheard of.

Thus, the above discussion suggests that while $\Lambda^{\star}_{ST}$ is expected to have a small Type I error rate and small false alarm rate, the statistic is expected to be most useful (in terms of being reliable and having a large power) when a small percentage of examinees have preknowledge of a moderately large number of items. In other cases, $\Lambda^{\star}_{ST}$ will have small power (an example of such a case is preknowledge on a small number of items by a small percentage of examinees) or will be unreliable (when a large percentage of examinees have preknowledge of some items).

Though this paper suggests a method that seems to be promising, this paper has several additional limitations. First, one could examine the consequences of misfit of the model on the properties of the new statistic in future research. Second, the statistic $\Lambda^{\star}_{ST}$ should be calculated for more data sets, both simulated and real. Especially, while the statistic applies to adaptive tests as well, we did not compute it from any data (real or simulated) originating from of an adaptive test and it is possible to compute $\Lambda^{\star}_{ST}$ using data originating from adaptive tests. Third, it is possible to compare the suggested approach to the Bayesian approach of detecting item preknowledge using mixture hierarchical IRT models (e.g., Lee & Wollack, 2017; Wang et al., 2018) that find both compromised items and aberrant examinees. Fourth, the item parameters were assumed known in the derivation of the null distution of the new statistic and it is possible to explore approaches to account for the uncertainty of the item parameters in the distribution of the new statistic. However, such an approach would almost surely have to be Bayesian. Finally, while the suggested approach is an important initial step towards detecting item preknowledge using both item scores and response times, the extension of the suggested approach to response-time models other than LNMRT and to the case when

the set of compromised items is unknown are potential areas of future research.

References

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.

Baker, F. B., & Kim, H., S. (2004). *Item response theory: Parameter estimation techniques (2nd ed.)*. New York, NY: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Boughton, K., Smith, J., & Ren, H. (2017). Using response time data to detect compromised items and/or people. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 177–190). Washington, DC: Routledge.

Box, G. E. P., & Muller, M. E. (1958). A note on the generation of random normal deviates. *The Annals of Mathematical Statistics*, *29*, 610–611. (doi=10.1214/aoms/1177706645)

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

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

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

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education*, *9*, 47–64. (doi=10.1207/s15324818ame0901_5)

Dykstra, R. (1991). Asymptotic normality for chi-bar-square distributions. *Canadian Journal of Statistics*, *19*, 297–306. (doi=10.2307/3315395)

Eckerly, C., Smith, R., & Lee, Y. (2018, October). *An introduction to item preknowledge detection with real data applications.* Paper presented at the Conference on Test Security, Park City, UT.

Eckerly, C. A. (2017). Detecting item preknowledge and item compromise: Understanding the status quo. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 101–124). Washington, DC: Routledge.

Fox, J.-P., Klein Entink, R. H., & Klotzke, K. (2017). *LNIRT: Lognormal response time item response theory models.* (R package version 0.2.0)

Fox, J.-P., & Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement*, *54*, 243–262. (doi=10.1111/jedm.12143)

Glas, C. A. W., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology*, *63*, 603–626. (doi=10.1348/000711009x481360)

Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying (ACT research report series no. 87-15).* Iowa City, IA: American College Testing.

Kasli, M., & Zopluoglu, C. (2018, October). *Do people with item pre-knowledge really respond faster to items they had prior access? An empirical investigation.* Paper presented at the Conference on Test Security, Park City, UT.

Klein Entink, 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*, 21–48. (doi=10.1007/s11336-008-9075-y)

Lee, S. Y., & Wollack, J. (2017, October). *A mixture model to detect item preknowledge using item responses and response times.* Paper presented at the Conference on Test Security, Madison, WI.

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral*

Statistics, *39*, 426–451. (doi=10.3102/1076998614559412)

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

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*(1), 56–74. (doi:10.1080/00273171.2014.962684)

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

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice*, *35*(1), 38–47. (doi=10.1111/emip.12102)

R Core Team. (2019). *R: A language and environment for statistical computing.* Vienna, Austria.

Silvapulle, M. J., & Sen, P. K. (2001). *Constrained statistical inference: Order, inequality, and shape constraints*. New York, NY: John Wiley & Sons, Inc.

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

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

Sinharay, S. (2019). *Detection of item preknowledge using response times.* (Manuscript under preparation)

Sinharay, S., Duong, M. Q., & Wood, S. W. (2017). A new statistic for detection of aberrant answer changes. *Journal of Educational Measurement*, *54*, 200–217. (doi=10.1111/jedm.12141)

Smith, R. W., & Davis-Becker, S. L. (2011, April). *Detecting suspect examinees: An application of differential person functioning analysis.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.

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

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308. (doi=10.1007/s11336-006-1478-z)

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

van Rijn, P. W., & Ali, U. S. (2017). A comparison of item response models for accuracy and speed of item responses with applications to adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *70*, 317–345. (doi=10.1111/bmsp.12101)

Veerkamp, W. J. J., & Glas, C. A. W. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, *25*, 373–389. (doi=10.2307/1165221)

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*, 469–501. (doi=10.3102/1076998618767123)

Wang, X., Liu, Y., & Hambleton, R. K. (2017). Detecting item preknowledge using a predictive checking method. *Applied Psychological Measurement*, *41*, 243–263. (doi=10.1177/0146621616687285)

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.

Zopluoglu, C. (2017). Similarity, answer copying, and aberrance: Understanding the status quo. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests*

(pp. 25–46). Washington, DC: Routledge.

Appendix A: Some Proofs Regarding the Asymptotic Null Distribution of $\Lambda^{\star}_{ST}$

Null Distribution of L_S^2 given $L_S > 0$: The joint probability of $L_S^2 < u^2$ and $L_S > 0$ is equal to

$$\begin{aligned} P(L_S^2 \leq u^2,\ L_S > 0) &= P(0 < L_S \leq u) \\ &= P(0 < Z \leq u), \end{aligned} \tag{A1}$$

where $Z \sim N(0,1)$ is a standard normal random variable. Equation A1 holds because L_S has a standard normal asymptotic null distribution (Sinharay, 2017a). Because of the symmetry of the standard normal distribution around zero, we have

$$\begin{aligned} P(L_S^2 \leq u^2,\ L_S > 0) &= P(0 < Z \leq u) \\ &= \frac{1}{2}P(-u < Z \leq u) \\ &= \frac{1}{2}P(Z^2 \leq u^2). \end{aligned}$$

We also have $P(L_S > 0) = \frac{1}{2}$, so

$$\begin{aligned} P(L_S^2 \leq u^2 | L_S > 0) &= \frac{P(L_S^2 \leq u^2,\ L_S > 0)}{P(L_S > 0)} \\ &= P(Z^2 \leq u^2) = P(\chi_1^2 \leq u^2). \end{aligned}$$

Therefore, the null distribution of L_S^2 given $L_S > 0$ is the χ_1^2 distribution. The null distribution of L_T^2 given $L_T > 0$ can be proved to be the χ_1^2 distribution in a similar manner.

Null Distribution of $L_S^2 + L_T^2$ given $L_S > 0$ and $L_T > 0$: To derive the asymptotic conditional distribution of $L_S^2 + L_T^2$ given $L_T > 0$ and $L_S > 0$ under the null hypothesis, we use the result that the polar coordinates of two independent normal random variables are independent of one another (Box & Muller, 1958). In terms of the statistics L_S and L_T, the polar coordinates are

$$\begin{aligned} r &= \sqrt{L_S^2 + L_T^2}, \text{ and} \\ \varphi &= \text{atan2}(L_T, L_S), \end{aligned}$$

where r is the distance of the point (L_s, L_T) from the origin and φ is the angle between the horizontal (L_S) axis and the ray going through the point (L_S, L_T). Note that atan2(x, y)

returns a single value θ such that for some $r > 0$, $x = r\sin(\theta)$ and $y = r\cos(\theta)$. Because L_S and L_T are standard normal variables under the null hypothesis, r^2 is a χ^2_2 random variable, and φ is uniform on $[0, 2\pi]$ (Box & Muller, 1958). Furthermore, r and φ are independent of one another and $L_S > 0$ and $L_T > 0$ if and only if $0 \leq \varphi \leq \frac{\pi}{2}$. Therefore, the conditional distribution of $L_S^2 + L_T^2$ given $L_S > 0$ and $L_T > 0$ is equivalent to the conditional distribution of r^2 given $0 \leq \varphi \leq \frac{\pi}{2}$. Because r^2 and φ are independent of one another, the conditional distribution of $L_S^2 + L_T^2$ given $L_S > 0$ and $L_T > 0$ is equal to the marginal distribution of r^2, which is χ^2_2.