Detecting Test Fraud Using Bayes Factors

Sandip Sinharay and Matthew S. Johnson,
Educational Testing Service

An Updated Version of this document appeared in Behaviormetrika. The website for the article is https://link.springer.com/article/10.1007%2Fs41237-020-00113-9

The citation for the article is: Sinharay, S., & Johnson, M. S. (2020). Detecting test fraud using Bayes factors. Behaviormetrika. Advance Online Publication. doi: 10.1007%2Fs41237-020-00113-9

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.

Detecting Test Fraud Using Bayes Factors

Sandip Sinharay & Matthew S. Johnson, Educational Testing Service

September 20, 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.

Detecting Test Fraud Using Bayes Factors

Abstract

According to Wollack and Schoenig (2018), score differencing is one of six types of statistical methods used to detect test fraud. In this paper, we suggested the use of Bayes factors (e.g., Kass & Raftery, 1995) for score differencing. A simulation study shows that the suggested approach performs slightly better than an existing frequentist approach. We also demonstrate the usefulness of the suggested approach using a real data example.

Key words: Likelihood ratio statistic, marginal likelihood, score differencing.

Acknowledgments

The authors wish to express sincere appreciation and gratitude to Wim van der Linden and Kazuo Shigemasu, the editors. The authors thank Sooyeon Kim, Carol Eckerly, and Daniel McCaffrey for their helpful comments on an earlier version. Any opinions expressed in this publication are those of the authors and not necessarily of ETS or of Institute of Education Sciences. The research was suppored by the Institute of Education Sciences, US Department of Education, through Grant R305D170026.

Producers and consumers of test scores are increasingly concerned about fraudulent behavior before and during the test. Such behavior is more likely to be observed when the stakes are high, such as in licensing, admissions, and certification testing (van der Linden, 2009). Cheating incidents such as educator cheating in Atlanta public schools (e.g., Kingston, 2013) and "cram schools" selling items on SATs (e.g., Strauss, 2014) recently made headline news. 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 among others that testing programs with high-stakes consequences should have defined procedures for detecting potential testing irregularities.

Naturally, there is a growing interest in statistical/psychometric methods for detecting fraudulent behavior on tests (e.g., Cizek & Wollack, 2017). Wollack and Schoenig (2018) categorized the statistical methods to detect test fraud/cheating into six categories. One of these six categories is "score differencing"—this category of methods essentially involves a test of the hypothesis of equal ability of an examinee over two sets of items S_1 and S_2 against the alternative hypothesis that the examinee's performance is better on one of these item sets. Score differencing can be performed to detect several types of test fraud including fraudulent erasures (e.g., Sinharay, Duong, & Wood, 2017), fraudulent and large gain scores (e.g., Fischer, 2003), and item preknowledge (e.g., Sinharay, 2017a, 2017b; Sinharay & Jensen, 2019).[1]

The existing methods for score differencing are mostly frequentist methods and the inferences from these methods are based on frequentist p-values. The use of these p-values may lead to a large proportion of false positives; Skorupski and Wainer (2017) provided an example where a statistic with a Type I error rate of 0.01 and power of 0.99 is expected to flag 1,386 examinees in a population of 70,000 examinees that includes 1% cheaters, but half of the flags are false positives. In addition, researchers such as van der Linden and Lewis (2015), Allen and Ghattas (2016), and Skorupski and Wainer (2017) argued

[1] Note that the term "score differencing" was used in only one of these references. However, the methods suggested in these references are various versions of "score differencing."

that a frequentist p-value corresponding to a statistic for detecting test fraud is an answer to the question "What is the probability of a significant value of the test statistic given that the examinee did not commit fraud?" that is not the question of interest in the context of detecting test fraud. Consequently, van der Linden and Lewis (2015), Allen and Ghattas (2016), Sinharay (2018), and Skorupski and Wainer (2017) encouraged more applications of Bayesian statistical methods to the detection of test fraud. In addition, a recent statement by American Statistical Association (Wasserstein & Lazar, 2016) included the recommendation that researchers and practitioners should explore Bayesian tools such as Bayes factors as alternatives to frequentist p-values.

However, Bayesian methods have rarely been applied in score differencing, with the exception of Wang, Liu, and Hambleton (2017). The goal of this paper is to suggest a new approach for score differencing using Bayesian methods.

Background: Score Differencing

Consider a test with I items each of which is dichotomously scored.[2] Let us assume that one is interested in score differencing, that is, in testing the equality of the performance on item sets S_1 and S_2 for an examinee whose true overall ability is θ. The sets S_1 and S_2 are non-overlapping and together constitute all items on the test. Let the true ability of the examinee on S_1 and S_2 respectively be denoted as θ_1 and θ_2. Typically, in score differencing, the null hypothesis is $\theta_1 = \theta_2$ and the alternative hypothesis is that the performance on one item set is better than that on the other due to reasons such as test fraud. Let us assume, without loss of generality, that the alternative hypothesis is that the performance on S_2 is better than that on S_1 for the examinee, or, in other words, that θ_2 is larger than θ_1.

Let the scores for the examinee on the I items be denoted by $X_1, X_2, ...X_I$. Let $\boldsymbol{X}$ denote the collection of scores for the examinee on all the items on the test. Let $\boldsymbol{X}_1 = \{X_i, i \in S_1\}$ and $\boldsymbol{X}_2 = \{X_i, i \in S_2\}$ respectively denote the collection of the scores of

[2] Although we deal with tests that include only dichotomous items, our suggested approach extends in a straightforward manner to tests that include polytomous items.

the examinee on the items in Sets 1 and 2. Let

$$P_i(\theta) = P(X_i = 1|\theta)$$

denote the probability of a correct answer on item i for an examinee with true ability θ. For example, for the 2-parameter logistic model (2PLM),

$$P_i(\theta) = \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]},$$

where a_i's and b_i's respectively are the slope and difficulty parameters of the items.

The likelihood of the examinee, denoted as $L(\theta; \boldsymbol{X})$, can be computed as

$$L(\theta; \boldsymbol{X}) = \prod_{i=1}^{I} P_i(\theta)^{X_i} (1 - P_i(\theta))^{1-X_i}. \tag{1}$$

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 item-set S_1 as $\hat{\theta}_1$, that from the scores on S_2 as $\hat{\theta}_2$, and that from the scores on all the items as $\hat{\theta}$.

Let us denote the log-likelihood for the examinee as $l(\theta; \boldsymbol{X})$, that is,

$$l(\theta; \boldsymbol{X}) = \log(L(\theta; \boldsymbol{X})).$$

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 S_1 and S_2 is given by

$$\begin{aligned}
\Lambda &= 2\left[l(\hat{\theta}_1; \boldsymbol{X}_1) + l(\hat{\theta}_2; \boldsymbol{X}_2) - l(\hat{\theta}; \boldsymbol{X})\right] \\
&= 2\sum_{i \in S_1} X_i \log \frac{P_i(\hat{\theta}_1)(1 - P_i(\hat{\theta}))}{P_i(\hat{\theta})(1 - P_i(\hat{\theta}_1))} + 2\sum_{i \in S_2} X_i \log \frac{P_i(\hat{\theta}_2)(1 - P_i(\hat{\theta}))}{P_i(\hat{\theta})(1 - P_i(\hat{\theta}_2))} \\
&\quad +2\sum_{i \in S_1} \log \frac{1 - P_i(\hat{\theta}_1)}{1 - P_i(\hat{\theta})} + 2\sum_{i \in S_2} \log \frac{1 - P_i(\hat{\theta}_2)}{1 - P_i(\hat{\theta})}.
\end{aligned} \tag{2}$$

To test the null hypothesis of equality of the examinee ability over S_1 and S_2 versus the alternative hypothesis that ability over S_2 is larger than that based on S_1, Sinharay (2017a) suggested the signed likelihood ratio (SLR) statistic given by

$$L_s = \begin{cases} \sqrt{\Lambda} \text{ if } \hat{\theta}_2 \geq \hat{\theta}_1, \\ -\sqrt{\Lambda} \text{ if } \hat{\theta}_2 < \hat{\theta}_1. \end{cases} \tag{3}$$

The statistic L_s has an asymptotic standard normal distribution (e.g., Sinharay, 2017a; Cox, 2006, p. 104) under the null hypothesis. A large value of L_s leads to the rejection of the null hypothesis of no difference in performance over S_1 and S_2. Researchers such as Sinharay (2017a), Sinharay (2017b), Sinharay and Jensen (2019), and Wang, Liu, Robin, and Guo (2019) found the Type I error rate and power of L_s to be quite satisfactory in comparison with those of the existing frequentist procedures for score differencing—so L_s will be used as the only frequentist procedure for score differencing in this paper.

As demonstrated by several researchers (e.g., Guo & Drasgow, 2010; Sinharay, 2017a; Sinharay & Jensen, 2019), statistics such as the L_s statistic can be used to detect several types of test fraud including fraudulent erasures, fraudulent and large gain scores, and item preknowledge. The item set S_2 in these three contexts would be the set of items with erasures, the set of items administered at the second time point, and the set of compromised items.

Bayes Factor

Definition

The Bayes factor (e.g., Kass & Raftery, 1995) is a Bayesian approach for model comparison. Let $\boldsymbol{y}$ denote the data and $\boldsymbol{\psi}$ denote the model parameters. Let $p(\boldsymbol{y}|\boldsymbol{\psi}, M_1)$ denote the distribution of the data given the parameters of model M_1 and $p(\boldsymbol{\psi}|M_1)$ denote the prior distribution under model M_1. Then, the Bayes factor in favor of model M_2 in comparison to M_1 is given by

$$BF_{21} = \frac{p(\boldsymbol{y}|M_2)}{p(\boldsymbol{y}|M_1)}, \tag{4}$$

where $p(\boldsymbol{y}|M_i)$ denotes the marginal probability of the data $\boldsymbol{y}$ under model M_i and can be computed as

$$p(\boldsymbol{y}|M_i) = \int_{\boldsymbol{\psi}} p(\boldsymbol{y}|\boldsymbol{\psi}, M_i)p(\boldsymbol{\psi}|M_i)d\boldsymbol{\psi}\cdot$$

The larger (smaller) the value of BF_{21}, the stronger (weaker) is the evidence in favor of model M_2 versus M_1.

If one assumes prior probabilities of $p(M_i)$ on model $M_i, i = 1, 2$, then one obtains

$$\frac{p(M_2|\boldsymbol{y})}{p(M_1|\boldsymbol{y})} = \frac{p(\boldsymbol{y}|M_2)}{p(\boldsymbol{y}|M_1)}\frac{p(M_2)}{p(M_1)},$$

that is,

$$\text{Posterior Odds in favor of Model 2} = BF_{21} \times \text{Prior Odds in favor of Model 2}\cdot \quad (5)$$

Thus, the Bayes factor can be interpreted as the ratio between the posterior odds and prior odds in favor of a model.

The Strength of the Evidence Provided by Bayes Factors

A large value of BF_{21} provides strong evidence in favor of model M_2 versus model M_1. Kass and Raftery (1995) provided the following guidelines on the relationship between the value of the Bayes factor and the strength of the evidence it provides in favor of Model 2 versus Model 1. Thus, for example, values of 3-20, 20-150, and larger than 150 of BF_{21}, or values of 1-3, 3-5, or larger than 5 of $\log(BF_{21})$, provide a positive, strong, and very strong evidence in favor of that model.

Table 1. Interpretation of the Bayes Factor.

Bayes factor	log of Bayes factor	Evidence
1-3	0-1	Not worth more than a bare mention
3-20	1-3	Positive
20-150	3-5	Strong
>150	>5	Very strong

Existing Applications to Educational and Psychological Measurement

Hoijtink, Mulder, van Lissa, and Gu (2019), Masson (2011), Morey, Romeijn, and Rouder (2016), Wetzels et al. (2011), and Wagenmakers (2007) provided widely accessible overviews of Bayes factors and described how they can be useful to researchers and practitioners in psychology. Researchers such as Fox, Mulder, and Sinharay (2017), Gu, Mulder, Deković, and Hoijtink (2014), Klugkist, Laudy, and Hoijtink (2005), Mulder et al. (2009), Schnbrodt, Wagenmakers, Zehetleitner, and Perugini (2017), and Verhagen, Levy, Millsap, and Fox (2016) showed how to use Bayes factors to test hypothesis regarding covariance structures underlying IRT models, evaluate inequality-constrained hypothesis, evaluate analysis of variance models with inequality constraints, evaluate hypothesis in repeated measurements, perform sequential hypothesis testing, and test for measurement invariance in IRT models. However, Bayes factors have not been applied to detection of test fraud or to score differencing.

Bayes Factor for Score Differencing

One can consider score differencing as a comparison of two models M_1 and M_2, where a common examinee ability (θ) underlies all the item scores ($\boldsymbol{X}$) of the examinee under M_1 and two different abilities, θ_1 and θ_2, underlie the scores ($\boldsymbol{X}_1$ and $\boldsymbol{X}_2$) of the examinee on item sets S_1 and S_2 under M_2. Thus, in score differencing, the model M_1 represents no performance difference, the model M_2 represents a possible performance difference, the data consist of $\boldsymbol{X}$ under both models M_1 and M_2, the parameters are θ under M_1 and θ_1 and θ_2 with the restriction $\theta_2 \geq \theta_1$ under M_2. Let us assume a standard normal prior distribution on θ under M_1. To define the prior distribution under M_2, let us assume that θ_1 and θ_2 are independent of each other and θ_1 follows the standard normal distribution and θ_2 follows a normal distribution with mean 0 and variance 10, but truncated so that $\theta_2 \geq \theta_1$. This joint prior distribution is essentially equal to $2\phi(\theta_1)\frac{1}{\sqrt{10}}\phi(\frac{\theta_2}{\sqrt{10}})$, where $\phi(.)$ denotes the probability density function of the standard normal distribution.[3] The assumption of

[3] Note that $\int_{\theta_1=-\infty}^{\theta_1=\infty}\int_{\theta_2=\theta_1}^{\theta_2=\infty} 2\phi(\theta_1)\frac{1}{\sqrt{10}}\phi(\frac{\theta_2}{\sqrt{10}})d\theta_1 d\theta_2 = 1$.

a variance of 10 on θ_2 acknowledges the possibility that under M_2, θ_2 may be large when there is preknowledge (for example, Sinharay, 2017a, reported values between 2.16 and 2.81 of estimates of θ_2 for three examinees who were flagged for cheating on a licensure examination).

Then, the Bayes factor for score differencing can be computed as

$$\begin{aligned} BF_{21} &= \frac{p(\boldsymbol{X}|M_2)}{p(\boldsymbol{X}|M_1)} \\ &= \frac{\int_{\theta_1=-\infty}^{\theta_1=\infty}\int_{\theta_2=\theta_1}^{\theta_2=\infty} 2L(\theta_1;\boldsymbol{X}_1)L(\theta_2;\boldsymbol{X}_2)\phi(\theta_1)\frac{1}{\sqrt{10}}\phi(\frac{\theta_2}{\sqrt{10}})d\theta_1 d\theta_2}{\int_{\theta=-\infty}^{\theta=\infty} L(\theta;\boldsymbol{X})\phi(\theta)d\theta}. \end{aligned} \tag{6}$$

Larger values of BF_{21} provide more evidence in favor of a significant score difference; the numbers in Table 1 can be used as guidelines on the strength of evidence in favor of a significant score difference provided by various values of BF_{21}.

An Illustration of the Application of Bayes Factors to Score Differencing

Consider a test with 20 items. Let us consider the true IRT model is the Rasch model and the true item difficulty is 0 for all items. Let us consider that score differencing has to be performed with the first 10 items and the last 10 items as the two item sets and that the alternative hypothesis is that the performance is better on the second set. Consider 6 examinees all of whom obtain a total (or raw) score of 10 on the test, but

- Examinee 1 obtains raw scores of 5 and 5 on item sets 1 and 2, respectively.
- Examinee 2 obtains raw scores of 4 and 6 on item sets 1 and 2, respectively.
- ...
- Examinee 6 obtains raw scores of 0 and 10 on item sets 1 and 2, respectively.

Table 2 provides the difference in raw score between the second half and the first half, $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}$, the SLR statistic provided by Equation 3, and the Bayes factor provided by Equation 6 for the six examinees. As one goes down the table, the score difference increases, that is, the evidence becomes stronger in favor of Model 2 that corresponds to a

Table 2. Results for six examinees in the Illustration.

Examinee	Score Diff	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}$	SLR	BF
1	0	0.00	0.00	0.00	0.00	0.68
2	2	-0.37	0.37	0.00	0.89	1.56
3	4	-0.77	0.77	0.00	1.81	5.64
4	6	-1.22	1.22	0.00	2.76	41.04
5	8	-1.85	1.85	0.00	3.81	793.5
6	10	-3.04	3.04	0.00	5.09	56373

Note: ‘Score Diff’ and BF respectively denote ‘Difference in the raw score’ and ‘Bayes factor’.

possible performance difference. As a consequence, both the SLR statistic and Bayes factor increases as one goes down the table. Noting that SLR statistic follows a standard normal distribution under the null hypothesis, the null hypothesis of no performance difference between the two halves of the test is not rejected for Examinees 1-2 and rejected for Examinees 3-6 at 5% level. Table 2 implies that the evidence in favor of Model 2 (or, a better performance on the second half) is not more than a bare mention for Examinees 1 and 2, positive for Examinee 3, strong for Examinee 4, and very strong for Examinees 5 and 6.

A Simulation Study

We used simulations based on real data to examine the properties of the suggested Bayes factors and to compare the properties of the Bayesian factors to those of the SLR statistic.

Study Design

The simulations were based on the item scores and response times of about 44,000 test takers on one form of a subject of a state test. The test consists of 75 multiple-choice items. There was no knowledge of examinees benefiting from any kind of test fraud on the test. The item parameters of the data set were estimated under the 2PLM. The MLE of ability parameter were computed for all examinees.

The data set was used to artificially create several simulated data sets that involve different extents of item preknowledge, which leads to a performance/score difference. The following two factors were varied in the simulations:

- the size of the set of compromised items (10, 20, or 30 items).
- the number of examinees in the sample who had item preknowledge as a percentage of those who did not have preknowledge (5, 10, or 20).

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

1. Randomly select 10,000 examinees (who comprise a little less than a quarter of all the examinees in the original data set) from the original data set. These 10,000 examinees will play the role of those who did not have item preknowledge.
2. From the rest of the original data set, randomly select 500, 1,000, or 2,000 examinees (that constitute 5, 10, or 20% of the 10,000) who would play the role of the cheaters, that is, those who had item preknowledge.
3. From the 75 items in the data set, randomly choose the 10, 20, or 30 items that would play the role of the compromised items.
4. For each combination of a compromised item and a cheater, artificially create item preknowledge by replacing the item scores of the cheaters on the compromised items by numbers randomly drawn from a Bernoulli distribution with success probability equal to the success probability under the 2PLM with the abovementioned estimated item parameters and ability equal to the estimated ability plus 2. Thus, it is assumed that the effect of preknowledge on an item is equivalent to a boost in the ability parameter for that item.
5. Compute the estimated item parameters for the 2PLM from the (changed) data set.

6. Compute the MLEs of the examinee ability (truncated between -4.0 and 4.0) on the whole test, compromised items, and non-compromised items from the data set using the item parameters computed in Step 5.

7. Compute the Bayes factor and the SLR statistic for score differencing for all the examinees in the (changed) data set using the estimated item parameters computed above.

Results from the Simulations

Figure 1 shows a scatter plot of the logarithm of Bayes factors versus the p-values for the SLR statistic for a 5,000 examinees randomly drawn from all simulated examinees. The true cheaters (those with preknowledge) are shown using black circles and the true non-cheaters are shown using gray circles. Two vertical dashed lines show the p-values of 0.01 and 0.05. The figure shows that

- The points for true non-cheaters mostly appear to the right and the bottom (that is, the Bayes factor is mostly small and p-value is mostly large for them) while those for the true cheaters mostly appear to the left

- In general, the Bayes factor increases as p-value decreases

- The average Bayes factor is about 1.5 and 7.0 for p-values of 0.05 and 0.01, respectively.

- Several points lie along a vertical line at p-value=0.5. These are outcomes of the statistic Λ in Equation 2 occasionally becoming negative.[4]

The distribution of the Bayes factors and p-values are not influenced much by the percent of cheaters in the data set, but substantially influenced by the number of compromised items. Therefore, for each value of the number of compromised items, we pooled the Bayes factors and p-values over the three levels of percent of cheaters. Table 3 shows the percentage of examinees with various levels of values of Bayes factors and p-values

[4] Sinharay (2017a) noted this phenomenon that occurs when $\hat{\theta}_1$ and $\hat{\theta}_2$ are very close—a conclusion of no significant score difference is made for the corresponding examinees.

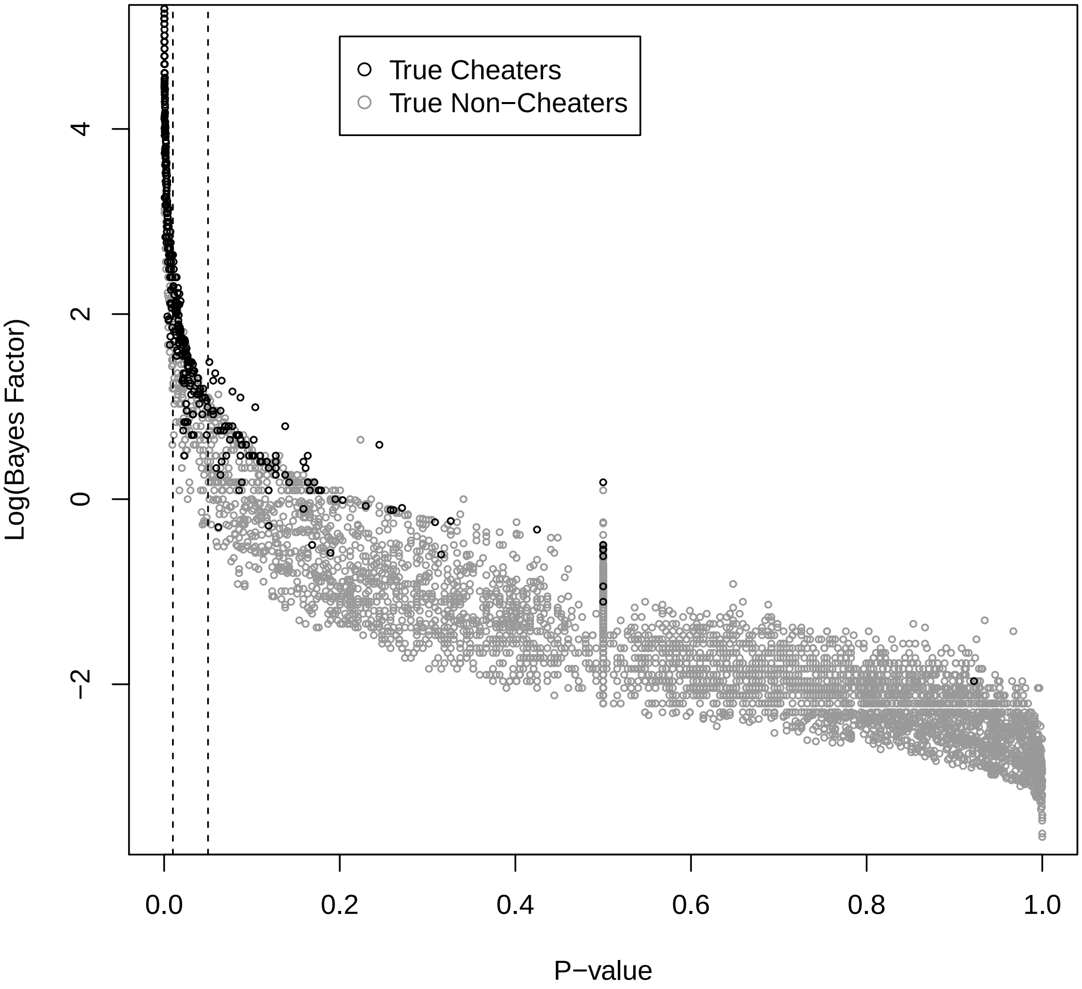

Figure 1. A scatter plot of the logarithm of Bayes factors versus the p-values for the SLR statistic.

for 10, 20, and 30 compromised items. The levels for Bayes factors are those from Table 1. The levels of p-values used in the table (< 0.001, 0.001-0.01, 0.01-0.05, and > 0.05) are guided by the traditional interpretation of them found in, for example, (e.g., Wasserman, 2004, p. 157) who mentioned that p-values of < 0.01, 0.01-0.05, and > 0.05 provide very strong evidence, strong evidence, and weak to no evidence against the null hypothesis.

Table 3. The percent of examinees for different combinations of P-values and Bayes factors.

NC	Bayes	P-values for True Non-Cheaters				P-values for True Cheaters			
	Factor	> 0.05	0.01-0.05	0.001-0.01	< 0.001	> 0.05	0.01-0.05	0.001-0.01	< 0.001
10	< 1	88	0	0	0	11	0	0	0
	1-3	8	1	0	0	23	4	0	0
	3-20	0	3	1	0	0	26	13	0
	20-150	0	0	0	0	0	0	12	4
	> 150	0	0	0	0	0	0	0	6
20	< 1	94	0	0	0	4	0	0	0
	1-3	2	2	0	0	9	3	0	0
	3-20	0	1	1	0	0	15	14	0
	20-150	0	0	0	0	0	0	15	12
	> 150	0	0	0	0	0	0	0	28
30	< 1	95	0	0	0	3	0	0	0
	1-3	1	2	0	0	6	2	0	0
	3-20	0	1	1	0	2	12	10	0
	20-150	0	0	0	0	0	0	14	11
	> 150	0	0	0	0	0	0	0	41

Note: “NC” means number of compromised items.

Four columns towards the left of Table 3 show percentages of examinees among the true non-cheaters and the four columns towards the right show percentages of examinees among the true cheaters. Rows 1-5, 6-10, and 11-15 show the percentages for 10, 20, and 30 compromised items, respectively. Note that the percentages add up to 100 for either the non-cheaters or cheaters for each number of compromised items. Table 3 shows that

- In agreement with Figure 1, the percentages of examinees are large for small Bayes factors and large p-values and also for large Bayes factors and small p-values.

- The p-value is larger than 0.05 and the Bayes factor is smaller than 1 for a large percentage of true non-cheaters (88, 94, and 95 percent, respectively, for 10, 20, and 30 compromised items), but for a small percentage of true cheaters.

- As the number of compromised items increases, the percentages for the true non-cheaters do not change much, but the percentage of more extreme p-value and Bayes factor increases for the true cheaters.

- When the p-value is between 0.01 and 0.05 (a range of values for which a frequentist often rejects the null hypothesis and, in this context, would often conclude that the corresponding examinee benefited from preknowledge), the Bayes factor is smaller than 3 (that is, provides evidence that is not worth more than a bare mention) about 25% of the times. Wetzels et al. (2011) also noted the tendency of the Bayes factor to be smaller than 3 when the p-value is between 0.01 and 0.05.

The comparison of the power of statistics for detecting aberrant examinees has been performed using receiver operating characteristics (ROC) curves at least since Drasgow, Levine, and Williams (1985). Given the values of a statistic (whose larger value indicates more aberrance) from a simulated data set, an ROC curve requires the computation of the following two quantities for several values of c:

- the false alarm rate (or "false positive rate" or "Type I error rate"), $F(c)$, which is the proportion of times when the statistic for a non-aberrant examinee is more than c

- the hit rate (or "true positive rate" or "power"), $H(c)$, which is the proportion of times when the statistic for an aberrant examinee is more than c

Then, a graphical plot is created in which $F(c)$ is plotted along the x-axis, $H(c)$ is plotted along the y-axis, and a line joins $\{F(c), H(c)\}$ for several values of c. The line is referred to as the *ROC curve*. Figure 2 shows the ROC curves for the SLR statistic (solid line) and Bayes factor (BF; dotted line) for the case of 10 compromised items and 10% aberrant examinees. A diagonal line is shown for convenience. It is possible to use the area under the ROC Curve (AUROC; Hanley & McNeil, 1982) of a statistic as a measure of how powerful the statistic is. The AUROC of a very powerful statistic is expected to be close to 1 because the hit rate of such a statistic will be close to 1 for most values of the false positive rate. In the context of detecting aberrant examinees, researchers such as Sinharay (2017b) used

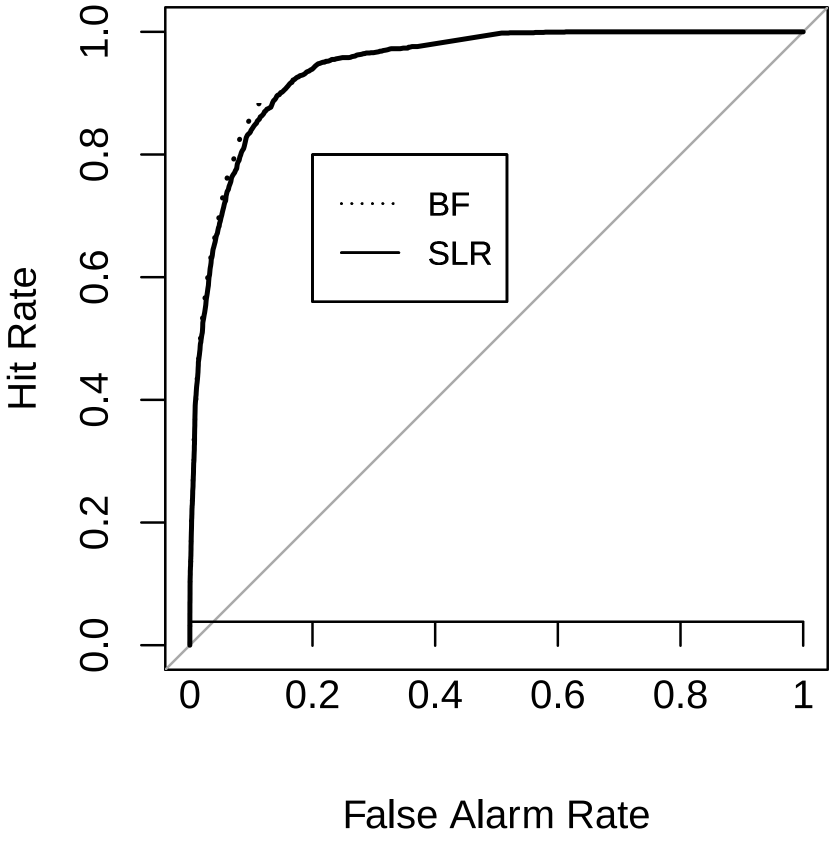

Figure 2. The ROC curve for 10 compromised items and 10% aberrant examinees.

truncated ROC areas, or areas under the ROC curves truncated between 0 and 0.1 and divided by 0.10—that is because false positive rates larger than 0.10 are hardly employed in the context of detecting aberrant examinees (e.g., Wollack, Cohen, & Eckerly, 2015). The truncated ROC area of a very powerful statistic is expected to be close to 1. The truncated ROC areas of the SLR statistic and Bayes factor are very close for all the simulation cases and the Bayes factor has slightly larger truncated ROC areas than the SLR statistic in a few simulation cases. The average truncated ROC areas of the SLR statistic and Bayes factor, averaged over all simulation cases, are 0.90 and 0.92, respectively. Thus, the simulations show that the Bayes factor seems to flag the cheaters a little more often compared to the

SLR statistic while not flagging the non-cheaters too often and provide some evidence of that the Bayes factor may be superior compared to the SLR statistic in some cases.

Real Data Example

Data

Let us consider item-response data from one form of a non-adaptive licensure assessment. The data set was analyzed in several chapters of Cizek and Wollack (2017) and also by Sinharay (2017a), and Sinharay and Jensen (2019). The form includes 170 operational items that are dichotomously scored. Item scores were available for 1,644 examinees for the form. The licensure organization who provided the data identified 61 items on the form as compromised. The organization also flagged 48 individuals on the form as possible cheaters from a variety of statistical analysis and a rigorous investigative process that brought in other information; these 48 examinees will be treated as true cheaters. As in Sinharay (2017a), the interest here will be in detecting item preknowledge.

Analysis and Results

The 2PLM was used for the analysis. The marginal maximum likelihood estimation procedure was used to estimate the item parameters from the data set and these estimates were used in the computation of the statistics. The values of the SLR statistic and the Bayes factor were computed for each individual in the data set. The MLEs of the abilities, restricted to the range -4.0 and 4.0, were used to compute the SLR statistic. The set of 109 non-compromised items was considered as the first set of items and the set of 61 compromised items were considered as the second set of items.

Figure 3 shows scatter plots of the SLR statistic versus the logarithm of the Bayes factor (left panel) and the p-values corresponding to the SLR statistic versus the logarithm of the Bayes factor (right panel) for all the examinees in the data set. In the figure, the gray circles correspond to the examinees who were not flagged by the licensure organization and the black circles correspond to the examinees who were flagged as possible cheaters by

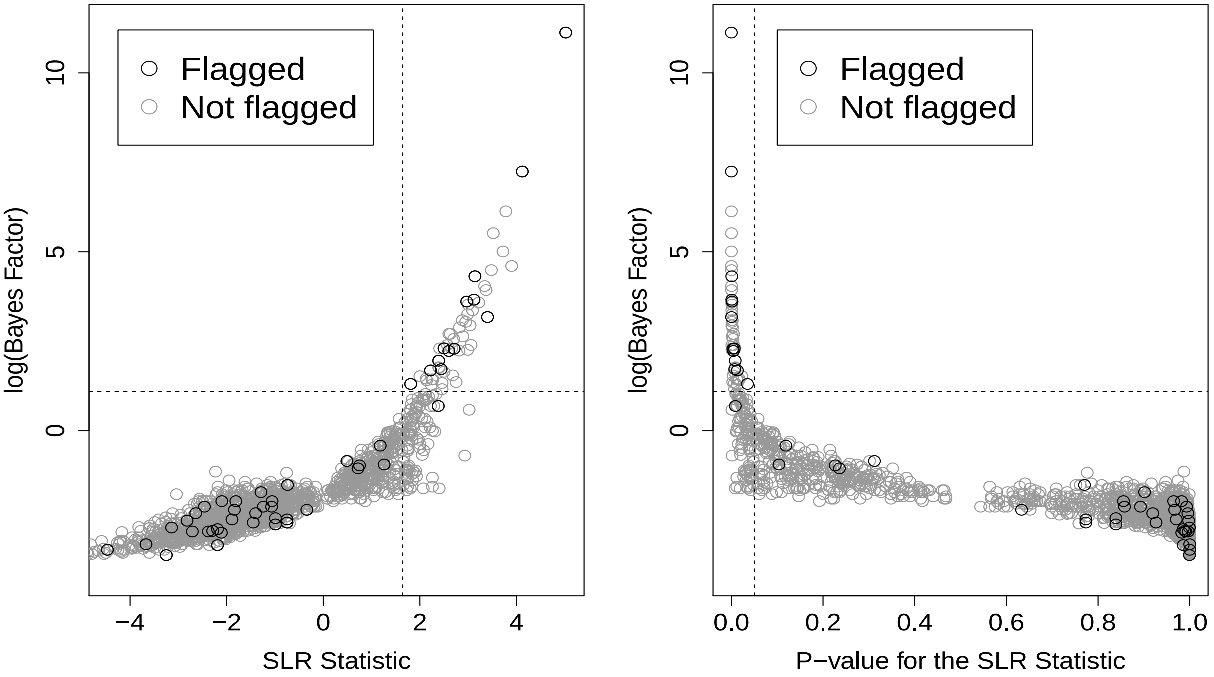

Figure 3. Scatter plots of the SLR statistic and the corresponding p-value versus the Bayes factor for the Real Data Example.

the licensure organization. In the left panel, horizontal and vertical dashed lines represent cutoffs of log(3) and 1.64 for the Bayes factor and the SLR statistic, respectively. In the right panel, horizontal and vertical dashed lines represent cutoffs of log(3) and 0.05 for the Bayes factor and the p-value. This choice of the cutoff for the Bayes factor is justified by the fact that in our simulations, the 95th percentile of the Bayes factor for true non-cheaters was close to 3, and also by findings of researchers such as Wetzels et al. (2011) who noted that p-values around 0.05 are roughly equivalent to Bayes factors around 3. The two statistics are positively correlated, that is, the Bayes factor increases as the SLR statistic increases. The right panel of the figure looks similar to Figure 1—so the relationship between the Bayes factor and SLR statistic is similar over the simulated and real data sets. Interestingly, for the examinee in the top right corner of the figure, the SLR statistic is 5.02

and the Bayes factor is about 68,000.

Table 4. The Percent of Examinees Above the Cutoff Values for the Licensure Data.

Examinees	SLR	Bayes Factor
Not Flagged	8.4	4.1
Flagged	29.2	27.1

The percent of examinees for whom the SLR statistic and the Bayes factor are above their respective cutoffs (1.64 and 3) are provided in Table 4. The first row of Table 4 shows the percents above the cutoff values among the examinees who were not flagged by the licensure organization. The second row of the table shows the percents above the cutoff values only among the 48 examinees who were flagged by the licensure organization. Table 4 shows that the SLR statistic is larger than the cutoff more often compared to the Bayes factor for both the "Not flagged" and "Flagged" group of examinees. Thus, the use of the Bayes factor with a cutoff of 3 would lead to a more conservative approach than the use of the SLR statistic with a cutoff of 1.64. While the conservativeness of the Bayes factor will protect the administrators from false positives, it will lead to fewer true positives.

Table 4 also shows that the percent above the cutoff for each statistic is much larger among the examinees flagged by the licensure organization (bottom row of the table) than among those not flagged (top row of the table)—this result provides some evidence that the statistics are somewhat successful—they are significant at a larger rate among the examinees who are true cheaters.

Note that several experts recommended against making conclusions by dichotomizing evidence using one frequentist or Bayesian statistic (e.g., Wasserstein & Lazar, 2016) and we agree with that viewpoint—Table 4 is just an attempt to compare the values of the SLR statistic and Bayes factor. In a real application, to determine whether an examinee was involved in test fraud, an investigator would most likely use the value of one of these statistics for the examinee as one piece of evidence along with other non-statistical evidence such as seating chart and proctor report (e.g., Tendeiro & Meijer, 2014).

Conclusions

In this paper, Bayes factors (e.g., Kass & Raftery, 1995) were suggested as an alternative tool for score differencing (Wollack & Schoenig, 2018). A simulation study was used to examine the performance of the Bayes factor in comparison to that of a frequentist statistic for score differencing. In a real-data application, the Bayes factor was found to lead to slightly smaller false positive rate and slightly smaller hit rate compared to a frequentist statistic for score differencing.

van der Linden and Lewis (2015) suggested the posterior odds of cheating for detecting various types of cheating on tests. They provided details on the computation of the posterior odds to detect fraudulent erasures. Given Equation 5, the Bayes factor is closely related to posterior odds. However, the computation of the posterior odds to detect fraudulent erasures in van der Linden and Lewis (2015) was predicated on a specialized IRT model that applies only to fraudulent erasures and the approach cannot be easily extended to score differencing.

In this paper, the cutoff for the Bayes factor was set equal to 3, which is the boundary between "non-positive" and "positive" evidence, in the real data example. This choice led to results that are comparable and close to those with frequentist p-values. In future research, other choices of the cutoff can be explored. It is possible to use a simulation-based cutoff—such a choice will lead to a false positive rate that is very close to the level of significance.

While this paper is one of the first to apply Bayesian methods to score differencing, it is possible to extend our research in several ways. First, more simulated data and real data should be analyzed using the method. Second, it is possible to compare the suggested Bayesian approach to other frequentist methods and to the Bayesian predictive checking method of Wang et al. (2017). Third, while some limited simulations (not reported here) shows the suggested Bayes factor to not be influenced much by the prior distributions on the ability parameters,[5] the sensitivity of the suggested Bayes factor to the prior distribution can be studied further. Fourth, it is possible to extend the approach to cases where both item scores and response times of examinees are available; the use of both scores and times could lead to a more powerful approach. Finally, other Bayesian approaches such as the use of the posterior probability (e.g., Hoijtink et al., 2019; Gelman et al., 2014) of a model given the data could be used to score differencing.

References

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

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

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.

[5] Though, in those simulations, we noticed the Bayes factor to be slightly more liberal with an increase in the prior variances of the ability distributions.

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.

Fischer, G. H. (2003). The precision of gain scores under an item response theory perspective: A comparison of asymptotic and exact conditional inference about change. *Applied Psychological Measurement*, *27*, 3–26.

Fox, J.-P., Mulder, J., & Sinharay, S. (2017). Bayes factor covariance testing in item response models. *Psychometrika*, *82*, 979–1006.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis (3rd edition).* New York, NY: Chapman and Hall.

Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*, 511–527.

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.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29-36.

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*. (Advance online publication. doi:10.1037/met0000201)

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

Kingston, N. (2013). Educator testing case studies. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 299–311). New York, NY: Routledge.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*, 477–493.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*, 679–690.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18.

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H. J., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, *53*, 530–546.

Schnbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*, 322–339.

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

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

Sinharay, S. (2018). Application of Bayesian methods for detecting fraudulent behavior on tests. *Measurement: Interdisciplinary Research and Perspective*, *16*, 100–113.

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.

Sinharay, S., & Jensen, J. L. (2019). Higher-order asymptotics and its application to testing the equality of the examinee ability over two sets of items. *Psychometrika*, *84*, 484–510.

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

Strauss, V. (2014). *The six-step SAT cheating operation in Asia and how to stop it.* (Retrieved from https://www.washingtonpost.com/news/answer-sheet/wp/2014/11/16/the-six-step-sat-cheating-operation-in-asia-and-how-to-stop-it/)

Tendeiro, J. N., & Meijer, R. R. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement*, *51*, 239–259.

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

van der Linden, W. J., & Lewis, C. (2015). Bayesian checks on cheating on tests. *Psychometrika*, *80*, 689–706.

Verhagen, J., Levy, R., Millsap, R. E., & Fox, J.-P. (2016). Evaluating evidence for invariant items: A Bayes factor applied to testing measurement invariance in IRT models. *Journal of Mathematical Psychology*, *72*, 171–182.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.

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

Wang, X., Liu, Y., Robin, F., & Guo, H. (2019). A comparison of methods for detecting examinee preknowledge of items. *International Journal of Testing*, *19*, 207–226.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. New York, NY: Springer.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*, 129–133.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology. *Perspectives on Psychological Science*, *6*, 291–298.

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

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.