


## The Use of the Posterior Probability in Score Differencing

Sandip Sinharay 

Matthew S. Johnson 

Educational Testing Service

*Score differencing is one of the six categories of statistical methods used to detect test fraud (Wollack & Schoenig, 2018) and involves the testing of the null hypothesis that the performance of an examinee is similar over two item sets versus the alternative hypothesis that the performance is better on one of the item sets. We suggest, to perform score differencing, the use of the posterior probability of better performance on one item set compared to another. In a simulation study, the suggested approach performs satisfactory compared to several existing approaches for score differencing. A real data example demonstrates how the suggested approach may be effective in detecting fraudulent examinees. The results in this article call for more attention to the use of posterior probabilities, and Bayesian approaches in general, in investigations of test fraud.*

Keywords: *Bayes factor; likelihood ratio statistic; score differencing*

Researchers such as van der Linden (2009) noted that an increasing concern of producers and consumers of test scores is fraudulent behavior before and during the test and that such behavior is more likely to be observed when the stakes are high, such as in licensing, admission, and certification testing. Naturally, there is an upswing in research on statistical methods and models that can be used to detect test fraud. The statistical methods to detect test fraud were divided into six categories by Wollack and Schoenig (2018). One of the categories is “score differencing,” which involves a test of the null hypothesis of equal ability of an examinee over two sets of items  $\mathcal{I}_1$  and  $\mathcal{I}_2$ . Score differencing can be used to detect several types of test fraud including item preknowledge (e.g., Sinharay, 2017a, 2017b; Sinharay & Jensen, 2019), fraudulent erasures (e.g., Sinharay et al., 2017), fraudulent gain scores (e.g., Fischer, 2003), and cheating on unproctored tests (e.g., Guo & Drasgow, 2010).

With the exceptions of Sinharay and Johnson (2020) and Wang et al. (2017), the currently used methods for score differencing are mostly frequentist and are dependent on (frequentist)  $p$  values. As researchers such as Allen and Ghattas (2016), Skorupski and Wainer (2017), and van der Linden and Lewis (2015)

noted, a frequentist  $p$  value 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?” which is not the question the investigators are interested in when they are trying to detect test fraud. The question of interest actually is “Given the available information, what is the chance that the examinee committed a test fraud?” and this question conforms more with a Bayesian approach than a frequentist approach. Consequently, van der Linden and Lewis (2015), Allen and Ghattas (2016), Sinharay (2018), and Skorupski and Wainer (2017) called for more applications of Bayesian statistical methods to the detection of test fraud. In addition, a recent statement by the American Statistical Association (Wasserstein & Lazar, 2016) included the recommendation that researchers and practitioners should explore approaches other than the frequentist  $p$  values, and Bayesian approaches are included in their list of “other approaches.”

However, Bayesian methods have rarely been applied in score differencing, with the exception of Sinharay and Johnson (2020), who suggested the use of Bayes factors, and Wang et al. (2017), who suggested the use of a Bayesian predictive checking methodology. The goal of this article is to suggest the Bayesian approach of using the posterior probability given the item score for score differencing.

The next section includes descriptions of score differencing and of the existing frequentist and Bayesian approaches for score differencing. The following section includes a description of our suggested approach of the use of posterior probability for score differencing. Simulated and real data sets are analyzed in the next two sections. The last section includes conclusions and recommendations.

## **Review of Score Differencing**

### *Description of Score Differencing*

Consider a test with  $N$  items, each of which can be a dichotomously or polytomously scored item. Let  $0, 1, \dots, m_i$  denote the possible scores on item  $i$ . Let us consider a randomly chosen examinee whose true overall ability is  $\theta$ . Score differencing for an examinee involves an examination of whether the examinee’s performance is equal over item sets  $\mathcal{I}_1$  and  $\mathcal{I}_2$ . The item sets  $\mathcal{I}_1$  and  $\mathcal{I}_2$  are nonoverlapping and together include all the  $N$  items. In most applications of score differencing, the sets  $\mathcal{I}_1$  and  $\mathcal{I}_2$  would be naturally defined. For example, Table 1 provides the sets  $\mathcal{I}_1$  and  $\mathcal{I}_2$  in four applications of score differencing.

Also, note that  $\mathcal{I}_1$  and  $\mathcal{I}_2$  could vary over the examinees. For example, they would be different over the examinees in the detection of item preknowledge on an adaptive test because of the administration of different items over examinees on adaptive tests, and they would be different over the examinees in the detection of fraudulent erasures because the set of items with erasures is typically different over the examinees.

TABLE 1.  
*The Item Sets in Various Applications of Score Differencing*

Application in Detection of	$\mathcal{I}_1$	$\mathcal{I}_2$
Item preknowledge	Noncompromised items	Compromised items
Fraudulent erasures	Nonerased items	Erased items
Fraudulent gain scores	Items on first administration	Items on second administration
Cheating on unproctored tests	Items on proctored test	Items on unproctored test

Let the true ability of the examinee on  $\mathcal{I}_1$  and  $\mathcal{I}_2$  be denoted as  $\theta_1$  and  $\theta_2$ , respectively. The null hypothesis of interest in score differencing can then be expressed as  $H_0 : \theta_1 = \theta_2$ . 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. Thus, the null and alternative hypotheses correspond to the answering behaviors of a noncheater and a cheater, respectively. Let us assume, without loss of generality, that the alternative hypothesis is that the performance on  $\mathcal{I}_2$  is better than that on  $\mathcal{I}_1$  for the examinee or that  $\theta_2 > \theta_1$ . For example, in an application of score differencing to the detection of item preknowledge, the alternative hypothesis is that the performance on the compromised items is better than that on the non-compromised items. The alternative hypothesis represents the situation where, due to test fraud, the examinee received a performance boost that is equivalent to an increase of  $\theta_2 - \theta_1$  in ability (whereas, without the fraud, the boost would be zero and  $\theta_2$  would be equal to  $\theta_1$ ).

Let  $y_1, y_2, \dots, y_N$  denote the scores for the examinee on the  $N$  items of the test and let  $(y_1, y_2, \dots, y_N)$  be denoted as  $\mathbf{y}$ . Let  $\mathbf{y}_1 = \{y_i, i \in \mathcal{I}_1\}$  and  $\mathbf{y}_2 = \{y_i, i \in \mathcal{I}_2\}$ , respectively, denote the collection of the scores of the examinee on the items in Sets 1 and 2. Let the probability of a score  $j$  on item  $i$  for the examinee be denoted as

$$P_{ij}(\theta) = P(y_i = j|\theta), j = 0, 1, 2, \dots, m_i; i = 1, 2, \dots, N,$$

where  $m_i$  is the maximum possible score on item  $i$ . For example, for the generalized partial credit model (Muraki, 1992),

$$P_{ij}(\theta) = \frac{\exp \left[ \sum_{h=0}^j a_i(\theta - b_{ih}) \right]}{\sum_{c=0}^{m_i} \exp \left[ \sum_{h=0}^c a_i(\theta - b_{ih}) \right]},$$

where  $a_i$  and  $b_{ih}$ , respectively, denote the slope and the location/threshold parameters of item  $i$ , and  $b_{i0} = 0$ .

Using the conditional independence assumption of item response theory (IRT), the likelihood of the examinee, henceforth denoted as  $L(\theta; \mathbf{y})$ , is given by

$$L(\theta; \mathbf{y}) = \prod_{i=1}^N \prod_{j=0}^{m_i} P_{ij}(\theta)^{d_j(y_i)}, \tag{1}$$

where  $d_j(y_i) = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{otherwise.} \end{cases}$

The above description encompasses dichotomous items as well. If item  $i$  is dichotomous, then  $m_i = 1$ , and

$$d_0(y_i) = 1 - y_i, d_1(y_i) = y_i, P_{i0}(\theta) = P(y_i = 0), \quad \text{and} \quad P_{i1}(\theta) = P(y_i = 1).$$

For example, if the two-parameter logistic model (2PLM) is used for item  $i$  that is dichotomous, then

$$P_{i1}(\theta) = \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]} \quad \text{and} \quad P_{i0}(\theta) = \frac{1}{1 + \exp[a_i(\theta - b_i)]},$$

where  $a_i$  and  $b_i$ , respectively, are the slope and difficulty parameters of item  $i$ . The Rasch (1960) model is a special case of the 2PLM with the  $a_i$ s being the same over all the items.

For an examinee, let us define the maximum likelihood estimate (MLE) or the weighted MLE (Warm, 1989) of the examinee ability from the scores on  $\mathcal{I}_1, \mathcal{I}_2$ , and all the items as  $\hat{\theta}_1, \hat{\theta}_2$ , and  $\hat{\theta}$ , respectively.

*A Frequentist Approach to Score Differencing*

Let us denote the log-likelihood of an examinee as  $l(\theta; \mathbf{y})$ , that is,

$$l(\theta; \mathbf{y}) = \log(L(\theta; \mathbf{y})).$$

The likelihood ratio test statistic (e.g., Finkelman et al., 2010; Guo & Drasgow, 2010) for testing the null hypothesis  $H_0 : \theta_1 = \theta_2$  is given by

$$\begin{aligned} \Lambda &= 2 \left[ l(\hat{\theta}_1; \mathbf{y}_1) + l(\hat{\theta}_2; \mathbf{y}_2) - l(\hat{\theta}; \mathbf{y}) \right], \\ &= 2 \left[ \sum_{i \in \mathcal{I}_1} \sum_{j=0}^{m_i} d_j(y_i) \log P_{ij}(\hat{\theta}_1) + \sum_{i \in \mathcal{I}_2} \sum_{j=0}^{m_i} d_j(y_i) \log P_{ij}(\hat{\theta}_2) - \sum_{i=1}^N \sum_{j=0}^{m_i} d_j(y_i) \log P_{ij}(\hat{\theta}) \right], \\ &= 2 \left[ \sum_{i \in \mathcal{I}_1} \sum_{j=0}^{m_i} d_j(y_i) \log \frac{P_{ij}(\hat{\theta}_1)}{P_{ij}(\hat{\theta})} + \sum_{i \in \mathcal{I}_2} \sum_{j=0}^{m_i} d_j(y_i) \log \frac{P_{ij}(\hat{\theta}_2)}{P_{ij}(\hat{\theta})} \right]. \end{aligned} \tag{2}$$

For score differencing, that is, for testing the null hypothesis  $H_0 : \theta_1 = \theta_2$  versus the alternative hypothesis  $H_1 : \theta_2 > \theta_1$ , Sinharay (2017a) suggested the signed likelihood ratio (SLR) statistic given by

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

When the log-likelihood  $l(\theta; \mathbf{y})$  originates from the commonly used IRT models, the statistic  $L_S$  has an asymptotic standard normal distribution under the null hypothesis (e.g., Sinharay, 2017a; Cox, 2006, p. 104). A large value of  $L_S$  leads to the rejection of the null hypothesis. Sinharay (2017a, 2017b) and Wang et al. (2019) demonstrated using real and simulated data that the performance of  $L_S$  was satisfactory compared to that of several existing statistics for detecting item preknowledge, and Sinharay and Jensen (2019) found  $L_S$  to have satisfactory Type I error rates and power in several applications of score differencing. Therefore,  $L_S$  is the only frequentist statistic for score differencing that is considered in this article.

*Existing Bayesian Approaches for Score Differencing*

*Bayes factor.* The Bayes factor (e.g., Kass & Raftery, 1995) is a Bayesian approach for model comparison and can be applied when one is interested in determining whether the model  $M_2$  fits the available data better than does model  $M_1$ . The Bayes factor in favor of model  $M_2$  in comparison to  $M_1$  is given by

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

where, for example,  $p(\mathbf{y}|M_1)$  denotes the marginal probability of the data  $\mathbf{y}$  under model  $M_1$  and can be computed as

$$p(\mathbf{y}|M_1) = \int_{\Psi} p(\mathbf{y}|\psi, M_1)p(\psi|M_1)d\psi,$$

where  $p(\mathbf{y}|\psi, M_1)$  is the distribution of the data, given the parameters  $\psi$  under model  $M_1$  and  $p(\psi|M_1)$  is the prior distribution under model  $M_1$ . The larger (smaller) the value of  $BF_{21}$ , the stronger (weaker) is the evidence in favor of model  $M_2$  versus  $M_1$ . Kass and Raftery (1995) provided the guidelines shown in Table 2 on the relationship between the value of the Bayes factor and the evidence it provides in favor of Model 2 versus Model 1.

Sinharay and Johnson (2020) noted that it is possible to consider score differencing as a comparison of two models  $M_2$  and  $M_1$ , where  $M_1$  represents the assumption that a common examinee ability ( $\theta$ ) underlies all the item scores ( $\mathbf{y}$ ) and  $M_2$  represents the assumption that two different abilities ( $\theta_1$  and  $\theta_2$ ) underlie the scores ( $\mathbf{y}_1$  and  $\mathbf{y}_2$ ) of the examinee on item sets  $\mathcal{I}_1$  and  $\mathcal{I}_2$ . Therefore, the likelihood functions of an examinee’s scores under  $M_1$  and  $M_2$  are  $L(\theta; \mathbf{y})$  and  $L(\theta_1; \mathbf{y}_1)L(\theta_2; \mathbf{y}_2)$ , respectively,  $M_1$  represents no performance difference, and  $M_2$  represents a possible performance difference. Note that even though  $M_1$  and  $M_2$  typically represent two models in the computation of Bayes factors, they

TABLE 2.  
Interpretation of the Bayes Factor

Bayes Factor	Log of Bayes Factor	Evidence
1–3	0–1	Weak/not worth more than a bare mention
3–20	1–3	Positive
20–150	3–5	Strong
>150	>5	Very strong

both are based on the same IRT model in score differencing; they are different in the sense that  $M_1$  involves one ability parameter ( $\theta$ ) while  $M_2$  involves two ability parameters ( $\theta_1$  and  $\theta_2$ ) for the same examinee.

Then, Sinharay and Johnson (2020) showed that the Bayes factor in the context of score differencing can be computed as

$$\begin{aligned}
 BF_{21} &= \frac{p(\mathbf{y}|M_2)}{p(\mathbf{y}|M_1)}, \\
 &= \frac{\int_{\theta_1=-\infty}^{\theta_1=\infty} \int_{\theta_2=\theta_1}^{\theta_2=\infty} L(\theta_1; \mathbf{y}_1)L(\theta_2; \mathbf{y}_2)p(\theta_1, \theta_2)d\theta_1d\theta_2}{\int_{\theta=-\infty}^{\theta=\infty} L(\theta; \mathbf{y})\phi(\theta)d\theta}, \tag{5}
 \end{aligned}$$

where  $p(\theta_1, \theta_2)$  is the joint prior distribution on  $\theta_1$  and  $\theta_2$ . For example, if the 2PLM is used, then  $L(\theta_1; \mathbf{y}_1) = \prod_{i \in \mathcal{I}_1} \frac{\exp[y_i a_i (\theta_1 - b_i)]}{1 + \exp[a_i (\theta_1 - b_i)]}$ , and the Bayes factor can be computed as

$$\begin{aligned}
 BF_{21} &= \frac{p(\mathbf{y}|M_2)}{p(\mathbf{y}|M_1)}, \\
 &= \frac{\int_{\theta_1=-\infty}^{\theta_1=\infty} \int_{\theta_2=\theta_1}^{\theta_2=\infty} \left[ \prod_{i \in \mathcal{I}_1} \frac{\exp[y_i a_i (\theta_1 - b_i)]}{1 + \exp[a_i (\theta_1 - b_i)]} \right] \left[ \prod_{i \in \mathcal{I}_2} \frac{\exp[y_i a_i (\theta_2 - b_i)]}{1 + \exp[a_i (\theta_2 - b_i)]} \right] p(\theta_1, \theta_2)d\theta_1d\theta_2}{\int_{\theta=-\infty}^{\theta=\infty} \left[ \prod_{i=1}^N \frac{\exp[y_i a_i (\theta - b_i)]}{1 + \exp[a_i (\theta - b_i)]} \right] \phi(\theta)d\theta}. \tag{6}
 \end{aligned}$$

A large value of  $BF_{21}$  will provide strong evidence in favor of a large score difference. The guidelines shown in Table 2 can be used to determine what value of Bayes factor is large.

*Predictive checking method.* Wang et al. (2017) suggested a Bayesian predictive checking method to detect item preknowledge—the method can be used in other

types of score differencing as well. In this method, one computes  $g(\theta_1|\mathbf{y}_1)$ , the posterior distribution of the examinee ability given the examinee's item scores on  $\mathcal{I}_1$ . Then, one computes the predictive distribution of a test statistic  $T(\mathbf{y}_2)$ , such as the raw score on  $\mathcal{I}_2$ , as

$$g(T(\mathbf{y}_2) = t_2|\mathbf{y}_1) = \int_{\theta_1} p(T(\mathbf{y}_2) = t_2|\theta_1)g(\theta_1|\mathbf{y}_1)d\theta_1, \quad (7)$$

where  $p(T(\mathbf{y}_2) = t_2|\theta_1)$  is the probability that the test statistic is equal to  $t_2$ , given  $\theta_1$ . Finally, a *predictive p value* is computed as the probability of the test statistic under the predictive distribution being more extreme than the actual observed value of the statistic. A small predictive  $p$  value, for example, one smaller than .05 or .01, indicates potential item preknowledge for the corresponding examinee (Wang et al., 2017). The predictive  $p$  value is often computed using a simulation where several draws of the test statistic are made from the abovementioned predictive distribution of  $T(\mathbf{y}_2)$ . This predictive  $p$  value is similar in spirit to the posterior predictive  $p$  value (e.g., Gelman et al., 2014, p. 146). Wang et al. (2019) found the performance of the predictive checking method to be similar to that of the SLR statistic and superior to that of another existing statistic. To compute the predictive  $p$  value, as in Wang et al. (2017) and Wang et al. (2019), we set

$$T(\mathbf{y}_2) = \sum_{i \in \mathcal{I}_2} y_i = \text{the raw score on } \mathcal{I}_2.$$

We computed  $p(T(\mathbf{y}_2)|\theta_1)$  using the recursive formula of Lord and Wingersky (1984) and approximated the integral in Equation 7 using the Riemann approximation (e.g., Thisted, 1988, p. 262).

### **A New Bayesian Approach for Score Differencing: Use of Posterior Probability**

Score differencing essentially is a test of a hypothesis, that of the equality of examinee ability over two sets of items, against a one-sided alternative hypothesis. Researchers such as Gelman et al. (2014, p. 95), Robert (2007, p. 226), and Stern (2005) suggested that a direct measure of the scientific evidence in favor of an alternative hypothesis and against the null hypothesis can be obtained as the posterior probability of the event corresponding to the alternative hypothesis, and Stern (2005) recommended the use of the posterior probability specifically for testing against one-sided alternative hypotheses. The remainder of this section includes (a) the definition and details on the computation of the posterior probability for score differencing, (b) a discussion on the choice of an appropriate cutoff for the posterior probability, (c) a discussion on the choice of the prior distributions while computing the posterior probability, and (d) an illustration using a hypothetical data set.

Definition and Computations

Let the joint posterior distribution of  $\theta_1$  and  $\theta_2$ , given the item scores for an examinee, be denoted as  $g(\theta_1, \theta_2|\mathbf{y})$ . Because of the local independence assumption under IRT models,  $g(\theta_1, \theta_2|\mathbf{y})$  can be computed as

$$g(\theta_1, \theta_2|\mathbf{y}) = \frac{L(\theta_1; \mathbf{y}_1)L(\theta_2; \mathbf{y}_2)p(\theta_1, \theta_2)}{\int_{\theta_1=-\infty}^{\theta_1=\infty} \int_{\theta_2=-\infty}^{\theta_2=\infty} L(\theta_1; \mathbf{y}_1)L(\theta_2; \mathbf{y}_2)p(\theta_1, \theta_2)d\theta_1d\theta_2} \tag{8}$$

According to the recommendations of Gelman et al. (2014, p. 95), Robert (2007, p. 226), and Stern (2005), a direct measure of the scientific evidence in favor of a significant score difference can be obtained from the posterior probability  $P(\theta_2 \geq \theta_1|\mathbf{y})$ , which, from Equation 8, can be computed as

$$\begin{aligned} P(\theta_2 \geq \theta_1|\mathbf{y}) &= \int_{\theta_1=-\infty}^{\theta_1=\infty} \int_{\theta_2=\theta_1}^{\theta_2=\infty} g(\theta_1, \theta_2|\mathbf{y})d\theta_1d\theta_2, \\ &= \frac{\int_{\theta_1=-\infty}^{\theta_1=\infty} \int_{\theta_2=\theta_1}^{\theta_2=\infty} L(\theta_1; \mathbf{y}_1)L(\theta_2; \mathbf{y}_2)p(\theta_1, \theta_2)d\theta_1d\theta_2}{\int_{\theta_1=-\infty}^{\theta_1=\infty} \int_{\theta_2=-\infty}^{\theta_2=\infty} L(\theta_1; \mathbf{y}_1)L(\theta_2; \mathbf{y}_2)p(\theta_1, \theta_2)d\theta_1d\theta_2} \end{aligned} \tag{9}$$

In Equation 9, the integrands in the numerator and denominator are the same, but the limits of integration are different.

The integrals in Equation 9 do not have closed forms—so one has to perform numerical integration to compute them. For example, the numerator in Equation 9 can be approximated using simple Riemann approximation (e.g., Thisted, 1988, p. 262), as

$$\begin{aligned} &\int_{\theta_1=-\infty}^{\theta_1=\infty} \int_{\theta_2=\theta_1}^{\theta_2=\infty} L(\theta_1; \mathbf{y}_1)L(\theta_2; \mathbf{y}_2)p(\theta_1, \theta_2)d\theta_1d\theta_2 \\ &\approx \sum_{k=1}^K \sum_{\substack{m=1 \\ \theta_{2m} > \theta_{1k}}}^M L(\theta_{1k}; \mathbf{y}_1)L(\theta_{2m}; \mathbf{y}_2)p(\theta_{1k}, \theta_{2m})\Delta_1\Delta_2, \end{aligned} \tag{10}$$

where  $\theta_{11}, \theta_{12}, \dots, \theta_{1k}, \dots, \theta_{1K}$  is a grid of  $K$  equispaced points,  $\theta_{21}, \theta_{22}, \dots, \theta_{2m}, \dots, \theta_{2M}$  is a grid of  $M$  equispaced points,  $\Delta_1 = \theta_{1,k+1} - \theta_{1k}$ , and  $\Delta_2 = \theta_{2,m+1} - \theta_{2m}$ . In the simulation study and real data example discussed later, we used 101 equispaced points between  $-5$  and  $5$  as  $\theta_{1k}$ s and  $\theta_{2m}$ s to perform the numerical integrations.

The Choice of the Prior Distribution on  $\theta_1$  and  $\theta_2$

We constructed the joint prior distribution on  $\theta_1$  and  $\theta_2$ ,  $p(\theta_1, \theta_2)$ , as the product of  $p(\theta_1)$ , the prior distribution on  $\theta_1$ , and  $p(\theta_2|\theta_1)$ , the prior distribution of  $\theta_2$  given  $\theta_1$ . We assumed that  $\theta_1$ , which, for example, reflects the performance



of an examinee under no test fraud, follows the standard normal distribution a priori, that is,

$$p(\theta_1) = \phi(\theta_1), \tag{11}$$

where  $\phi(\theta_1) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{\theta_1^2}{2}}$ . The conditional prior distribution  $p(\theta_2|\theta_1)$  in an investigation of score differencing, especially to detect test fraud, should ideally incorporate the fact that  $\theta_2$  is considerably larger than  $\theta_1$  for the cheaters. Therefore, we assumed that for the cheaters,  $\theta_2$ , given  $\theta_1$  follows a normal distribution with the mean of  $\theta_1 + \mu$  and standard deviation (SD) of  $\sigma_c$ , and is truncated to the left at  $\theta_1$  a priori; that is, for the cheaters,

$$p(\theta_2|\theta_1) = \frac{k}{\sigma_c} \phi\left(\frac{\theta_2 - (\theta_1 + \mu)}{\sigma_c}\right) I(\theta_2 > \theta_1), \tag{12}$$

where

$$k = \left[ \int_{\theta_2=\theta_1}^{\infty} \frac{1}{\sigma_c} \phi\left(\frac{\theta_2 - (\theta_1 + \mu)}{\sigma_c}\right) d\theta_2 \right]^{-1},$$

and  $I(\theta_2 > \theta_1)$  is equal to 1 if  $\theta_2 > \theta_1$  and zero otherwise, and  $\mu$  is a large positive number. In addition, given the earlier framing of the score differencing problem, in which the null hypothesis was stated as  $H_0 : \theta_1 = \theta_2$ , one may be tempted to make the assumption that a priori,  $\theta_2$  is equal to  $\theta_1$  for the noncheaters. However, we avoided making the assumption because, under a Bayesian framework,  $\theta_1$  and  $\theta_2$  are continuous random variables so that the probability is 0 that  $\theta_1 = \theta_2$ . Instead, we make the assumption that for the noncheaters, given  $\theta_1$ ,  $\theta_2$  is not exactly equal to, but is practically equal to,  $\theta_1$ , or,  $\theta_2 - \theta_1$  is not exactly equal to zero but is practically equal to zero a priori. Specifically, we assume that for the noncheaters, given  $\theta_1$ ,  $\theta_2 - \theta_1$  follows a normal distribution with mean of 0 and SD of  $\sigma_{nc}$  a priori, or

$$p(\theta_2|\theta_1) = \frac{1}{\sigma_{nc}} \phi\left(\frac{\theta_2 - \theta_1}{\sigma_{nc}}\right). \tag{13}$$

The range of values of  $\theta_2 - \theta_1$  over which the distribution provided by Equation 13 has nonnegligible mass represents the values that we think are practically equivalent to zero. The concept of *practically equal to* is borrowed from the concept underlying the *region of practical equivalence* (ROPE), *indifference zone*, and *region of equivalence* (e.g., Carlin & Louis, 2008; Kruschke, 2018). Each of ROPE, indifference zone, and region of equivalence refers to a range of parameter values that are practically equivalent. For example, as Carlin and Louis (2008) described, if one is testing the null hypothesis that the difference between the mean for a treatment and a placebo is zero, then one typically does not care whether to use the treatment or placebo if the difference in the means

falls in the ROPE or indifference zone or region of equivalence that is of the form  $(-\epsilon, \epsilon)$ .

Information provided by Equations 12 and 13 (for the cheaters and noncheaters, respectively) was utilized by assuming that the prior distribution of  $\theta_2$ , given  $\theta_1$  is a mixture of two normal distributions; the first is a normal distribution with mean  $\theta_1$  and  $SD \sigma_{nc}$ , and the second is a normal distribution with mean  $\theta_1 + \mu$  and  $SD \sigma_c$ , truncated below at  $\theta_1$ . The first and second components of the mixture, respectively, represent the distribution of  $\theta_2$ , given  $\theta_1$  for a noncheater and a cheater. The joint prior distribution of  $\theta_1$  and  $\theta_2$  is therefore given by

$$p(\theta_1, \theta_2) = \phi(\theta_1) \left[ \tau \frac{1}{\sigma_{nc}} \phi\left(\frac{\theta_2 - \theta_1}{\sigma_{nc}}\right) + (1 - \tau) \frac{k}{\sigma_c} \phi\left(\frac{\theta_2 - (\theta_1 + \mu)}{\sigma_c}\right) I(\theta_2 > \theta_1) \right], \tag{14}$$

where  $\tau$  represents the weight provided to the first component of the mixture. The value of  $\tau$  should represent the investigator’s belief about the percentage of noncheaters in the sample.

The top panel of Figure 1 shows the densities of the two components (provided by Equations 12 and 13) of the mixture in terms of the difference  $\theta_2 - \theta_1$ . The bottom panel of the figure shows the kernel density estimates of the distribution of a sample of 5,000 values of  $\theta_2 - \theta_1$  simulated from the abovementioned mixture prior distribution (dashed line) and the values of  $\hat{\theta}_2 - \hat{\theta}_1$  for the real data set analyzed later in this article (solid line). The values of  $\tau$ ,  $\mu$ ,  $\sigma_{nc}$ , and  $\sigma_c$  in Equation 14 were set equal to 0.95, 2.0, 0.5, and 0.5, respectively, in the computations leading to Figure 1 that was created using the R function “density” (e.g., R Core Team, 2019). The closeness of the two curves in the bottom panel indicates that the prior distribution reflects reality accurately.<sup>1</sup> The solid line in the bottom panel in Figure 1 also indicates that  $\theta_1$  is unlikely to be exactly equal to  $\theta_2$  even for noncheaters (if it were equal, the corresponding density would have had a sharp spike at 0), which lends support to the distribution assumed in Equation 13.<sup>2</sup> In the rest of the article, the prior distribution is assumed to be the one given by Equation 14, with  $\tau = 0.95$ ,  $\mu = 2.0$ ,  $\sigma_{nc} = 0.5$ , and  $\sigma_c = 0.5$ . Appendix A includes a small simulation study to examine the sensitivity of the posterior probability and the Bayes factor to various choices of  $\tau$  and  $\mu$  in Equation 14 while fixing  $\sigma_{nc} = 0.5$  and  $\sigma_c = 0.5$ —the results in the appendix indicate that the choices of these two constants have only a small effect on the two statistics.

### *The Choice of an Appropriate Cutoff for the Posterior Probability*

To use the posterior probability in score differencing, one needs an appropriate cutoff value so that the individuals with values of the posterior probability above this cutoff can be considered to have a statistically significant score difference. The choice of the cutoff should ideally be guided by (Bayesian) decision

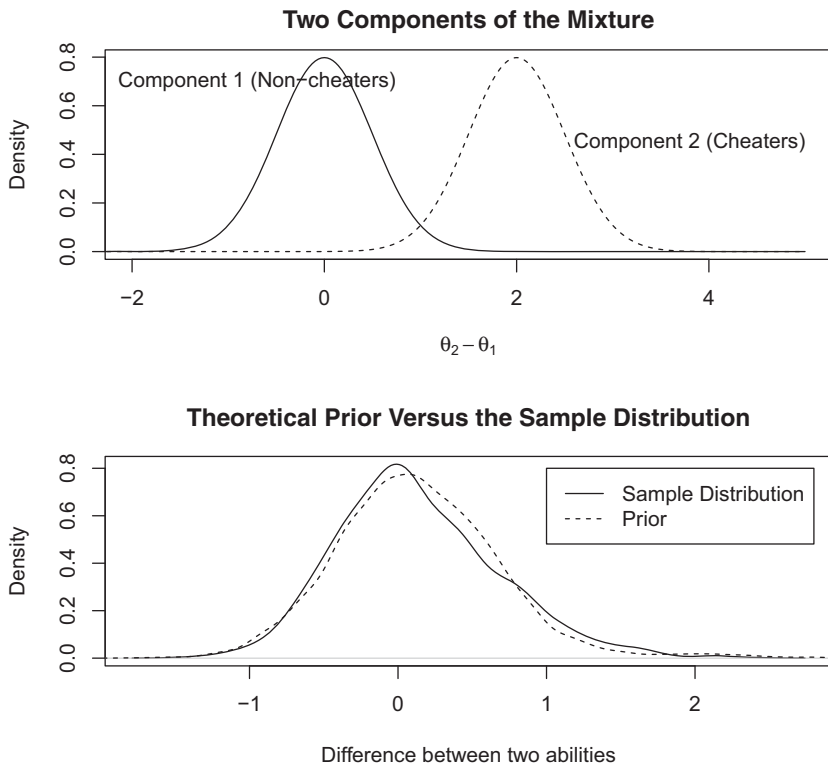


FIGURE 1. *The components of the prior distribution and the theoretical prior versus the empirical values.*

theory (e.g., Robert, 2007, p. 51) and the specific application at hand. For example, Robert (2007, p. 225) and Johnson and Sinharay (2016, p. 249) noted that to apply Bayesian decision theory to a hypothesis-testing problem, one should assign losses of  $c_1$  and  $c_2$  to false positive and false negative (or Type I and Type II) errors. Then, one should minimize the “posterior expected loss” to obtain the “Bayes rule” or “Bayes estimator,” which, in our context, is given by

$$\text{Reject the null hypothesis if } P(\theta_2 \geq \theta_1 | y) > \frac{c_1}{c_1 + c_2} = \frac{\frac{c_1}{c_2}}{\frac{c_1}{c_2} + 1}.$$

Therefore, in an application, the choice of the cutoff for the posterior probability would ideally depend on  $\frac{c_1}{c_2}$ , which is the comparative severity of the false positive and false negative errors. For example, if the test administrators think that a false positive error is 19 times as costly as a false negative error, then the cutoff

would be .95, whereas if the test administrators think that a false positive error is 99 times as costly as a false negative error, then the cutoff would be .99. Given the observation by, for example, Wollack et al. (2015) that methods for detection of test fraud are typically applied with conservative levels, it is more likely that a large value of  $\frac{c_1}{c_2}$  would be used in determining a cutoff for the posterior probability.

*Reconcilability of Evidence From Posterior Probability and Frequentist Approaches*

Berger and Sellke (1987), Casella and Berger (1987), and Pratt (1965) discussed the issue of reconcilability (or the lack of it) of frequentist test statistics/ $p$  values and posterior probabilities for testing against one-sided alternative hypotheses and two-sided alternative hypotheses. Berger and Sellke (1987) showed that frequentist  $p$  values and posterior probabilities are usually irreconcilable, or appear to provide different extents of evidence for the same data set, when the alternative hypothesis is two sided, that is, of the type  $H_1 : \theta \neq 0$ . In contrast, Casella and Berger (1987) and Pratt (1965) showed that the two probabilities are often reconcilable when the alternative hypothesis is one-sided and of the type  $H_1 : \theta > 0$ . In the context of score differencing, the findings of Berger and Sellke (1987), Casella and Berger (1987), and Pratt (1965) imply that a posterior probability would be reconcilable with the (frequentist)  $L_S$  statistic because the alternative hypothesis underlying  $L_S$  is  $H_1 : \theta_2 > \theta_1$ , which is a one-sided alternative hypothesis. Therefore, our suggested posterior probabilities are expected to provide evidence that is mostly reconcilable with the evidence provided by the  $p$  value corresponding to the  $L_S$  statistic.

*A Simple Illustration*

Consider a test with 20 items. Let us consider that the Rasch model fits the data from the test and that the estimated 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 seven examinees all of whom obtain a raw score of 3 on the first 10 items on test but obtained raw scores of 3, 4, 5, 6, 7, 8, and 9 on the last 10 items on the test.

Table 3 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, the  $p$  value for the SLR statistic ( $p$  value), the predictive  $p$  value (PrP), the Bayes factor given by Equation 5 (BF), and the posterior probability given by Equation 9 (PP) for the examinees. The R code for computing the posterior probability for Examinee 1 in Table 3 is provided in Appendix B.

TABLE 3.  
*Results for Seven Examinees*

Examinee	Score Diff.	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}$	SLR	<i>p</i> Value	PrP	BF	PP
1	0	-.76	-0.76	-.80	0.00	.50	.60	0.71	.44
2	1	-.76	-0.37	-.59	0.45	.32	.45	0.82	.54
3	2	-.76	0.0	-.39	0.91	.18	.28	1.10	.65
4	3	-.76	0.37	-.19	1.35	.09	.19	2.30	.74
5	4	-.76	0.76	.00	1.81	.04	.07	2.44	.83
6	5	-.76	1.22	.19	2.29	.01	.04	4.13	.90
7	6	-.76	1.85	.39	2.84	.00	.01	9.74	.96

*Note.* Score Diff. = difference in the raw score; *p* Value = *p* value for the SLR statistic; PrP = predictive *p* value; BF = Bayes factor; PP = posterior probability.

As the score difference (shown in Column 2 of Table 3) increases, the methods are expected to find stronger evidence in favor of a large score difference. So, it is not surprising that each statistic provides strong evidence of a significant score difference in the bottom rows of the table. One using the SLR statistic would not reject the null hypothesis of no performance difference between the two halves of the test for Examinees 1 through 4 and would reject the null hypothesis for Examinees 5 through 7 at 5% level. If one uses the cutoff of .95 for the posterior probability, then one would conclude that there is no evidence of a performance difference for Examinees 1–6 and some evidence of a performance difference for Examinee 7. Thus, the SLR statistic (or, equivalently, the frequentist *p* value) and the posterior probability may lead to different conclusions for some examinees.

### Simulations

Simulated data that involved different extents of score differences were used to compare the properties of the posterior probability to those of three existing approaches for score differencing. It was assumed in the simulations that the score differences originated from preknowledge of compromised items.

#### *Design*

All simulations involved a nonadaptive assessment that includes 100 dichotomous items. The true item parameters were randomly drawn from the estimated item parameters of the item pool of one subject of a state test.<sup>3</sup> The true abilities of the examinees were simulated from a standard normal distribution.

The following two factors were varied in the simulations:

- the number of compromised items (10, 20, or 30 items)<sup>4</sup>; for each simulated data set, the compromised items were randomly selected out of the 100 items and

- the number of examinees who had item preknowledge (the *cheaters*) as a percentage of those who did not have preknowledge (5%, 10%, or 20%).

The simulation factors were crossed with each other. Thus, the number of simulation conditions was nine. For each simulation condition, 100 data sets were simulated; the number of noncheaters in each data set was 2,000 so that the number of cheaters in a data set was 100, 200, or 400 in the various simulation conditions. The item scores of the noncheaters (or those without item preknowledge) on all items and of the cheaters (those with item preknowledge) on the uncompromised items were simulated from the 2PLM. The item scores of each cheater on a compromised item was simulated using the 2PLM but using a value of ability that is obtained by adding 2.0 on the  $\theta$  scale to the true ability of the cheater or by shifting the ability of the cheater to the right by 2.0. Item response data under aberrant responding has been simulated after shifting the examinee ability (or a “ $\theta$  shift”) by researchers such as Glas and Dagohoy (2007). The simulation of item scores after a  $\theta$  shift recreates the scenario that item preknowledge leads to a boost in the ability.

For each simulated data set, the following computational steps were performed:

1. Compute the estimated item parameters using the marginal maximum likelihood estimation procedure.
2. For each examinee, compute the SLR statistic. The MLE of ability, restricted to the range  $-4.0$  and  $4.0$ , was used to compute the SLR statistic. The item parameters computed in the previous step were used in these calculations.
3. For each examinee, compute the Bayes factor (Sinharay & Johnson, 2020), posterior probability, and the predictive  $p$  value (Wang et al., 2017) using Equations 6, 9, and 7, respectively.

For each simulation condition, the values of the four statistics over the 100 simulated data sets were used to compare their performances.

### Results

Figure 2 shows a scatterplot of the posterior probability ( $y$ -axis) versus  $1 - p$  value for the SLR statistic ( $x$ -axis) for the examinees in the simulation case with 30 compromised items and 10% examinees with preknowledge. We decided to plot the posterior probability versus  $1 - p$  value because an increasing value of each of these statistics indicates an increasing score difference. Each circle in the plot corresponds to one examinee. The gray circles correspond to the examinees who are true noncheaters, and the black circles correspond to the examinees who are true cheaters. The figure includes horizontal and vertical dashed lines representing cutoffs of .95 for the two statistics and also includes a diagonal line. The cutoff of .95 was used for  $1 - p$  value because a  $p$  value smaller than .05 is equivalent to  $1 - p$  value being larger than .95. The two plotted quantities seem

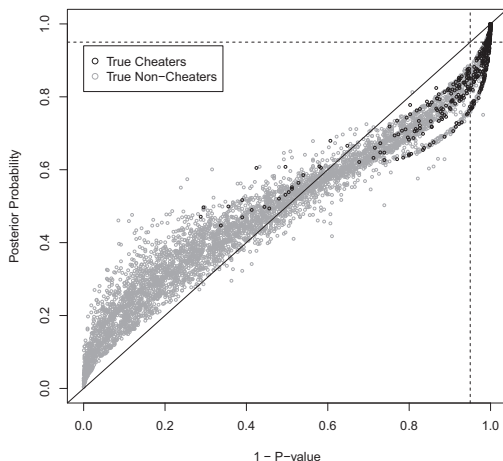


FIGURE 2. A scatterplot of  $1 - p$  value versus posterior probability for the simulations.

to be good agreement. Both are mostly smaller than the cutoff for the true noncheaters, and both are often larger than the cutoff for the true cheaters. This agreement is not a surprise, given the finding of Casella and Berger (1987) of reconcilability of evidence from posterior probabilities and frequentist  $p$  values for testing a one-sided null hypothesis. The posterior probability shows a tendency to be smaller than  $1 - p$  value for both the true cheaters and true noncheaters in the right side of the figure and larger than  $1 - p$  value in the left side of figure, which means that if the same cutoff (of, say, .95 or .99) is used for both, then the use of the posterior probability (rather than the frequentist statistic) will lead to a more conservative detection of item preknowledge.

The comparison of the power of statistics for detecting aberrant examinees has been performed using receiver operating characteristics (ROC) curves at least since Drasgow et al. (1985). Given the values of a statistic (whose larger value indicates more aberrance) from a data set for which the identities of the true aberrant and nonaberrant examinees are known, an ROC curve requires the computation of the following two quantities for several values of  $y$ :

- the false alarm rate (or “false positive rate” or “Type I error rate”),  $F(y)$ , which is the proportion of times when the statistic for a nonaberrant examinee is larger than  $y$  and
- the hit rate (or “true positive rate” or “power”),  $H(y)$ , which is the proportion of times when the statistic for an aberrant examinee is larger than  $y$ .

Then, a graphical plot is created in which  $F(y)$  is plotted along the  $x$ -axis,  $H(y)$  is plotted along the  $y$ -axis, and a line joins  $\{F(y), H(y)\}$  for several values of  $y$ .

These lines together constitute the *ROC curve*. Appendix C shows the ROC curve from one condition of the simulation study.

The area under the ROC Curve (AUROC; e.g., Sinharay, 2017b) of a statistic is a measure of how powerful the statistic is. In the context of detecting aberrant examinees, researchers such as Sinharay (2017b) used *truncated ROC areas*, or areas under the ROC curves truncated between 0 and .1 and divided by .10—that is because false positive rates larger than .10 are hardly employed in the context of detecting aberrant examinees (Wollack et al. 2015). The truncated ROC area of a very powerful statistic is expected to be close to 1. The truncated ROC areas of all the statistics were computed for all the simulation conditions.

When the number of compromised items was fixed, the truncated ROC area of the statistics was not affected by the percentage of examinees benefiting from preknowledge—so the truncated ROC areas were averaged over the three levels of this percentage. The average truncated ROC areas of the statistics for the various number of compromised items are shown in Figure 3. In the figure, the *x*-axis represents the number of compromised items and the *y*-axis represents the average truncated ROC area. The average truncated ROC area for the posterior probability, SLR statistic, Bayes factor, and predictive checking are joined by a solid line, dashed line, dotted line, and a dotted dashed line, respectively. The figure shows that the average truncated ROC area increases as the number of compromised items increases.

The average truncated ROC areas of the four statistics are very close for any given number compromised items, all lying in a narrow interval of width about .02. The average areas of the posterior probability are the largest by a small margin followed by that of the SLR statistic. The average truncated ROC area of the SLR statistic is the largest among the four statistics for 10 compromised items but close to the smallest for 30 compromised items. The average truncated ROC areas of the posterior probability, SLR statistic, Bayes factor, and predictive checking method, averaged over all simulation cases, are .87, .86, .86, and .85, respectively.

Note that the comparative performance of the approaches was very similar (results not reported) in another set of simulations that were very similar to the above except that the item parameters were not estimated in each iteration.

## **Real Data Example**

### *Data*

We analyzed item response data from one form of a nonadaptive licensure assessment. The source of the data set is Cizek and Wollack (2017, p. 14). Researchers such as Sinharay (2017a), Sinharay and Jensen (2019), and Zoplouglu (2017) analyzed the same data set to detect various types of test fraud. The test form comprises 170 operational items that are dichotomously scored. The sample size for the form is 1,644. A total of 61 items on the form were identified



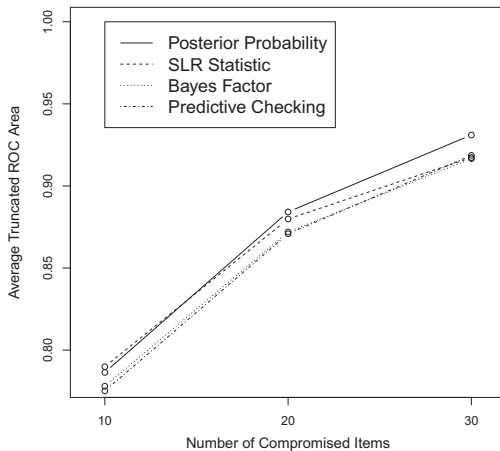


FIGURE 3. Average truncated receiver operating characteristics areas for the four statistics for the simulation study.

as compromised by the organization that provided the data. In addition, 48 examinees were flagged by the organization 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, it is reasonable to treat these examinees as true cheaters. As in Sinharay (2017a) and Sinharay and Jensen (2019), the interest here is in detecting item preknowledge, that is, detecting the examinees who may have benefited from the preknowledge of the 61 compromised items.

### Analysis and Results

Although the Rasch model is operationally used in the assessment, the 2PLM was found to fit the data better and was used for the analysis here. The item parameters were estimated using the marginal maximum likelihood estimation procedure from the data set using the R package *ltm* Version 1.1-1 (Rizopoulos, 2006) and were used in the computation of the SLR statistic and the posterior probability. We then computed the values of the SLR statistic, Bayes factor, predictive probability, and posterior probability for each examinee in the data set. The MLEs of the abilities, truncated between  $-4$  and  $4$ , were used to compute the SLR statistic. To perform score differencing, the first set of items ( $\mathcal{I}_1$ ) comprised the set of 109 noncompromised items and the second set of items ( $\mathcal{I}_2$ ) comprised the set of 61 compromised items.

Figure 4 shows a scatterplot of the posterior probability versus  $1 -$  the  $p$  value for the SLR statistic for the examinees in the data set. The black and gray circles correspond to examinees who were flagged (48 of them) and not flagged (1,596

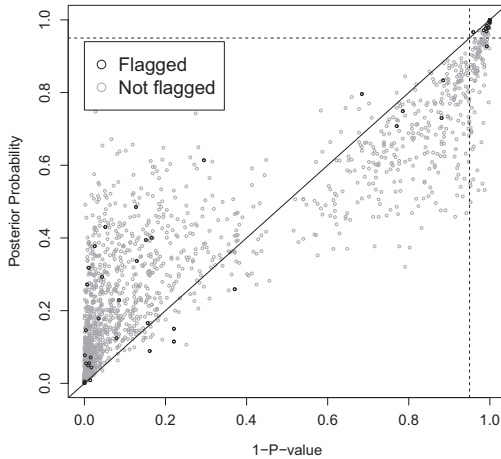


FIGURE 4. A scatterplot of  $1 - p$  value versus posterior probability for the real data example.

of them), respectively, by the licensure organization. The figure includes horizontal and vertical dashed lines representing cutoffs of .95 for the two statistics. A diagonal line is also provided. The two plotted quantities are mostly in agreement with each other; that is, the posterior probability is mostly large for the examinees for whom the  $p$  value is small. As in Figure 2, there is a tendency for the posterior probability to be smaller than  $1 - p$  value in the right side of Figure 4 and larger than  $1 - p$  value in the left side of Figure 4. Also, among the flagged examinees, the posterior probability is larger than .95 whenever the  $p$  value is smaller than .05 (or  $1 - p$  value is larger than .95) except for one examinee. However, an interesting pattern is visible toward the right of the plot. All of the gray circles to the right of the vertical dashed line and below the horizontal dashed line belong to examinees who are not flagged by the licensure organization, but the frequentist  $p$  value is smaller than .05 and the posterior probability is smaller than .95 for them. Thus, a frequentist using a  $p$  value at 5% level would conclude that these examinees benefited from item preknowledge while a Bayesian using a posterior probability with a cutoff of .95 would not.

It is possible to draw an ROC curve and compute the truncated ROC areas for the statistics for the licensure data sets by treating the flagged and nonflagged examinees as true cheaters and noncheaters, respectively. These areas for the posterior probability, Bayes factor, SLR statistic, and predictive checking were .62, .61, .60, and .60, respectively, so that the area for the SLR statistic is slightly larger than those for the other statistics for the data set. Figure 5 shows the ROC curves for the posterior probability and SLR statistic, truncated to show the false alarm rates between 0 and .10.

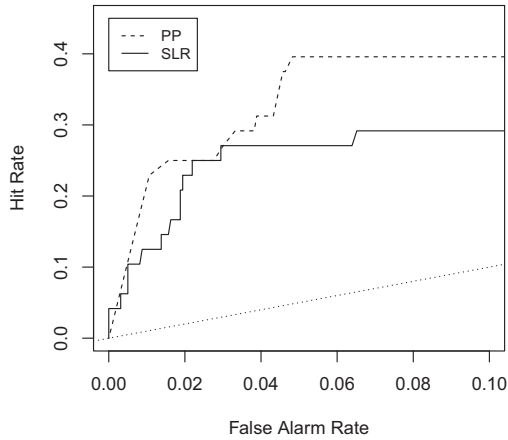


FIGURE 5. The receiver operating characteristics curve for the posterior probability (PP) and the signed likelihood ratio (SLR) statistic for the real data example.

### Conclusions

In this article, we suggested posterior probabilities as an alternative tool for score differencing (Wollack & Schoenig, 2018). These probabilities are less likely to be misinterpreted than frequentist  $p$  values and are intended to provide direct evidence in favor of a significant score difference. In a simulation study and in a real data application, the posterior probability was found to have a slightly larger AUROC curve compared to several existing approaches.

Although only nonadaptive tests were considered in the simulations and the real data example of this article, the posterior probability can be computed for adaptive tests as well. Sinharay (2017a) discussed how to apply the  $L_S$  statistic for adaptive tests—the application involved the computation of the likelihood over a set of items received by an examinee. Once the likelihood is computed, Equation 9 can be applied to compute the posterior probability for an adaptive test when, for example, a subset of items administered to an examinee is found to have been compromised. However, the number of compromised items that each examinee receives on an adaptive test will most often be very small (possibly with the exception of multistage tests where one or more first- or second-stage unit/module was compromised) and the posterior probability will not be a powerful tool for score differencing for adaptive tests.

While the posterior probability has the natural interpretation of being a probability and is bounded between 0 and 1, it is possible to convert it to *posterior odds* that is given by

$$\text{Posterior odds} = \frac{\text{Posterior probability}}{1 - \text{posterior probability}},$$

and use the posterior odds instead of the posterior probability (e.g., Edwards et al., 1963). For example, the posterior odds are equal to 1, 9, 19, and 99, when the posterior probability is equal to .5, .9, .95, and .99, respectively.

There exist several approaches that are somewhat similar to the posterior probability suggested in this article. 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, but the computation was predicated on a specialized IRT model that applies only to fraudulent erasures and cannot be easily extended to score differencing in general. The posterior probability of answer copying, suggested by Allen and Ghattas (2016), is conceptually similar to the posterior probability suggested in this article but cannot be used for score differencing.<sup>5</sup> Skorupski and Wainer (2017) suggested the posterior probability of cheating (PPoC) of an individual as  $P(C|T \geq t)$ , where  $C$  is the event that the examinee is a true cheater,  $T$  is the random variable corresponding to the test statistic, and  $t$  denotes the value of  $T$  for the individual. The PPoC is conceptually similar to the posterior probability suggested in this article. Skorupski and Wainer (2017) showed that the PPoC can be expressed as

$$\text{PPoC} = 1 - \frac{\text{Frequentist p-value} \times P(\text{non-cheater})}{P(T \geq t)},$$

where  $P(\text{noncheater})$  is the prior probability of noncheaters in the population. Table 18.4 of Skorupski and Wainer (2017) showed that the choice of  $P(\text{noncheater})$  may be fairly influential on the PPoC. In contrast, the posterior probability suggested in this article is less dependent on the prior distributions. The quantity  $\tau$  in the prior distribution in this article is like  $P(\text{noncheater})$  in the expression of PPoC, but Appendix A of this article shows that the extent of sensitivity of the posterior probability to  $\tau$  is considerably smaller than that of PPoC on  $P(\text{noncheater})$ .

The new approach can be applied only in the context of one set of statistical methods (score differencing) out of six mentioned by Wollack and Schoenig (2018). In addition, the approach can be used to detect preknowledge only when the set of compromised items is known as in the real example discussed earlier. The new approach should not be used as the sole source evidence of test fraud in operational testing. Instead, as recommended by, for example, Hanson et al. (1987) and Holland (1996), the new approach should be employed as a part of quality control and/or as secondary evidence, along with other statistics and nonstatistical evidence, in investigations of test fraud.

Although the research reported in this article represents one of the first applications of Bayesian methods to score differencing, the article has several limitations and it is possible to extend the 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 other (potentially new) Bayesian methods for score differencing. Third, although the results in Appendix A provide some evidence that the posterior probability is not influenced much by the joint prior distribution on the ability parameters, especially for large  $\mathcal{I}_1$  and  $\mathcal{I}_2$ , it is possible to perform a more detailed examination of the influence of the joint prior distribution on the posterior probability. Fourth, it is possible to extend the approach to utilize both item scores and response times on computerized tests; such an approach would involve an examination of whether the examinees perform better and faster on a subset of items and may be more powerful than one based only on item scores. Similarly, it is possible to extend this approach to multivariate ability—the computations are likely to be more involved, especially in the presence of within-item multidimensionality. Fifth, it is possible to perform more research on the choice of an appropriate cutoff for the posterior probability. Sixth, although a simple Riemann approximation was used to approximate the integrals in Equation 9, it is possible to explore the use of other numerical integration approaches (e.g., Givens & Hoeting, 2013, pp. 129–195). Finally, although this article focuses on  $P(\theta_2 \geq \theta_1 | \mathbf{y})$ , a more direct measure of the scientific evidence in favor of a significant score difference would be  $P(C | \mathbf{y}, \mathbf{Z})$  where  $C$  is the event that the examinee is a true cheater and  $\mathbf{Z}$  quantifies other information (like test center information, proctor report etc.). However, the computation of  $P(C | \mathbf{y}, \mathbf{Z})$  would be extremely difficult as argued by experts such as Holland (1996)—one reason of the difficulty is the lack of, for example, a statistical model for the behavior of an examinee who commits test fraud. So, we focus on  $P(\theta_2 \geq \theta_1 | \mathbf{y})$  rather than  $P(C | \mathbf{y}, \mathbf{Z})$  in this article and believe to have demonstrated that  $P(\theta_2 \geq \theta_1 | \mathbf{y})$  may provide useful evidence regarding test fraud and should be considered for inclusion in the practitioner’s toolkit for detecting test fraud.

## Appendix A

---

### *Sensitivity to the Prior Distributions*

To examine the sensitivity of the Bayes factor and the posterior probability to the prior distribution, we computed these two quantities for five different test length conditions. The size of  $\mathcal{I}_1$  was 10, 20, 40, 40, and 40 in the five conditions while that of  $\mathcal{I}_2$  was 10, 20, 10, 20, and 40. Thus, the total number of items on the test is 20, 40, 50, 60, and 80 in the five cases. The set  $\mathcal{I}_1$  included the first several items in all the cases.

The Rasch model was assumed to hold with all item difficulties being equal to 0. The scores  $y_i, i = 1, 2, \dots, I$  were set so that the raw scores on  $\mathcal{I}_1$  and  $\mathcal{I}_2$  in

TABLE A1.  
The Bayes Factor and Posterior Probability for Different Prior Distributions

Values of $\tau$ and $\mu$	Sizes of $\mathcal{I}_1$ and $\mathcal{I}_2$									
	(10, 10)		(20, 20)		(40, 10)		(40, 20)		(40, 40)	
	BF	PP	BF	PP	BF	PP	BF	PP	BF	PP
.9, 1	1.84	.74	6.40	.89	1.36	.79	5.09	.92	46.3	.97
.9, 2	1.69	.72	5.78	.88	1.24	.77	4.51	.92	38.7	.97
.95, 1	1.75	.72	5.88	.88	1.29	.77	4.63	.91	41.1	.97
.95, 2	1.68	.71	5.56	.87	1.23	.76	4.34	.91	37.3	.96
.99, 1	1.68	.70	5.45	.86	1.24	.75	4.27	.90	36.9	.96
.99, 2	1.66	.70	5.39	.86	1.23	.75	4.21	.90	36.2	.96

Note. BF = Bayes factor; PP = posterior probability.

the five test length conditions were (5, 7), (10, 15), (20, 7), (20, 15), and (20, 30), respectively. Thus, there were score differences of various extent in all the cases. The frequentist  $p$  values for the five test length conditions do not depend on the joint prior distribution for  $\theta_1$  and  $\theta_2$  and were equal to .18, .05, .13, .03, and .01, respectively. Six joint prior distributions of  $\theta_1$  and  $\theta_2$ , all special cases of Equation 14, were considered, with the values of  $\sigma_{nc}$  and  $\sigma_c$  set equal to .5, and the values of  $\tau$  and  $\mu$  given by (a) .9 and 1.0, (b) .9 and 2.0, (c) .95 and 1.0, (d) .95 and 2.0, (e) .99 and 1.0, and (f) .99 and 2.0.

Table A1 shows the values of the Bayes factor and posterior probability for all the above-mentioned prior distributions for all test length conditions. Each row of the table shows the values of these two statistics for one prior distribution for the five test length conditions. Table A1 shows that the joint prior distribution has a small effect on Bayes factor and posterior probability, with both statistics becoming more conservative as either of  $\tau$  or  $\mu$  increases. For the fifth test length condition (that involves the largest  $\mathcal{I}_1$  and  $\mathcal{I}_2$ ), the joint prior distribution has a very small effect on Bayes factor and posterior probability and especially on the posterior probability. This finding implies that the posterior probability is not likely to be influenced much by the prior distribution for large  $\mathcal{I}_1$  and  $\mathcal{I}_2$ , which is the case of the real data example in this article.

## Appendix B

### The R Code Used in the Illustration

The R code for computing the posterior probability for Examinee 1 in Table 3 is provided below. The code makes use of the R function *integral2* in the R package *pracma* (Borchers, 2019).

```

library(pracma)
pr2pl=function(t,a,b){return(1/(1+exp(a*(b-t))))}
logpr1=function(u,t,a,b) {p=pr2pl(t,a,b)
                        return(ifelse(u==1,log(p),log(1-p)))}
u=rep(c(rep(1,3),rep(0,7)),2)
a=rep(1,20)
b=rep(0,20)
s1=1:10
Joint=function(t1,t2){LL=0
for (j in s1)
  {LL=LL+logpr1(u[j],t1,a[j],b[j])}
for (j in setdiff(1:length(u),s1))
  {LL=LL+logpr1(u[j],t2,a[j],b[j])}
p2=0.95*dnorm(t2,mean=t1,sd=0.5)+0.05*dnorm(t2,mean=t1+2,sd=0.5)*ifelse(t2>t1,1,0)
return(exp(LL)*dnorm(t1,mean=0,sd=1)*p2)}
min=-5
max=5
ymin=function(t1) t1
num=integral2(Joint,min,max,ymin,max,vectorized=FALSE)
den=integral2(Joint,min,max,min,max,vectorized=FALSE)
PostProb=num$Q/den$Q

```

### Appendix C

#### An Example ROC Curve

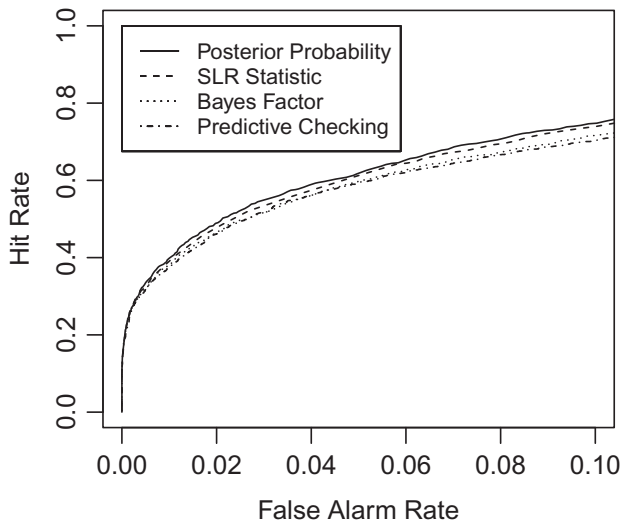


FIGURE C1. The receiver operating characteristics curve for the four statistics for one simulation case.

Figure C1 shows the ROC curve for the posterior probability (solid line), signed likelihood ratio (SLR) statistic (dashed line), Bayes factor (dotted line), and the

predictive checking method (dotted dashed line) for the simulation case with 10 compromised items and 5% examinees having preknowledge. The curve is truncated between the values of 0 and .01 of the false alarm rate ( $x$ -axis). The curve for the posterior probability is the highest, followed by that of the SLR statistic.

### **Acknowledgments**

The authors wish to express sincere appreciation and gratitude to Steven Culpepper, the editor, and the three anonymous reviewers for their helpful comments. The authors would also like to thank Rebecca Zwick, John Donoghue, and Bingchen Liu for their helpful comments on an earlier version and would like to thank Andrew Gelman for his helpful advice on the choice of the prior distribution used in this article.


### **Declaration of Conflicting Interests**


The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The authors prepared the work as employees of Educational Testing Service. Any opinions expressed in this publication are those of the authors and not necessarily of Educational Testing Service or Institute of Education Sciences.

### **Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D170026.

### **ORCID iDs**

Sandip Sinharay  <https://orcid.org/0000-0003-4491-8510>

Matthew S. Johnson  <https://orcid.org/0000-0003-3157-4165>

### **Notes**

1. In similar plots with these values of  $\tau$ ,  $\mu$ ,  $\sigma_{nc}$ , and  $\sigma_c$  for two other test data sets for which a set of items was known to be compromised (these plots are not included in this article and can be obtained from the authors upon request), the prior distribution reflected reality accurately.
2. Although the solid line in the bottom panel of Figure 1 is created from the estimates of  $\theta_1$  and  $\theta_2$ , the estimates are expected to be very close to the corresponding true values given that both  $\mathcal{I}_1$  and  $\mathcal{I}_2$  include a large number of items.
3. The use of two other sets of estimated item parameters and a set of simulated item parameters did not affect the comparative performance of the statistics (results not included here and can be obtained from the authors).
4. Thus, the number of uncompromised items was 90, 80, or 70.
5. Wollack and Schoenig (2018) included the methods to detect answer copying in a separate category than those for 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.
- Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association, 82*, 112.
- Borchers, H. W. (2019). *pracma: Practical numerical math functions* (R package version 2.2.5). Available at <https://CRAN.R-project.org/package=pracma>.
- Carlin, B. P., & Louis, T. A. (2008). *Bayesian methods for data analysis*. Chapman & Hall.
- Casella, G., & Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association, 82*, 106–111.
- Cizek, G. J., & Wollack, J. A. (2017). *Handbook of detecting cheating on tests*. Routledge.
- Cox, D. R. (2006). *Principles of statistical inference*. 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.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193–242.
- 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.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Chapman & Hall.
- Givens, G. H., & Hoeting, J. A. (2013). *Computational statistics*. Wiley.
- Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika, 72*, 159–180.
- 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)*. American College Testing.
- Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support* (ETS Research Report No. RR-94-4). Educational Testing Service.
- Johnson, M. S., & Sinharay, S. (2016). Bayesian estimation. In W. van der Linden (Ed.), *Handbook of item response theory, Volume 2: Statistical tools* (pp. 237–257). Chapman & Hall/CRC.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.
- Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science, 1*, 270–280.

- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercntile observed-score “equatings.” *Applied Psychological Measurement*, 8, 453–461.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Pratt, J. W. (1965). Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society: Series B*, 27, 169–192.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.
- R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria. Available at <https://www.R-project.org/>.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1–25.
- Robert, C. P. (2007). *The Bayesian choice* (2nd ed.). Springer.
- 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.
- Sinharay, S., & Johnson, M. S. (2020). Detecting test fraud using Bayes factors. *Behaviormetrika*, 47, 339–354.
- Skorupski, W. P., & Wainer, H. (2017). The case for Bayesian methods when investigating test fraud. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 214–231). Routledge.
- Stern, H. S. (2005). Model inference or model selection: Discussion of Klugkist, Laudy, and Hoijtink (2005). *Psychological Methods*, 10, 494–499.
- Thisted, R. A. (1988). *Elements of statistical computing: Numerical computation*. Chapman & Hall.
- 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.
- 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.
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA’s statement on p-values: Context, process, and purpose. *The American Statistician*, 70, 129–133.

- 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). Sage.
- Zopluoglu, C. (2017). Similarity, answer copying, and aberrance: Understanding the status quo. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 25–46). Routledge.

### **Authors**

SANDIP SINHARAY is a Distinguished Presidential Appointee at Educational Testing Service, MS-12T, Rosedale Road, Princeton, NJ 08541, USA; email: [ssinharay@ets.org](mailto:ssinharay@ets.org). His research interests include item response theory, assessment of model fit, reporting of subscores, statistical methods for detecting test fraud, and Bayesian methods.

MATTHEW S. JOHNSON is a principal research director at Educational Testing Service, MS-12T, Rosedale Road, Princeton, NJ 08541, USA; email: [msjohnson@ets.org](mailto:msjohnson@ets.org). His research interests include the development of hierarchical models for educational statistics, statistical methods for large-scale assessments such as NAEP, Bayesian statistics, and measures for reliability and validity for complex assessments.

Manuscript received December 12, 2019

First revision received April 2, 2020

Second revision received June 9, 2020

Third revision received August 3, 2020

Accepted August 14, 2020