Detection of Item Preknowledge Using Response Times

Sandip Sinharay, Educational Testing Service

An Updated Version of this document will appear in the Applied Psychological Measurement. The website for the journal is https://journals.sagepub.com/home/apm

The citation for the article is: Sinharay, S. (in press). Detection of item preknowledge using response times. Applied Psychological Measurement.

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.

Detection of Item Preknowledge Using Response Times

Sandip Sinharay, Educational Testing Service

November 19, 2019

Note: Any opinions expressed in this publication are those of the author and not necessarily of Educational Testing Service.

Detection of Item Preknowledge Using Response Times

Abstract

Benefiting from item preknowledge (e.g., McLeod, Lewis, & Thissen, 2003) is a major type of fraudulent behavior during educational assessments. This paper suggests a new statistic that can be used for detecting the examinees who may have benefitted from item preknowledge using their response times. The statistic quantifies the difference in speed between the compromised items and the non-compromised items of the examinees. The distribution of the statistic under the null hypothesis of no preknowledge is proved to be the standard normal distribution. A simulation study is used to evaluate the Type I error rate and power of the suggested statistic. A real data example demonstrates the usefulness of the new statistic that is found to provide information that is not provided by statistics based only on item scores.

Key words: Item compromise; likelihood ratio statistic; Wald statistic.

Item preknowledge refers to some examinees having prior access to test questions and/or answers before taking the test. For example, Educational Testing Service (ETS) discovered in 2002 that students in several countries were benefiting from websites showing live items used in the Graduate Record Examination (GRE); the phenomenon was so widespread that average scores on GRE Verbal increased by 100 points (out of a possible 800 points) in one country and 50 points in another (Kyle, 2002). The leaked/shared/memorized items are usually referred to as "compromised" items. The focus of this paper will be on detecting examinees who may have benefited from item preknowledge. This paper considers only the case when the investigator knows which items are compromised.

Research on detection of item preknowledge has mostly been based on the item scores of the examinees. Researchers such as Drasgow, Levine, and Zickar (1996), McLeod et al. (2003), Shu, Henson, and Luecht (2013), and Sinharay (2017a) suggested a variety of methods based on item scores to detect item preknowledge. Sinharay (2017a) suggested the L_s statistic, which is based on the likelihood ratio; Sinharay (2017a) and Sinharay (2017b) demonstrated that the L_s statistic performed satisfactorily in detecting item preknowledge. However, given the current popularity of online testing, response times are being recorded for an increasing number of tests (e.g., Wang, Xu, Shang, & Kuncel, 2018) and researchers have realized the importance of response times in detecting various types of test fraud including item preknowledge. Consequently, researchers such as Fox and Marianti (2017), Sinharay (2018), Toton and Maynes (2019), van der Linden and Guo (2008), and Wang et al. (2018) have suggested a variety of approaches that can be used to detect item preknowledge based on response times. This paper suggests a simple frequentist approach to detect item preknowledge using response times. The new approach is essentially an examination of whether the examinees answer the compromised items faster in comparison to the non-compromised items.

The next section includes reviews of (a) a popular response time model (RTM), (b) the existing approaches for estimation of the parameters of the model, and (c) the existing approaches for detection of item preknowledge using response times. The Methods section includes the description of a new statistic for detection of item preknowledge and of its null

distribution. The Simulation section includes a comparison of the Type I error rate and power of the new approach to those of two existing approaches. The Real Data section includes an application of the new approach to an operational data set. Discussion and conclusions are provided in the last section.

Background

The Lognormal Model for Response Times

Let us consider a test that includes I items. Let t_i denote the response time of a randomly chosen examinee[1] on item i, where $i = 1, 2, \cdots, I$. Let us define

$$y_i = \log(t_i)\cdot$$

Under the lognormal model for response times (LNMRT; van der Linden, 2006), y_i, $i = 1, 2, \ldots, I$, are independent given τ and

$$y_i|\tau \sim \mathcal{N}\left(\beta_i - \tau, \frac{1}{\alpha_i^2}\right), \tag{1}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. The parameter τ is the examinee's speed parameter; a larger value of the parameter results in smaller expected response times on all the items for the examinee. The parameter β_i is the time-intensity parameter for item i; a larger value of the parameter results in larger expected response times for all examinees on the item. The parameter α_i is the discrimination parameter for item i; a larger value of the parameter leads to more information on and hence smaller standard error of the examinee speed parameters. To estimate the item parameters of the LNMRT using a marginal maximum likelihood approach or to perform a Bayesian inference on the examinee ability, one assumes a prior distribution $g(\tau)$ on τ. As is common in applications of LNMRT (see, for example, van der Linden & Guo, 2008), $g(\tau)$ is assumed to be the normal distribution with mean 0 and variance σ^2 in this paper.

[1]No subscript is used here for the examinees because the existing statistics and the new statistic will be described for one randomly chosen examinee.

The LNMRT is arguably one of the most popular RTMs. The model was considered, either to analyze only the response times, or to analyze the response times and item scores, by, for example, Bolsinova and Tijmstra (2018), Boughton, Smith, and Ren (2017), Glas and van der Linden (2010), Qian, Staniewska, Reckase, and Woo (2016), Sinharay (2018), van der Linden (2007), van der Linden (2009), van der Linden (2016), van der Linden and Glas (2010), and van der Linden and Guo (2008). Bolsinova and Tijmstra (2018, p. 13) commented that the LNMRT is used in most applications of RTMs.

Estimation of the Item and Examinee Parameters of the LNMRT

A Gibbs sampler (e.g., Gelman et al., 2014, p. 276) was suggested by van der Linden (2006) to estimate the item parameters of the LNMRT. That approach has been used in most applications of the model and the R package LNIRT (Fox, Klein Entink, & Klotzke, 2017) can be used to implement the Gibbs sampler. Glas and van der Linden (2010) suggested an approach to compute the marginal maximum likelihood estimates (MMLEs) of the item parameters when the LNMRT is used along with the three-parameter logistic model (3PLM) to jointly analyze both response times and item scores. Finger and Chee (2009) showed how one can use factor analysis to obtain the MMLEs of the item parameters of the LNMRT when it is used as a stand-alone model, as in van der Linden (2006). The R package *lavaan* (Rosseel, 2012), which is used to perform factor analysis and structural equation modeling (SEM), was used in this paper, both in the simulation study and real data analysis, to estimate the item parameters of the LNMRT.

van der Linden (2006) showed that given α_i^2's and β_i's, the MLE of the person speed parameter τ for the LNMRT can be obtained as

$$\hat{\tau} = \frac{\sum_i \alpha_i^2(\beta_i - y_i)}{\sum_i \alpha_i^2}. \tag{2}$$

Equation 2 was used in this paper (both in the simulation study and real data analysis) to estimate the person speed parameters of the LNMRT. Because $\hat{\tau}$ is a linear combination of normal random variables y_i's, it has a normal distribution (because of, for example,

Theorem 2.4.1 of Anderson, 1984, p. 25) with mean and variance given by

$$\mathrm{E}(\hat{\tau}) = \tau \text{ and } \mathrm{Var}(\hat{\tau}) = \frac{1}{\sum_i \alpha_i^2} \tag{3}$$

when the LNMRT fits the data.

Detection of Item Preknowledge Using Response Times: A Review

Let c denote the set of compromised items that was administered to the randomly chosen examinee considered above. Let $\overline{c}$ denote the set of non-compromised items that were administered to the examinee. Together, c and $\overline{c}$ constitute all the I items administered to the examinee. Let $\boldsymbol{y}_c$ and $\boldsymbol{y}_{\overline{c}}$ denote the collection of logarithms of response times of the examinee on the items in c and $\overline{c}$, respectively.

Sinharay (2018) suggested for the LNMRT a person-fit statistic χ_{pf} that is given by

$$\chi_{pf} = \sum_i \alpha_i^2 (y_i - \beta_i + \hat{\tau})^2, \tag{4}$$

and showed that when the LNMRT fits the data, χ_{pf} follows the χ^2 distribution with $I-1$ degrees of freedom. The χ_{pf} statistic can be used to detect item preknowledge. Marianti, Fox, Avetisyan, Veldkamp, and Tijmstra (2014) and Fox and Marianti (2017) suggested a Bayesian person-fit analysis approach that was found to perform very similarly, but slightly worse than the χ_{pf} statistic by Sinharay (2018)—so their Bayesian approach is not considered henceforth.

A Bayesian approach was suggested by van der Linden and Guo (2008) to determine if the response time of an examinee-item combination is aberrant and the approach can be used to detect item preknowledge. It was proved by van der Linden and Guo (2008) that the posterior distribution of the predicted value of the log-response time on item i conditional on $\boldsymbol{y}_{-i} = (y_1, y_2, \cdots, y_{i-1}, y_{i+1}, \cdots, y_I)$, is normal. Then, the standardized residual is computed as

$$e_i = \frac{y_i - E(y_i|\boldsymbol{y}_{-i})}{\sqrt{\mathrm{Var}(y_i|\boldsymbol{y}_{-i})}}. \tag{5}$$

If the absolute value of e_i is larger than an appropriate quantile of the standard normal distribution, the response time for the examinee for item i is considered aberrant. One can compute the e_i's for an examinee over all the compromised items and then combine information over these items for the examinee to assess item preknowledge, as in Boughton et al. (2017, p. 181) and Qian et al. (2016). In this paper, an examinee is flagged as having item preknowledge if at least one e_i for a compromised item is statistically significant and negative for a compromised item, similar to how van der Linden and Guo (2008, p. 382) suggested detecting item preknowledge.

Lee and Wollack (2017) and Wang et al. (2018) used a mixture hierarchical IRT model, which is fitted using the Bayesian Markov chain Monte Carlo algorithm (e.g., Gelman, Carlin, Stern, & Rubin, 2003), to determine whether the response time and item score for an item-examinee combination are aberrant. Wang et al. (2018) showed that the approach outperforms the approach of van der Linden and Guo (2008). This approach can be used to detect item preknowledge and does not require the assumption of known compromised items.

Toton and Maynes (2019) suggested an approach to detect item preknowledge that does not require fitting any model to the data. The approach involves a comparison of an examinee's response time on an item to the average response time of all examinees who did not have preknowledge of the item, conditioned on whether the item was answered correctly and incorrectly. This approach is simple, but requires a group of examinees who did not have item preknowledge.

The only frequentist approach that can be used to detect item preknowledge for RTMs is the one suggested by Sinharay (2018). This lack of frequentist approaches is surprising given the existence of several frequentist approaches to assess, for example, item fit (e.g., Glas & van der Linden, 2010; Ranger & Ortner, 2012), fit of the local independence assumption (Glas & van der Linden, 2010), independence of responses and response times (van der Linden & Glas, 2010), and differential item functioning (Glas & van der Linden, 2010) for RTMs. In addition, the existing approaches that can be used to detect item preknowledge based on response times are all designed to detect a variety of aberrant

responses (or, a variety of person misfit) and are expected to have low power for detecting item preknowledge. This expectation is based on the finding by researchers such as Glas and Dagohoy (2007) and Sinharay (2017a) that person-fit statistics based on item scores have much smaller power compared to statistics for detecting item preknowledge based on item scores.

Detection of Item Preknowledge Using Item Scores

Several methods (e.g., Drasgow et al., 1996; McLeod et al., 2003; Shu et al., 2013; Sinharay, 2017a) exist for detecting item preknowledge using only item scores.

Let x_i denote the score of a randomly chosen examinee on item i. Let $\boldsymbol{x}_c$ and $\boldsymbol{x}_{\overline{c}}$ respectively denote the collection of scores of the examinee on the items in c and $\overline{c}$.

For an examinee, let us define the maximum likelihood estimate (MLE) or the weighted maximum likelihood estimate (WLE; Warm, 1989) of the examinee ability from the scores on c as $\hat{\theta}_c$, that from the scores on $\overline{c}$ as $\hat{\theta}_{\overline{c}}$, and that from the scores on all the items as $\hat{\theta}$.

The likelihood ratio test (LRT) statistic (e.g., Finkelman, Weiss, & Kim-Kang, 2010; Guo & Drasgow, 2010) for testing the null hypothesis of equality of the examinee ability over c and $\overline{c}$ is given by

$$\Gamma = 2[\ell(\hat{\theta}_c; x_i, i \in c) + \ell(\hat{\theta}_{\overline{c}}; x_i, i \in \overline{c}) - \ell(\hat{\theta}; x_i, i = 1, 2, \ldots, I)], \tag{6}$$

where

$$\ell(\hat{\theta}_c; x_i, i \in c) = \text{log-likelihood of the scores on } c \text{ at } \hat{\theta}_c,$$

$$\ell(\hat{\theta}_{\overline{c}}; x_i, i \in \overline{c}) = \text{log-likelihood of the scores on } \overline{c} \text{ at } \hat{\theta}_{\overline{c}},$$

$$\text{and } \ell(\hat{\theta}; x_i, i = 1, 2, \ldots, N) = \text{log-likelihood of the scores on all the items at } \hat{\theta}.$$

Letting $P_i(x_i|\hat{\theta}_c)$ denote the likelihood of x_i given $\hat{\theta}_c$, one obtains

$$\ell(\hat{\theta}_c; x_i, i \in c) = \sum_{i \in c} \log P_i(x_i|\hat{\theta}_c).$$

Then the LRT statistic given in Equation 6 can be expressed as

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

Sinharay (2017a) suggested the signed likelihood ratio statistic given by

$$L_s = \begin{cases} \sqrt{\Gamma} \text{ if } \hat{\theta}_c \geq \hat{\theta}_{\bar{c}}, \\ -\sqrt{\Gamma} \text{ if } \hat{\theta}_c < \hat{\theta}_{\bar{c}} \end{cases}$$

for detecting item preknowledge. The statistic L_s has an asymptotic standard normal distribution (e.g., Sinharay, 2017a; Cox, 2006, p. 104) under the null hypothesis of no item preknowledge. A large value of L_s leads to the rejection of the null hypothesis of no item preknowledge.

Sinharay (2017a) and Sinharay (2017b) demonstrated that the L_s statistic performed quite well in comparison to existing statistics in detecting item preknowledge.

Method: A New Statistic Based on Response Times

If some examinees benefited from item preknowledge, it is likely that they would perform faster on the compromised items in comparison to the non-compromised items. Kasli and Zopluoglu (2018) and Toton and Maynes (2019) analyzed real data sets involving item compromise and found that those with item preknowledge answered the compromised items faster than the rest. Consequently, the speed parameter (τ) of the examinees with item preknowledge would not be equal to their original speed parameters, but would be larger on average than the latter on the compromised items. This phenomenon is very similar to item preknowledge leading to examinee-ability estimates being larger on the compromised items than on non-compromised items (e.g., Sinharay, 2017a). Thus, it is possible to determine whether examinees benefited from item preknowledge by examining whether their speed parameters are larger on the compromised items than on the non-compromised items. Let τ_c and $\tau_{\bar{c}}$ respectively denote an examinee's true speed parameters on the compromised and non-compromised items, respectively and let $\hat{\tau}_c$ and $\hat{\tau}_{\bar{c}}$ 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.

One way to detect item preknowledge using RTMs is to test the null hypothesis $H_0 : \tau_c = \tau_{\bar{c}}$ versus the alternative hypothesis $H_1 : \tau_c > \tau_{\bar{c}}$. It is reasonable to test this

hypothesis using the likelihood ratio test (e.g., Cox & Hinkley, 1974; Lehmann & Romano, 2005; Rao, 1973) or LRT given the satisfactory performance of LRTs in a wide variety of hypothesis testing problems (e.g., Casella & Berger, 2002, p. 374). The LRT statistic for testing $H_0 : \tau_c = \tau_{\overline{c}}$ versus the alternative hypothesis $H_1' : \tau_c \neq \tau_{\overline{c}}$ is given by

$$\Lambda = 2[\ell(\hat{\tau}_c; y_i, i \in c) + \ell(\hat{\tau}_{\overline{c}}; y_i, i \in \overline{c}) - \ell(\hat{\tau}; y_i, i = 1, 2, \ldots, I)], \tag{7}$$

where, for example,

$\ell(\hat{\tau}_c; y_i, i \in c) =$ log-likelihood of the log-response times on the items in c, computed at $\hat{\tau}_c$.

For the LNMRT (van der Linden, 2006), one can express $\ell(\hat{\tau}_c; y_i, i \in c)$ as

$$\begin{aligned}
\ell(\hat{\tau}_c; y_i, i \in c) &= \sum_{i \in c} \left[-\frac{1}{2}\log(2\pi) + \log(\alpha_i) - \frac{\alpha_i^2}{2}(y_i - \beta_i + \hat{\tau}_c)^2 \right] \\
&= \sum_{i \in c} \left[-\frac{1}{2}\log[2\pi] + \log(\alpha_i) - \frac{\alpha_i^2}{2}\hat{\tau}_c^2 - \frac{\alpha_i^2}{2}(y_i - \beta_i)^2 + \hat{\tau}_c \alpha_i^2 (y_i - \beta_i) \right] \\
&= \sum_{i \in c} \left[-\frac{1}{2}\log[2\pi] + \log(\alpha_i) \right] - \hat{\tau}_c^2 \sum_{i \in c} \frac{\alpha_i^2}{2} - \sum_{i \in c} \frac{\alpha_i^2}{2}(y_i - \beta_i)^2 + \hat{\tau}_c \sum_{i \in c} \alpha_i^2 (y_i - \beta_i) \\
&= \sum_{i \in c} \left[-\frac{1}{2}\log[2\pi] + \log(\alpha_i) \right] - \hat{\tau}_c^2 \sum_{i \in c} \frac{\alpha_i^2}{2} - \sum_{i \in c} \frac{\alpha_i^2}{2}(y_i - \beta_i)^2 + \hat{\tau}_c^2 \sum_{i \in c} \alpha_i^2 \\
&= \sum_{i \in c} \left[-\frac{1}{2}\log[2\pi] + \log(\alpha_i) \right] + \hat{\tau}_c^2 \sum_{i \in c} \frac{\alpha_i^2}{2} - \sum_{i \in c} \frac{\alpha_i^2}{2}(y_i - \beta_i)^2, \qquad (8)
\end{aligned}$$

where the penultimate equality holds because of Equation 2.

Then one obtains from Equation 7 that

$$\Lambda = \hat{\tau}_c^2 \sum_{i \in c} \alpha_i^2 + \hat{\tau}_{\overline{c}}^2 \sum_{i \in \overline{c}} \alpha_i^2 - \hat{\tau}^2 \sum_{i=1}^{I} \alpha_i^2. \tag{9}$$

To test $H_0 : \tau_c = \tau_{\overline{c}}$ versus $H_1 : \tau_c > \tau_{\overline{c}}$, one can use the signed likelihood ratio statistic that is given by

$$\Lambda_s = \begin{cases} \sqrt{\Lambda} \text{ if } \hat{\tau}_c \geq \hat{\tau}_{\overline{c}}, \\ -\sqrt{\Lambda} \text{ if } \hat{\tau}_c < \hat{\tau}_{\overline{c}} \end{cases} \tag{10}$$

(Cox, 2006; Sinharay, 2017a). The appendix includes an R (R Core Team, 2019) function for computing Λ_s from a data set.

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

$$Z = \frac{\hat{\tau}_c - \hat{\tau}_{\overline{c}}}{\sqrt{\mathrm{Var}(\hat{\tau}_c) + \mathrm{Var}(\hat{\tau}_{\overline{c}})}} = \frac{\hat{\tau}_c - \hat{\tau}_{\overline{c}}}{\sqrt{[\sum_{i \in c} \alpha_i^2]^{-1} + [\sum_{i \in \overline{c}} \alpha_i^2]^{-1}}}.$$

Noting that the log-likelihood of the response times of a person is quadratic in the speed parameter (see Equation 8), the equality of Λ_s and Z agrees with the result of Buse (1982) that for quadratic log-likelihoods, the Wald test and the LRT are identical.

The statistic Λ_s follows the standard normal distribution for large c and $\overline{c}$ under the null hypothesis of no item preknowledge (e.g., Cox, 2006, p. 104). In this case, it is possible to obtain a distributional result that is more general. Because Λ_s is identical to Z in this case, and because Z is a linear combination of normal random variables (see Equation 2) divided by its standard deviation (Equation 3) under the null hypothesis, Z and hence Λ_s follows the standard normal distribution under the null hypothesis even when the test is not long.

Simulation Based on Real Data

It is not known whether any RTM perfectly reflects reality or fits real data adequately. For example, even though the LNMRT is quite popular, researchers such as Bolsinova and Tijmstra (2018) and Ranger (2013) pointed to some limitations of the model. Therefore, to examine the properties of Λ_s, simulations based on real data were used rather than simulations based on data generated from a RTM. For comparison purposes, the properties of χ_{pf} and the Bayesian residuals of van der Linden and Guo (2008) were also examined.

Simulation Design

The starting point of this examination was a real data set that consisted of the response times of more than 18,000 test takers on a computerized test for English proficiency. The test includes 34 operational items that are all multiple-choice. The mean response times on the operational items were between 21 and 52 seconds and the mean per-item response

times of the persons on the operational items were between 9 and 53 seconds. There were no evidence of item compromise or item preknowledge for the test. The item parameters of the LNMRT were estimated once from the whole data set (of 18,000 test takers) using the R package *lavaan* (Rosseel, 2012) and then these estimates were used in the next steps of the study. The R codes for estimating the item parameters for the data set using the *lavaan* package are included in the appendix. The item fit statistic for the LNMRT of Glas and van der Linden (2010) was statistically significant for 3 out of 34 items, or 8.8% items at 5% level, which indicates that the LNMRT shows some misfit, but is not too unreasonable for these data.

Then the following steps were performed 100 times for different choices of the size of the set of compromised items, c (2, 4, 7, or 10 items), and a quantity d (with values 0, 1, 2, or 3) that determines the speed of those with preknowledge on the compromised items:

1. Randomly select 10,000 examinees from the original data set.

2. From the 10,000 examinees, randomly identify 1,000 examinees who would be treated as those with item preknowledge; the remaining 9,000 examinees would be treated as not having item preknowledge.

3. Randomly choose the items that would constitute c (that is, from the 34 items in the data set, choose the 2, 4, 7, or 10 items that would be treated as compromised).

4. For each item in c and each examinee with item preknowledge, reset the logarithm of the response time to be its actual value minus d times the standard deviation (over all examinees) of the logarithm of the response times for the item. This step artificially creates a data set with item preknowledge.

5. Compute the MLEs of the person speed parameters (τ_c, $\tau_{\bar{c}}$, and τ) from the (changed) data set.

6. Compute the Λ_s and χ_{pf} statistics and the Bayesian residuals for all the examinees in the (changed) data set using the person parameter estimates computed in the previous

step and using Equations 10, 4, and 5 above and Equations 14 and 15 of van der Linden (2006).

Note that when d is 0, the response times are actually not changed in Step 4 and the statistics are computed from data sets that actually do not include any preknowledge. The simulations for these cases allowed us to approximate the Type I error rate of the statistics as the proportion of all examinees that had a significant value of the statistic. The simulations for the cases with $d > 0$ allowed us to approximate the power of the statistics as the proportion of examinees with item preknowledge that had a significant value of the statistic.

Results from the Simulation

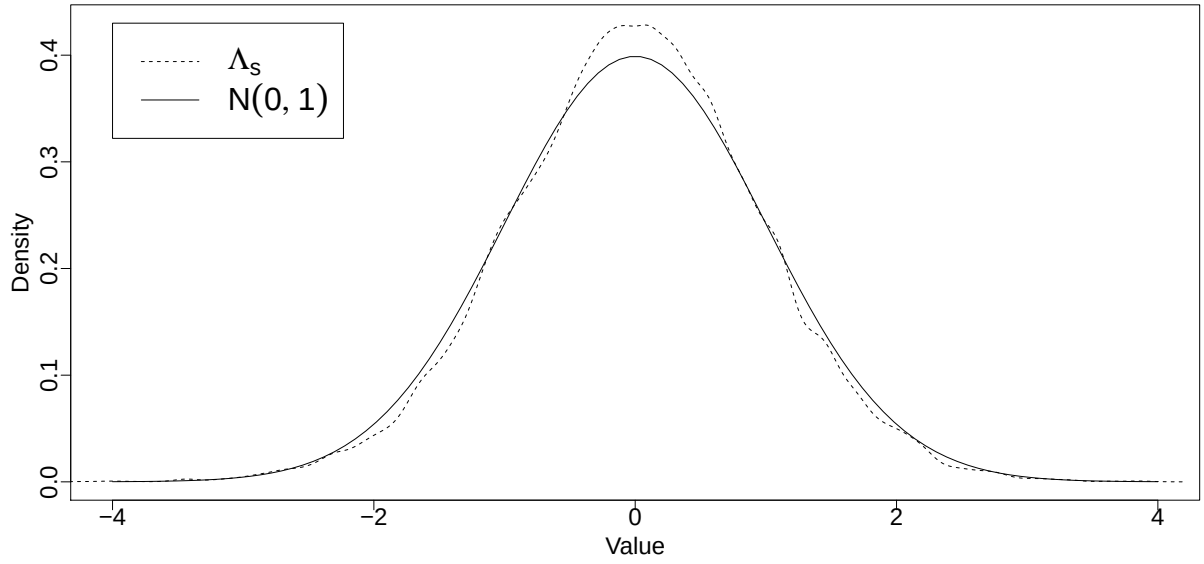

Figure 1: The kernel-density estimate of the distribution of Λ_s for the case of 10 compromised items.

Figure 1 shows (using a dashed line) the kernel-density estimate[2] of the distribution

[2]The figure was created using the function "density" in the R software (R Core Team, 2019).

of the values of Λ_s for the simulation case of d=0 for the case of 10 compromised items. The theorized standard normal null distribution is also shown (using a solid line) in the figure for convenience. The distribution of the values of Λ_s is very close, especially at the right tail, to the corresponding theorized null distribution. Thus, the standard normal null distribution of Λ_s seems to adequately hold for data that involve no preknowledge.

Table 1: The Type I Error Rates at 1% Level.

Statistic	2 items	4 items	7 items	10 items
χ_{pf}	0.063	0.063	0.063	0.062
Bayesian residuals	0.017	0.028	0.048	0.055
Λ_s	0.005	0.006	0.009	0.008

Table 1 shows the Type I error rates at 1% level of χ_{pf}, Bayesian residuals, and Λ_s for different numbers of compromised items. Wollack, Cohen, and Eckerly (2015) commented that methods for detection of test fraud are typically applied with conservative levels—that is why results are reported for 1% level rather than the customary 5% level. The Type I error rates of χ_{pf} are considerably larger than the nominal level. Presumably, this is due to a general misfit of the LNMRT to the data as well as the presence of person misfit other than item preknowledge in the data set. The Type I error rates of the Bayesian residuals are also inflated. On the contrary, the Type I error rates of Λ_s are always smaller than the nominal level, which provides favorable evidence for Λ_s given that the data used to compute these rates are not simulated, but real data; the rates become closer to the nominal level as the number of compromised items increases.

Figure 2 shows the power at 1% level of Λ_s, Bayesian residuals, and χ_{pf}, to detect item preknowledge for different combinations of values of number of items compromised and d. The four panels of the figure show the power of the statistics when the number of compromised items (shown in the title of each panel) is 2, 4, 7, and 10, respectively. In each panel, the value of d is shown along the X-axis and the power is shown along the Y-axis. The power for Λ_s, Bayesian residuals, and χ_{pf} are shown using hollow circles, hollow triangles, and plus signs respectively, joined by a solid line.

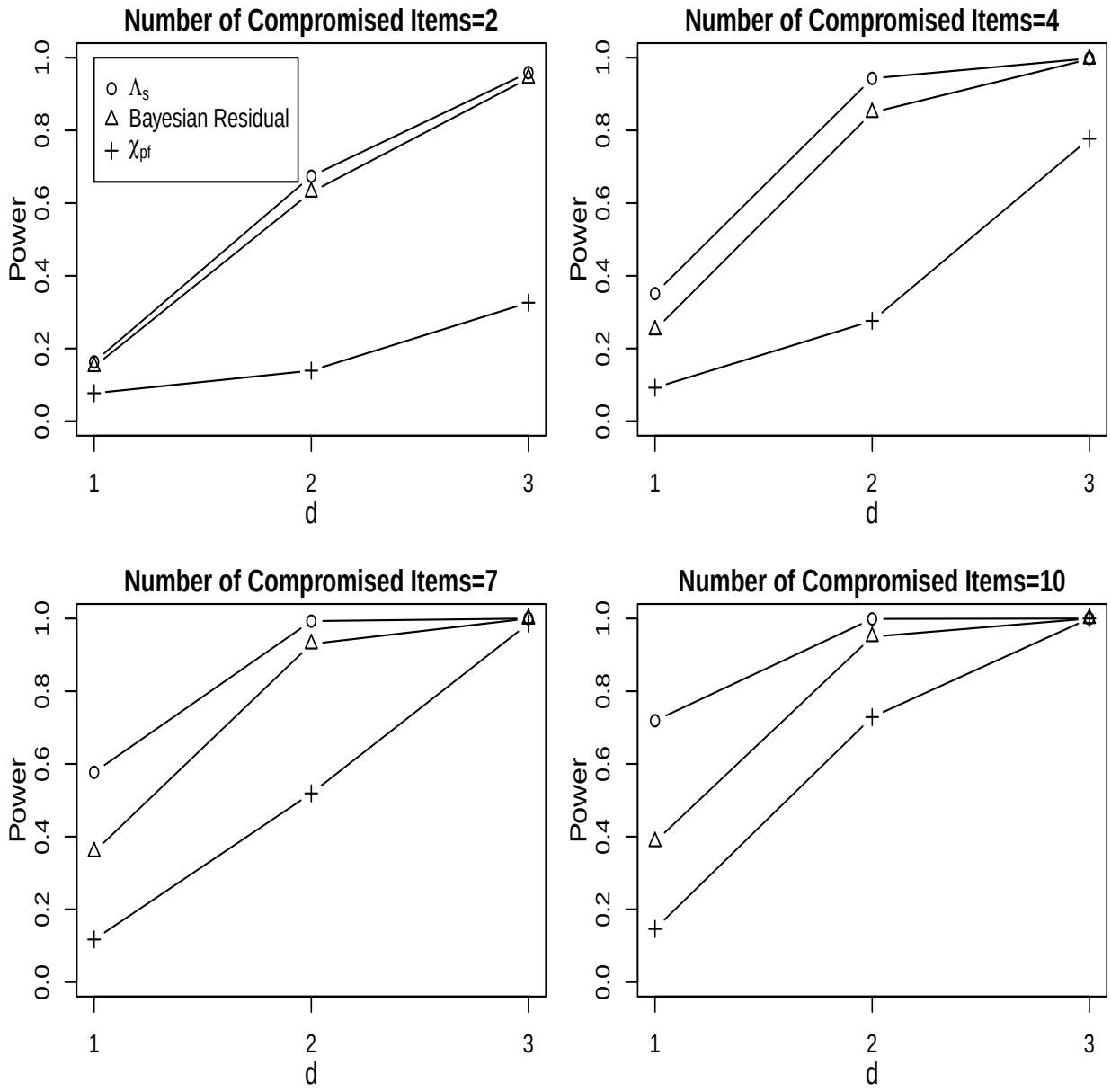

Figure 2: The power of the statistics at 1% level.

Figure 2 shows that the power of Λ_s is considerably larger than that of χ_{pf} and the Bayesian residuals. The smaller power of χ_{pf} is expected given the common knowledge that person-fit statistics have small power against specific alternatives (e.g., Glas & Dagohoy, 2007; Sinharay, 2017a). The figure also shows that the power of each statistic increases as the number of compromised items increases and as d increases. The power of Λ_s is larger than 0.8 when d is 2 or 3 and the number of compromised items is 4 or larger.

Real Data Example

Let us consider data from two forms of a non-adaptive licensure test. The data sets (or other data sets similar to these two) were analyzed in several chapters of Cizek and Wollack (2017) and also in Fox and Marianti (2017) and Sinharay (2017a). Each test form includes 170 operational items. Item scores and response times were available for 1,624 and 1,629 examinees, respectively, for Forms 1 and 2. The licensure organization who provided the data identified as compromised 63 and 61 items, respectively, on the forms. The organization also flagged 41 and 42 examinees (among the above-mentioned 1,624 and 1,629 examinees), respectively, as possible cheaters from a variety of statistical analysis and a rigorous investigative process that brought in other information; given the rigor of the investigative process, these examinees will be treated as truly aberrant.

The LNMRT was fitted to the data sets (and its item parameters estimated) using the R package *lavaan* (Rosseel, 2012). The item fit statistic of Glas and van der Linden (2010) was statistically significant for 7.6% items at 5% level, which indicates that the LNMRT fits these data not too poorly although there is some evidence of misfit of the model. Then the values of χ_{pf} (Marianti et al., 2014), Bayesian Residuals (van der Linden & Guo, 2008), and Λ_s were computed from the two data sets. In addition, because the item scores were available for the data sets, the unidimensional two-parameter logistic model was fitted to the item scores using the R package mirt (Chalmers, 2012) and the L_s statistic (Sinharay, 2017a) was computed for all the examinees.

Table 2: The Percent of Examinees for Whom χ_{pf}, Bayesian Residuals, Λ_s, and L_s Were Significant for the Real Data.

Level	Form 1								Form 2							
	Not Flagged				Flagged				Not Flagged				Flagged			
	χ_{pf}	e_i	Λ_s	L_s	χ_{pf}	e_i	Λ_s	L_s	χ_{pf}	e_i	Λ_s	L_s	χ_{pf}	e_i	Λ_s	L_s
0.1%	10	2	1	0	27	5	22	10	10	2	2	1	17	6	19	14
1%	15	2	3	3	29	10	29	19	14	1	3	3	17	7	21	19

The rounded percentages of examinees for whom χ_{pf}, the Bayesian Residual (e_i), L_s, and Λ_s were significant at significance levels of 0.1% and 1% for the two forms are provided

in Table 2. For each form, the first four columns include the percents significant among the examinees who were not flagged by the licensure organization and the last four columns include the percents significant only among the 41 or 42 examinees who were flagged as possible cheaters by the licensure organization; thus, for example, the percent 27 in sixth column of the first row of numbers denotes that among the 41 examinees flagged by the licensure organization, χ_{pf} was significant at 0.1% level for 11 examinees (note that $11/41 \approx 0.27$).

Table 2 shows that the values of percent significant among the non-flagged examinees for χ_{pf} are much larger than those for the other two statistics and also larger than the significance level. This finding is in agreement with the inflated Type I error rate of χ_{pf} in the simulation studies described earlier.

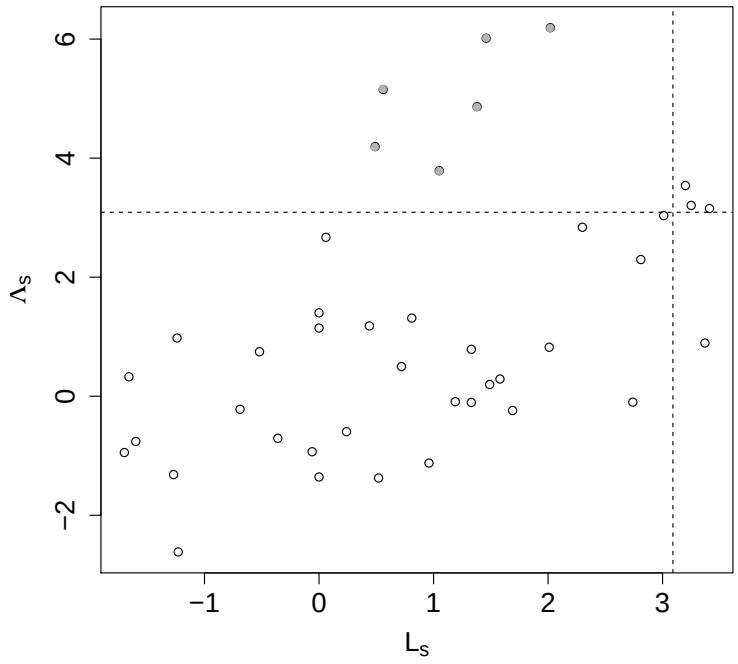

Figure 3: A scatter-plot of Λ_s versus L_s for the 41 flagged examinees for Form 1.

In Table 2, the percents of significant values for Λ_s are close to those for L_s and the Bayesian residuals for the non-flagged examinees, but are considerably larger than those for L_s and the Bayesian residuals for the flagged examinees. Thus, Λ_s seems to provide useful

information that is not provided by other existing statistics for the data.

Further insight is provided by Figure 3 that shows the values of the Λ_s statistic (along the Y-axis) versus those of the L_s statistic (along the X-axis) for the 41 examinees who were flagged by the licensure organization for Form 1. Each circle (either hollow or filled in gray) in the figure shows the combination of values of L_s and Λ_s for one examinee. For example, the topmost circle (filled in gray) in the plot corresponds to an examinee for whom L_s and Λ_s are 1.99 and 6.13, respectively. Horizontal and vertical dashed lines are shown at the 99.9th percentile of the standard normal distribution; any value larger than this quantile is statistically significant at 0.1% significance level. The figure shows that for the flagged examinees, the two statistics are positively correlated (the correlation coefficient is 0.48)[3], indicating that among the flagged examinees, those who performed better on the compromised items were also faster on those items in general. The value of Λ_s is significant at 0.1% level for nine examinees (those corresponding to the points above the horizontal dashed line). Interestingly, each of these nine examinees performed better on the compromised items than on the non-compromised items, which is evident from L_s being positive for all of them. Also, L_s is not significant at 0.1% level for six of these nine examinees (corresponding to the six circles filled in gray)—so Λ_s provides additional evidence (over and above L_s) of item preknowledge for these six examinees. The fact that only one among Λ_s and L_s is significant for a few flagged examinees (note the one flagged examinee for whom L_s is significant while Λ_s is not) indicates that each of Λ_s and L_s provides some unique information regarding item preknowledge—so using both of them may be a prudent strategy in investigations of item preknowledge.

In Table 2, the percents significant for each statistic are much larger among the examinees flagged by the licensure organization than among those not flagged—this result provides some evidence that the statistics are somewhat successful—they are significant at a larger rate among the examinees who are truly aberrant. Note that item compromise was not the only reason of flagging by the licensure organization; for example, researchers

[3]The two statistics are positively correlated for the non-flagged examinees (correlation=0.20) and whole sample (correlation=0.28) as well.

such as Zopluoglu (2017) found the values of answer-copying statistics to be statistically significant for some of these flagged examinees. Therefore, percents considerably smaller than 100 for the flagged examinees in Table 2 is not a severe limitation of the statistics.

Conclusions and Recommendations

This paper suggested a frequentist approach to detect item preknowledge based on response times. The distribution of the suggested statistic under the null hypothesis is proved to be a standard normal distribution irrespective of the test length and the number of compromised items. Simulations based on real data show that the Type I error rate of the new statistic is close to the nominal level and the power of the statistic is larger than that of existing statistics. An encouraging aspect of the new statistic is that the statistic appears to have satisfactory power in several cases even for the conservative significance level of 1% (see Figure 2). The new statistic can be calculated very easily, as is clear from the computer code that is provided—so the statistic may become useful to those interested in detection of test fraud.

Though the new statistic seems promising in detecting test fraud, it should not be used as a sole measure to detect test fraud. Experts such as van der Linden and Guo (2008) suggested using statistics based on response times to detect aberrant examinee behavior as a part of quality control and the new statistic can be used in the same manner. van der Linden and Guo (2008) also warned against the mechanical use of statistics based on response times in high-stakes contexts such as detection of cheating because of the presence of false alarms of these statistics. A wise strategy in high-stakes contexts would involve the use of the new statistic and/or other statistics for detection of test fraud as secondary evidence, as recommended by experts such as Hanson, Harris, and Brennan (1987).

The statistic Λ_s can only be applied when only 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. In addition, Λ_s can only be applied when the set of compromised items is known; researchers such as

Drasgow et al. (1996), Sinharay (2017a), Shu et al. (2013), and van der Linden and Guo (2008) considered this case.[4] 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). Cizek and Wollack (2017, p. 14) and Eckerly, Smith, and Lee (2018) described real data sets where the set of compromised items was known. The case of known compromised items may also arise when the test administrators have applied a method for detection of compromised items (e.g., that suggested by Veerkamp & Glas, 2000) to flag several items that may have been compromised. In cases when the set of compromised items is not precisely known, Λ_s 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 compromised by Smith and Davis-Becker (2011)—the old and new items would respectively play the roles of the compromised and non-compromised items.

Item parameters were assumed known (and estimated from a previous calibration) in the derivation of the distribution of the new statistic and in the simulation and no adjustment is made to the distribution of Λ_s to account for the uncertainty in the estimates of the item parameters. This assumption is common in various person-level analysis such as erasure analysis (e.g., Wollack et al., 2015), person-fit analysis (e.g., Snijders, 2001), and detection of item preknowledge using item scores (e.g., Sinharay, 2017a). In addition, this assumption of known item parameters is reasonable in several contexts such as in computerized adaptive testing where item parameters are assumed known (and estimated from a previous calibration) and in cases where the proportion of examinees with item preknowledge is small. However, if the proportion of examinees with item preknowledge is large, then the assumption may lead to undesirable consequences regarding detection of 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 if a large number of examinees

[4]McLeod et al. (2003) and Wang et al. (2018) considered the case when the set of compromised items is unknown.

have preknowledge of those items because they would answer those items faster. As a consequence, the speed-parameter estimate based on the compromised items ($\hat{\tau}_c$) would be substantially underestimated for those with preknowledge and without preknowledge—this underestimation would lead to smaller power and a false alarm rate that is smaller than nominal level. This phenomenon was verified in an additional set of simulations in which the item parameters were estimated between the fourth and fifth steps of the above simulation. In these additional simulations,[5] the comparative performance of the statistics was very similar to those reported in this paper, but the false alarm rate of Λ_s was smaller than the nominal level and the power of the statistic was smaller than those reported in Figure 2. One possible solution in the face of a severe extent of item preknowledge involves the four-step purification process of (a) estimating item parameter from the full sample, (b) computing Λ_s 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 Λ_s, and (d) computing Λ_s for the full sample using the item-parameter estimates computed in the previous step. Such 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 percent of examinees benefiting from item preknowledge is very large (say, larger than 50%), then even a purification would not work well and retesting all examinees would be the only reasonable choice. However, tests for which a large proportion of examinees benefited from item preknowledge are very rare, if not unheard of. The effect of the assumption of known item parameters on the properties of the new statistic and new approaches for accounting for the uncertainty of the item parameters in the distribution of the new statistic may be explored in future research.

The statistic Λ_s is expected to have small power when the number of compromised items is small because the estimate of the examinee speed parameter for the compromised items ($\hat{\tau}_c$) would have a large variance in this case. Overall, Table 3 reflects a rough guideline about the performance of Λ_s for different percentages of items that are compromised and

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

different percentages of examinees who have preknowledge for tests in which the item parameters are estimated from the examinee sample. The table shows that the statistic

Table 3: A Rough Guide to the Application of Λ_s.

% Items	% of Examinees with Preknowledge		
Compromised	Small	Moderately large	Large
Small	Low power	Low power	Unreliable result
Moderately large	Large power	Low power	Unreliable result
Large	Low power	Unreliable result	Unreliable result

should have best performance in the form of large power when the percent of examinees with preknowledge is small and the percent of items that are compromised is moderately large. In four cases, the statistic is expected to have low power due to reasons like too few compromised items leading to inaccurate estimation of τ. In four other cases including three with a large percent of examinees with preknowledge, the statistic would lead to unreliable results and should not be used. If accurate estimates of item parameters are available (for example, on a computerized adaptive test), then the performance of Λ_s would not depend on the percent of examinees with preknowledge and would only depend on the percent of compromised items an examinee answers in a manner shown in the second column of Table 3.

The results on the distribution of Λ_s were derived under the assumption that the LNMRT fits the data adequately. Therefore, one should assess the overall fit of the LNMRT to the data set before applying the Λ_s statistic. If the LNMRT does not fit the data overall (due to, for example, a violation of the local independence assumption), then the null distribution of Λ_s may not be standard normal and the use of the statistic could lead to erroneous conclusions. The percentage of standardized residuals (van der Linden & Guo, 2008) over all examinee-item combinations and the item fit statistic of Glas and van der Linden (2010) may be used to assess the fit of the LNMRT before computing Λ_s for the data. However, the simulations based on the real data showed that the distribution of Λ_s statistic under no item preknowledge was close to the standard normal distribution for a real data set even though the LNMRT showed a small extent of misfit to the data—this

result shows that Λ_s may be robust to model misfit that is typically observed for real data.

This paper has several limitations and, consequently, leaves plenty of room for future research. First, the suggested statistic should be computed for more simulated and real data sets. Second, only the LNMRT was considered in this paper—extension of the suggested statistic to other types of RTMs would be a potential area of future research. It is anticipated that for other RTMs, the suggested statistic would have a standard-normal distribution under the null hypothesis only for long tests because of the central limit theorem. Third, though the simulation study provided some evidence that the new statistic is robust to misfit of the LNMRT, it is possible to further examine the consequences of misfit of the LNMRT on the properties of the new statistic in future research. Fourth, this paper only deals with the case of a known set of compromised items. Extension of the suggested statistic to the case of unknown compromised items is a potential area for further research. Finally, extension of the suggested approach to detect item preknowledge using both response times and item scores would be a possible area for further research. Sinharay and Johnson (2019) made some progress along this line of research.

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis, 2nd edition.* New York, NY: Wiley.

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*, 13–38.

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.

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician*, *36*(3), 153–157.

Casella, G., & Berger, R. L. (2002). *Statistical inference (2nd edition).* Pacific Grove, CA: Duxbury.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

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.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London, UK: Chapman and Hall.

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)

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.

Finger, M. S., & Chee, C. S. (2009, April). *Response-time model estimation via confirmatory factor analysis.* Paper presented at the Annual meeting of the National Council of Measurement in Education, San Diego, CA.

Finkelman, M., Weiss, D. J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the adaptive measurement of change. *Applied Psychological Measurement*, *34*, 238–254.

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)

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. New York, NY: Chapman and Hall.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. New York, NY: Chapman and Hall.

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for

polytomous items. *Psychometrika*, *72*, 159–180.

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)

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*, 351–364.

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.

Kyle, T. (2002). *Cheating scandal rocks GRE, ETS.* http://thedartmouth.com/2002/08/09/cheating-scandal-rocks-gre-ets/. (Retrieved May 12 2016)

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.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). New York, NY: Springer-Verlag.

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)

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.

Ranger, J. (2013). A note on the hierarchical model for responses and response times in tests of van der Linden (2007). *Psychometrika*, *78*, 538–544.

Ranger, J., & Ortner, T. (2012). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling*, *54*, 128–148.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: John Wiley.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

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

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. (2018). A new person-fit statistic for the lognormal model for response times. *Journal of Educational Measurement*, *55*, 457–476.

Sinharay, S., & Johnson, M. S. (2019). 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*. (Advance online publication. doi:10.1111/bmsp.12187)

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.

Snijders, T. (2001). Asymptotic distribution of person-fit statistics with estimated person parameter. *Psychometrika*, *66*, 331–342.

Toton, S. L., & Maynes, D. D. (2019). Detecting examinees with pre-knowledge in experimental data using conditional scaling of response times. *Frontiers in Education*, *4*.

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

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

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247–272.

van der Linden, W. J. (2016). Lognormal response-time model. In W. van der Linden (Ed.), *Handbook of item response theory, Volume 1. Models.* Boca Raton, FL: Chapman and Hall/CRC.

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*, 120–139.

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

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.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory.

Psychometrika, *54*, 427–450.

Wollack, J. A., Cohen, A. S., & Eckerly, C. A. (2015). Detecting test tampering using item response theory. *Educational and Psychological Measurement*, *75*, 931–953.

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: R Code to Estimate Item Parameters and Compute the New Statistic

```
# R Subroutine to estimate item parameters of the LNMRT; ltimes is the matrix of
# log-response times
library(lavaan)
ly <- data.frame(ltimes)# ltimes (of dimension nxI) includes log-response times
I=34#I, the number items, is 34 for the data set in the Simulation Study
model=paste("f1=~",paste0("a*X",1:(I-1),"+",collapse=""),paste("a*X",I,sep=""))
fit <- cfa(model, data = ly, meanstructure = TRUE, auto.var= TRUE)
pars <- coef(fit)#pars[35:69], sqrt(1/pars[1:34]), and pars[35] include
#                 estimated beta's, alpha's, and sigma_squared

# R Subroutine to compute the new statistic; ltimes is the matrix of
# log-response times, comp is the set of compromised items, alpha and
# beta are item parameters of the log-normal response time model
Lambdas=function(ltimes,comp,alpha,beta){
ncomp=setdiff(1:ncol(ltimes),comp)
tcomp=PPest(alpha[comp],beta[comp],ltimes[,comp])
tncomp=PPest(alpha[ncomp],beta[ncomp],ltimes[,ncomp])
tall=PPest(alpha,beta,ltimes)
return((tcomp-tncomp)/sqrt(1/sum((alpha[comp])^2) + 1/sum((alpha[ncomp])^2)))}
# PPest is the subroutine to compute estimated person parameters
PPest=function(alpha,beta,ltimes)
{tauhat=rep(sum(alpha*alpha*beta)/sum(alpha*alpha),nrow(ltimes))
   -ltimes%*%(alpha*alpha)/(sum(alpha*alpha))
 return(tauhat)}
```