Which Statistic Should Be Used to Detect Item Preknowledge When the Set of Compromised Items Is Known?

Applied Psychological Measurement

2017, Vol. 41(6) 403–421

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621617698453

journals.sagepub.com/home/apm

Sandip Sinharay¹

Abstract

Benefiting from item preknowledge is a major type of fraudulent behavior during educational assessments. Belov suggested the posterior shift statistic for detection of item preknowledge and showed its performance to be better on average than that of seven other statistics for detection of item preknowledge for a known set of compromised items. Sinharay suggested a statistic based on the likelihood ratio test for detection of item preknowledge; the advantage of the statistic is that its null distribution is known. Results from simulated and real data and adaptive and nonadaptive tests are used to demonstrate that the Type I error rate and power of the statistic based on the likelihood ratio test are very similar to those of the posterior shift statistic. Thus, the statistic based on the likelihood ratio test appears promising in detecting item preknowledge when the set of compromised items is known.

Keywords

cheating, fraud, test security

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

This article considers only the case when the investigator knows which items are compromised; researchers such as Drasgow, Levine, and Zickar (1996), Belov (2016), and Sinharay (2017a) considered this case. Typically, such a case arises when the test administrators become aware after an administration about some items possibly being compromised. Such a case may

¹Pacific Metrics Corporation, Monterey, CA, USA

Corresponding Author:

Sandip Sinharay, Educational Testing Service, 660 Rosedale Rd., Princeton, NJ 08540-2218, USA.

Email: ssinharay@ets.org

also arise when the test administrators have applied a method to detect compromised items (e.g., that suggested by Veerkamp & Glas, 2000) to find several items that may have been compromised. The real data example later in this article demonstrates that such a situation can arise in practice.

Belov (2016) suggested the *posterior shift statistic* (PSS) for detecting item preknowledge. He also compared the performance of the PSS with that of seven other statistics for detection of item preknowledge and found the PSS to perform the best on average. Sinharay (2017a) showed that a statistic based on the likelihood ratio test (LRT; e.g., Cox & Hinkley, 1974; Rao, 1973) performs quite well in detecting item preknowledge. Because Belov (2016) did not consider the statistic suggested by Sinharay (2017a) and Sinharay (2017a) did not consider the PSS, a natural question is “How does the statistic suggested by Sinharay (2017a) perform in comparison to the PSS (Belov, 2016)?” This article attempts to answer that question by comparing the performances of the PSS and the statistic based on the LRT,¹ both for adaptive and nonadaptive tests, using both simulated and real data.

The next section includes some background material including a description of the PSS (Belov, 2016) and the statistic based on the LRT (Sinharay, 2017a). Then, the Type I error rates and power of these two statistics are compared for data simulated from both nonadaptive tests and for CATs. In the “A Real Data Application” section, the two statistics are computed for two real data sets from a licensure test. Conclusions and recommendations are provided in the last section.

As in Belov (2016), this article involves the assumption that the test items have been calibrated using item response theory (IRT), and the item parameters are known (and are equal to the estimates computed from a previous IRT calibration). Furthermore, this article considers only dichotomous items.

Background

Let θ denote the true ability and x_i denote the score of an examinee on item i , $i = 1, 2, \dots, I$, on a test, and let $P_i(x_i|\theta)$ denote the probability of the score x_i on item i for ability θ . For example, if the three-parameter logistic model (3PLM) is used, then

$$P_i(x_i = 1|\theta) = g_i + (1 - g_i) \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \text{ and } P_i(x_i = 0|\theta) = 1 - P_i(1|\theta),$$

where a_i , b_i , and g_i , respectively, denote the slope, difficulty, and guessing parameters of item i . Let c denote the (known) set of items that are compromised and were administered to the examinee. Note that c could vary over the examinees because, for example, on a CAT, an examinee may have been administered only a subset of all compromised items. Let u denote the set of noncompromised items that were administered to the examinee. Let $\mathbf{x}_c$ and $\mathbf{x}_u$, respectively, denote the collection of scores of the examinee on the items in c and u , that is, $\mathbf{x}_c = \{x_i, i \in c\}$ and $\mathbf{x}_u = \{x_i, i \in u\}$.

Review of the PSS

If an examinee is aberrant, or, benefitted from item preknowledge, then he or she is expected to perform significantly better on c than on u , and the posterior distribution of the examinee's ability based on c is expected to be located far to the right of that based on u for that examinee. Belov (2016) suggested the PSS, which quantifies the distance between the two above-mentioned posterior distributions, to detect item preknowledge. To compute the PSS, one first

computes a discrete approximation of the posterior distribution of the examinee's ability given the scores on c over a grid of J equispaced values $\theta_1 > \theta_2 > \dots > \theta_J$ under the assumption that the prior distribution is uniform. Let us denote the difference between two successive values of the grid as h (i.e., $h = \theta_1 - \theta_2$). The probability mass function (pmf) of the discrete approximation of the posterior distribution at θ_j given the scores on the items in c , denoted $F_c(\theta_j)$, is computed as

$$F_c(\theta_j) = \frac{\prod_{i \in c} P_i(x_i | \theta_j)}{\sum_{m=1}^J \prod_{i \in c} P_i(x_i | \theta_m)}, \quad j = 1, 2, \dots, J. \quad (1)$$

Similarly, the pmf of the discrete approximation of the posterior distribution at θ_j given the scores on the items in u , denoted $F_u(\theta_j)$, is computed by replacing c by u in the above equation. Then, one finds the integer k such that $F_c(\theta_j) \geq F_u(\theta_j)$, $j = 1, 2, \dots, k$. Thus, k corresponds to the value on the grid of θ_j s so that F_c is larger than or equal to F_u for that value and any larger value. If $F_c(\theta_1) < F_u(\theta_1)$ or $F_c(\theta_j) = F_u(\theta_j)$ for $j = 1, 2, \dots, k$, then k is set equal to 0. Finally, the PSS for the examinee is computed as

$$\text{PSS} = h \sum_{j=1}^k [F_c(\theta_j) - F_u(\theta_j)]. \quad (2)$$

The PSS is set to 0 if $k = 0$.² If an examinee is aberrant, then the PSS for that examinee is likely to be large; otherwise, the PSS is likely to be small.

Some Limitations of the PSS

One limitation of the PSS is that it does not have a known distribution under the null hypothesis, making it difficult to determine appropriate critical values for the statistic. In his simulation study, Belov (2016) used simulations using the true item parameters to determine critical values of the PSS. Belov (2016) did not discuss how critical values of the PSS can be computed for real data for which true item parameters are unknown. It is possible to use simulations using the estimated item parameters to compute critical values of the PSS for real data; this strategy is used in the real data examples later in this article. However, a discussion in Belov (2015) implies that simulated data may incorrectly represent real data, which may adversely influence the performance for real data of test statistics whose null distribution is determined from simulations; the same discussion implies that the performance of the PSS for real data may be unsatisfactory. Belov (2015) also recommended the use, for real data, of three statistics (D_1 , D_2 , and D_3) that have a known null distribution over his D statistic whose null distribution is unknown even though D_1 , D_2 , and D_3 are much less powerful than D .

A second limitation of the PSS is revealed by taking a deeper look at Figure 1 of Belov (2016), which is supposed to correspond to the case when the examinee has some item pre-knowledge. The figure shows F_c and F_u for abilities between -3 and 3 . However, it is clear that near $\theta = 3$, F_c is decreasing sharply whereas F_u is not; so, if the posterior distributions are computed for abilities between 3 and 5 (abilities between -5 and 5 were considered above Equation 3 of Belov, 2016), F_c will almost surely be smaller than F_u for ability $= 5$; in that case, the value of k would be equal to 0, and hence, PSS would actually be 0, contrary to the positive value mentioned by Belov (2016). This problem is further demonstrated in Figure 1 that shows the values of F_c and F_u for $J = 101$ in the left panel and the logarithms of the F_c and

Figure 1. Two examples where the PSS may be misleading.

Note. PSS = posterior shift statistic.

F_u in the right panel for two aberrant examinees for simulated data. The items in c are more discriminating than those in u for both of these examinees.³ For the first examinee (represented in the top two panels), c and u include 10 items and 14 items, respectively. Thus, 42% items are compromised for the examinee; this percentage is close to the value of 37% for one of the real data sets discussed later. For the examinee, PSS is 0 because F_c is smaller than F_u for ability = 5 (this is more clear from a look at the right panel that shows that $\log(F_c) < \log(F_u)$ for ability close to 5). Also note that if one does not inspect F_c and F_u for abilities larger than 3 for the first examinee, then, because $F_c > F_u$ at ability = 3, it would appear, incorrectly, that the PSS is a large positive number, just like what would appear from Figure 1 of Belov (2016). The bottom panels show F_c , F_u , and their logarithms for another aberrant examinee for whom c and u include 10 and 30 items, respectively (thus, 25% items are compromised). The distribution F_u for this examinee is bimodal with two modes around 0 and 3, respectively—this causes F_u to be larger than F_c around 2.9, which in turn causes the PSS to be 0 up to four decimal places. Although these examples may be somewhat extreme, they lead to graphical plots that are very similar to the one that Belov (2016) used to illustrate the PSS. Therefore, a safe practice to avoid incorrect flagging for those interested in using the PSS is to examine whether such a problem occurs for their own data.

The third limitation of the PSS originates from the fact that its maximum value for a data set is h . The PSS will achieve its maximum value when F_c and F_u will be almost nonoverlapping, with F_c being to the right of F_u ; in that case, F_u will be very close to 0 whenever F_c is positive, which will cause $\sum_{j=1}^k [F_c(\theta_j) - F_u(\theta_j)]$ to be approximately equal to $\sum_{j=1}^k F_c(\theta_j) \approx 1.0$, and then from Equation 2, the PSS will be approximately equal to h .

Table 1 shows some common choices of the number of ability levels and ability ranges used in the context of IRT models in the first two columns and the resulting maximum values of PSS

Table 1. How the Maximum Value of PSS Depends on the Choice of Ability Levels.

Number of ability levels	Ability range	Maximum value of PSS
4I	$\{-3, 3\}$	0.15
4I	$\{-5, 5\}$	0.25
10I	$\{-3, 3\}$	0.06
10I	$\{-5, 5\}$	0.10

Note. PSS = posterior shift statistic.

in the third column. The table shows that for the same data set and the same aberrant examinee, two investigators could come up with a PSS between 0.06 or 0.25 depending on the choices of the number of ability levels and the range of ability. This can be a potential source of confusion for practitioners while deciding what can be treated as a large value of the PSS.

Review of a Statistic Based on the LRT

The rationale behind the PSS is that the posterior distributions based on c and u are close to each other resulting in a small value of the PSS for an examinee who did not benefit from item preknowledge, but the posterior distribution based on c is located far to the right of that based on u resulting in a large value of the PSS for an examinee who benefited from item preknowledge.

Chang and Stout (1993) proved that the posterior distribution of the ability on a large set of items is close to a normal distribution whose mean is equal to the maximum likelihood estimate (MLE) of the examinee ability based on these items. Therefore, if c and u are large sets and the posterior distributions based on c and u are close to each other for an examinee, then the above-mentioned result of Chang and Stout (1993) implies that the examinee's ability estimates based on c and u would be close to each other as well; the result also implies that if the posterior distribution based on c is located far to the right of that based on u , then the examinee's ability estimate based on c is expected to be substantially larger than that based on u .

Thus, instead of testing for the closeness of the posterior distributions (i.e., the basis of the PSS) based on c and u , it is possible to detect item preknowledge by testing whether the true abilities based on c and u are significantly different. A widely used statistical test that can be used to test for the equality of the model parameters underlying two subsets of data (or two samples) is the LRT (e.g., Cox & Hinkley, 1974; Lehmann & Romano, 2005; Rao, 1973).

For an examinee, let us denote the MLE or the weighted maximum likelihood estimate (WLE; Warm, 1989) of the examinee ability based on the scores on c , u , and all items as $\hat{\theta}_c$, $\hat{\theta}_u$, and $\hat{\theta}_0$, respectively. Let the true values corresponding to $\hat{\theta}_c$ and $\hat{\theta}_u$ be denoted by θ_c and θ_u , respectively. Given the earlier discussion, the null hypothesis of no item preknowledge can be denoted as $H_0 : \theta_c = \theta_u$, and the alternative hypothesis of some item preknowledge can be denoted as $H_A : \theta_c > \theta_u$.

The LRT statistic (e.g., Cox & Hinkley, 1974; Lehmann & Romano, 2005; Rao, 1973) for testing H_0 is given by

$$L = 2[L(\hat{\theta}_c; x_i, i \in c) + L(\hat{\theta}_u; x_i, i \in u) - L(\hat{\theta}_0; x_i, i = 1, 2, \dots, I)], \quad (3)$$

where, for example, $L(\hat{\theta}_c; x_i, i \in c) = \log$ -likelihood of the scores on c at $\hat{\theta}_c$.

Because, for example, $L(\hat{\theta}_c; x_i, i \in c) = \sum_{i \in c} \log P_i(x_i | \hat{\theta}_c)$, the LRT statistic given in Equation 3 can be expressed as

$$L = 2 \left[\sum_{i \in c} \log P_i(x_i | \hat{\theta}_c) + \sum_{i \in u} \log P_i(x_i | \hat{\theta}_u) - \sum_{i=1}^I \log P_i(x_i | \hat{\theta}_0) \right]. \quad (4)$$

The statistic provided in Equation 4 is appropriate for testing against a two-sided alternative hypothesis; however, in the present study, the alternative hypothesis is one-sided—the interest is in testing against the alternative $H_A : \theta_c > \theta_u$ (and not $\theta_c < \theta_u$).

For testing the above-mentioned one-sided alternative hypotheses, Sinharay (2017a) suggested the use of the signed LRT statistic that is given by

$$L_s = \begin{cases} \sqrt{L} & \text{if } \hat{\theta}_c \geq \hat{\theta}_u, \\ -\sqrt{L} & \text{if } \hat{\theta}_c < \hat{\theta}_u. \end{cases} \quad (5)$$

Thus, the absolute value of L_s is equal to the square root of L , and L_s is positive if the examinee's estimated ability based on c is larger than or equal to that based on u and negative otherwise. A large positive value of L_s (that means a large value of L accompanied with $\hat{\theta}_c > \hat{\theta}_u$) leads to the rejection of the null hypothesis of no item preknowledge and the flagging of the corresponding examinee for possibly benefiting from item preknowledge.

Sinharay (2017a) showed, using both theoretical results and simulations, that L_s has an asymptotic $\mathcal{N}(0, 1)$ distribution under the null hypothesis of no item preknowledge for both nonadaptive tests and CATs. Sinharay, Duong, and Wood (in press) suggested a statistic very similar to the L_s statistic to detect fraudulent answer changes and found the null distribution of the statistic to be very close to the standard normal distribution.

In theory, L_s has a standard normal null distribution only for a large number of compromised and noncompromised items. In practice, the number of noncompromised items is expected to be large, but the number of compromised items can often be small. However, Sinharay (2017a) showed that in a simulation for CATs, the null distribution of L_s was close to the standard normal distribution for as few as four compromised items. The null distribution of L_s for a few compromised items would be examined in simulations later in this article.

For the examinees considered in Figure 1, L_s is 2.30 and 1.92, respectively, and hence is statistically significant at the .05 level, which is justified because of the examinees' better performance on the compromised items.

A Graphical Demonstration

Let us assume that two examinees took a 20-item test calibrated using the Rasch model and the difficulty parameters of all items are equal to 0. Let us further assume that c includes 10 items for both examinees. Suppose that the two examinees answered 11 and 10 items correctly on the test, that the first examinee is nonaberrant and answered six out of the 10 compromised items correctly, and that the second examinee is aberrant and answered eight out of the 10 compromised items correctly.

The top panel of Figure 2 shows for the nonaberrant examinee the posterior distribution on the noncompromised items or $F_u(\theta_j)$ s and that on the compromised items or $F_c(\theta_j)$ s. A grid of $J = 101$ equispaced values of θ_j between -5 and 5 were used in the computations. The bottom panel shows a similar plot for the aberrant examinee. The two posterior distributions are much closer for the nonaberrant examinee than for the aberrant examinee. The PSS is 0.026 for the nonaberrant examinee and a much larger 0.096 for the aberrant examinee (note that the maximum value of the PSS in this case is 0.10).

The MLEs of the ability are $\hat{\theta}_c = 0.41$, $\hat{\theta}_u = 0.00$, $\hat{\theta}_0 = 0.20$ for the nonaberrant examinee and $\hat{\theta}_c = 1.39$, $\hat{\theta}_u = -1.39$, and $\hat{\theta}_0 = 0.0$ for the aberrant examinee. Vertical lines are drawn in

Figure 2. Posterior distributions of two examinees.

Figure 2 at the estimates $\hat{\theta}_c$ (dashed line) and $\hat{\theta}_u$ (solid line). Thus, $\hat{\theta}_c$ is close to $\hat{\theta}_u$ for the non-aberrant examinee, but much larger than $\hat{\theta}_u$ for the aberrant examinee.

The value of L_s for the nonaberrant examinee is 0.45. The value of L_s for the aberrant examinee is 2.78, which is larger than the 99th percentile of the $\mathcal{N}(0, 1)$ distribution—so L_s is statistically significant for the examinee at the .01 level.

Thus, the PSS and L_s are conceptually similar and would be able to detect item preknowledge, but they work slightly differently. For an aberrant examinee, the posterior distribution $F_c(\theta_j)$ will be to the far right of $F_u(\theta_j)$ that will result in a large value of the PSS, and $\hat{\theta}_c$ will be larger than $\hat{\theta}_u$ that will result in a large positive value of L_s .

Annotated R (R Core Team, 2015) code for computing L_s and PSS for the two examinees is provided in the appendix.

An Effect Size Associated With L_s

The statistic L_s is based on the concept of significance testing and does not incorporate an effect size—so practitioners may find the statistic even more helpful if an effect size can be attached to L_s . One way to compute a rough effect size, assuming that a table is available to convert an ability estimate to a score scale for the test concerned, would involve (a) the computation of two scaled scores: one corresponding to $\hat{\theta}_0$ and another corresponding to $\hat{\theta}_u$, and (b) computing the difference of these two scaled scores. The resulting difference (or a standardized version obtained by dividing the difference by the standard deviation), which is an estimate of what the examinee actually scored minus what he or she would have scored if he or she had not benefited from item preknowledge, would provide some idea on how much the examinee benefited from item preknowledge. For example, if a testing program uses the transformation $500 + 100\theta$ to convert an estimated ability to the reported score scale (of about 200 to 800), then the rough effect sizes (in the form of a difference) for the examinees corresponding to the nonaberrant and aberrant examinees considered in Figure 2 would be 20 and 278, respectively.⁴

A Brief Review of Other Methods of Detecting Item Compromise

Belov (2016) considered seven statistics in addition to the PSS. Drasgow et al. (1996) suggested an optimal person-fit index that is essentially a likelihood ratio statistic to detect cheating (where item preknowledge could be one form of cheating) on a known subset of items; the index is a special case of the likelihood ratio statistics for optimal appropriateness measurement (Levine & Drasgow, 1988). The statistic based on Neyman–Pearson lemma of Belov (2016) is equivalent to the index of Drasgow et al. (1996), as demonstrated by Sinharay (2017b). Belov (2016) found the PSS to have larger power than all the other seven statistics including the optimal index (Drasgow et al., 1996) in his simulations; hence, these seven other statistics are not considered further in this article. McLeod, Lewis, and Thissen (2003) suggested a Bayesian method to detect item preknowledge in CATs; they did not assume that the set of compromised items is known. Researchers such as Segall (2002) and Eckerly, Babcock, and Wollack (2015) suggested IRT models for detection of item preknowledge for nonadaptive tests. Van der Linden and van Krimpen-Stoop (2003) suggested a Bayesian method that uses item scores and response times to detect item preknowledge for nonadaptive tests. Qian, Staniewska, Reckase, and Woo (2016) used a similar approach in which the residuals from the joint application of the Rasch or the two-parameter logistic model to item scores and the response time model of van der Linden (2006) to item-response times were used to flag examinees for possible item preknowledge in both adaptive and nonadaptive tests. These methods are not considered henceforth in this article.

A Simulation Study

A simulation study was performed to compare the performance of the PSS with that of L_s for both nonadaptive tests and CATs.

Design of the Simulation

The design of the simulation was exactly the same as that described in the “Computer Simulations” section of Belov (2016) with only two differences:

- The true item parameters for the nonadaptive tests were simulated from univariate normal distributions with means and variances identical to those of the item pool of Belov (2016).⁵

- The true item parameters of the item pool for the CATs were simulated from univariate normal distributions with means and variances identical to those of the item pool of Belov (2016).

Belov (2016) used the estimated item parameters of several disclosed items of the Law School Admission Test (LSAT) as true item parameters in his simulations.

As in Belov (2016), the test length was 100 items for the nonadaptive tests and 50 items for the CATs, the number of nonaberrant examinees in each replication was 1,000, the number of replications was 10, and the following three factors were manipulated in a $3 \times 2 \times 3$ fully crossed design resulting in 18 simulation cases:

- The uncertainty in the set of compromised items was varied between 0%, 25%, and 50%.
- The distribution of the true ability of the aberrant examinees was a $\mathcal{N}(0, 1)$ distribution or a $U(-3, 0)$, that is, uniform distribution between -3 and 0 .
- The number of aberrant examinees as a percentage of the number of nonaberrant examinees was varied between 5, 10, and 20.

The first factor (uncertainty) reflects the phenomenon that in practice, the set of items that are actually compromised may differ from the set of items that the investigator believes is compromised, and different aberrant examinees may have preknowledge of different sets of items. To simulate uncertainty, the aberrant examinees were divided into 10 aberrant groups of equal size, where group m , $m = 1, 2, \dots, 10$, had preknowledge of a unique random subset of items, c_m . Each c_m had the same size as c , but included $(100 - P)\%$ of the compromised items (randomly drawn)⁶ when the uncertainty in the compromised items of the simulation case is $P\%$. Thus, for the case with 0% uncertainty, c_m is equal to c .

The $\mathcal{N}(0, 1)$ ability distribution represents a sample of regular examinees, and the $U(-3, 0)$ ability distribution represents low-ability examinees who are usually more likely to cheat on examinations. The number of compromised items was 26 for nonadaptive tests and 13 for CATs. The WLEs of the ability were used in the computation of L_s . The ability levels 5.0, 4.9, $\dots$, -4.9 , -5.0 were used to compute the PSS. The item scores of the aberrant examinees on the compromised items were simulated as random draws from a Bernoulli distribution with .9 as the success probability (that implies that aberrant examinees are 90% likely to produce a correct answer on compromised items); all the other item scores were drawn from the 3PLM. One hundred replications of some simulation cases were performed, but the areas under the receiver operating characteristic (ROC) curves (that are described shortly and are the main quantity of interest in the simulation study) up to two decimal places were identical to those with 10 replications; therefore, the number of replications was set equal to 10, the same number as in Belov (2016).

Results

Figure 3 shows a scatterplot of the values of the PSS (x axis) versus those of L_s (y axis) for a sample of 1,000 examinees drawn from all the examinees (both aberrant and nonaberrant) simulated for nonadaptive tests. The figure shows that L_s is large when the PSS is large. The PSS is bounded above by $h = 0.10$ as discussed earlier.

Areas under the ROC curves (e.g., Karabatsos, 2003) truncated between 0 and 0.10 and divided by 0.10, denoted as *truncated ROC areas*, were computed as in Belov (2016) for the two statistics for all the simulation conditions.

The truncated ROC areas hardly varied as the percentage of aberrant examinees varied over 5, 10, and 20; therefore, the areas were averaged over the three values of this factor. Figure 4

Figure 3. A scatterplot of the PSS versus L_s for nonadaptive tests.

Note. PSS = posterior shift statistic.

shows these (averaged) truncated ROC areas for different values of the other two manipulated factors (uncertainty in the compromised items and ability distribution) for the two statistics for nonadaptive tests in the left panel and CATs in the right panel. The x axis denotes the percentage of uncertainty (0%, 25%, or 50%) in the set of compromised items, and the y axis denotes the (averaged) truncated ROC areas. In any panel, the two dotted lines denote the areas for the $U(-3, 0)$ distribution, and the two dashed lines denote the areas for the $\mathcal{N}(0, 1)$ distribution of the ability of aberrant examinees. Among a set of two adjacent dotted or dashed lines, the line joining the three hollow circles represents the areas for the three levels of uncertainty for the PSS and that joining the hollow triangles represents the areas for L_s . Thus, for example, the average truncated ROC area for the PSS for CAT for the standard normal distribution and 0% uncertainty is, from the right panel, about 0.75 (slightly larger than the value of 0.66 in Belov, 2016).⁷

The figure shows that there is hardly a winner with respect to the average truncated ROC area among the two statistics. The areas are almost the same for the two statistics for all simulation cases except for 25% and 50% uncertainty under the $U(-3, 0)$ distribution for CATs in which the areas for L_s are slightly larger than those for the PSS.⁸ The average truncated ROC areas, averaged over all conditions, are 0.647 for the PSS and 0.652 for L_s .

The effects of the manipulated factors on the power (or truncated ROC area) of the statistics are very similar to those in Belov (2016)—the power of each statistic is larger as the extent of uncertainty in the compromised items becomes smaller, when the abilities of the aberrant examinees follow the $U(-3, 0)$ rather than the $\mathcal{N}(0, 1)$ distribution, and for nonadaptive tests rather than CATs.

The Case of a Few Compromised Items

To further examine the null distribution, Type I error rate, and power of L_s for small sizes of c (i.e., for the case when the number of compromised items is small), the above simulations were repeated with only the following changes:

- c was assumed to include only five items (instead of 25 in the simulations above) for nonadaptive tests.

Figure 4. The average truncated ROC areas of the two statistics.
Note. ROC = receiver operating characteristic; PSS = posterior shift statistic; CAT = computerized adaptive test.

- The total number of compromised items was assumed to be four (instead of 13 in the simulations above) for CATs, which resulted in sizes 1, 2, 3, and 4 of c .

The solid line in Figure 5 shows the kernel density estimate of the distribution⁹ of the observed values of L_s from all nonaberrant examinees for CATs with one to four compromised items. The probability density function of the standard normal distribution is also shown using a dashed line for comparison. Although the distribution of L_s seems too peaked around 0, the distribution is not too far from the standard normal distribution, and the right tails of the two distributions are almost on top of each other. The 90th, 95th, and 99th percentiles of the distribution of L_s are 1.28, 1.64, and 2.35, respectively, which are all very close to the corresponding quantiles of a standard normal distribution. The distribution of L_s for nonaberrant examinees appears very similar to that in Figure 5 for the nonadaptive tests with five compromised items. Thus, the simulations show that the null distribution of L_s is not too far from the standard normal distribution for a few compromised items.

Table 2 displays the Type I error rates and power at three levels of significance (.001, .01, and .05) when the size of c is 1 to 4 in the case of the CATs with four compromised items. The Type I error rates were computed as the percentage of nonaberrant examinees (over all levels of uncertainty, aberrant distribution, and percentage of aberrant examinees) whose values of L_s were larger than the appropriate quantile of the standard normal distribution. The power was computed as the percentage of aberrant examinees (over all levels of uncertainty, aberrant distribution, and percentage of aberrant examinees) whose values of L_s were larger than the appropriate quantile of the standard normal distribution. The numbers of examinees for each size of c are provided in the second and sixth columns (under the column heading “ n ”) of the table; the numbers in the second column add up to 180,000 because there are 10 replications, each

Figure 5. Distributions of L_s for nonaberrant examinees for CATs for one to four compromised items. Note. CAT = computerized adaptive test.

Table 2. Type I Error Rates and Power of L_s for CATs With Four Compromised Items.

Size of c	n	Type I error rate			n	Power		
		$\alpha = .001$	$\alpha = .01$	$\alpha = .05$		$\alpha = .001$	$\alpha = .01$	$\alpha = .05$
1	10,381	.007	.012	.015	901	.00	.07	.07
2	4,283	.001	.001	.036	1,203	.09	.18	.33
3	13,412	.010	.015	.028	1,922	.17	.25	.47
4	151,924	.003	.011	.055	16,956	.18	.41	.53

Note. CAT = computerized adaptive test.

involving 1,000 nonaberrant examinees, for each of 18 simulation cases. The table shows that L_s becomes more conservative as the size of c decreases. For example, at level .05, the Type I error rate and power of L_s are 0.055 and 0.53 when the size of c is 4 but only 0.015 and 0.07 when the size of c is 1; one source of this conservativeness is that the average standard error (SE) of the WLE is 1.35, 0.97, 0.95, and 0.92 when the size of c is 1, 2, 3, and 4, respectively, which means that the ability estimates become more imprecise as the number of compromised items becomes smaller. Although Table 2 indicates that L_s suffers from small power for very small sizes of c , it also indicates that L_s does not incorrectly flag too many nonaberrant examinees in these cases. The Type I error rates of L_s were close to the nominal level and equal to 0.001, 0.012, and 0.055, respectively, at levels .001, .01, and .05 for nonadaptive tests with five compromised items.

Figure 6, in a manner similar to Figure 4, shows the average truncated ROC areas of the PSS and L_s for these simulations. Although the areas in Figure 6 are mostly smaller than those in Figure 4, the average truncated ROC areas of PSS and L_s are very close in Figure 6, just as in

Figure 6. The average truncated ROC areas of the two statistics for a few compromised items.
Note. ROC = receiver operating characteristic; PSS = posterior shift statistic; CAT = computerized adaptive test.

Figure 4. The average truncated ROC areas for a few compromised items, averaged over all conditions, are 0.529 for the PSS and 0.540 for L_s .

The above results demonstrate that (a) the asymptotic standard normal null distribution of L_s seems to hold for examinees who receive four or more compromised items, (b) the Type I error rate of L_s is smaller than the nominal level and the power is small for examinees who receive fewer than four compromised items, and (c) L_s is as powerful as the PSS for a few compromised items.

A Real Data Application

Let us consider item-response data from two forms of a nonadaptive licensure test. The data sets were analyzed in several chapters of Cizek and Wollack (2017) and also in Sinharay (2016). Both forms include 170 operational items that are dichotomously scored; 87 items are common between the two forms. Item scores were available for 1,636 examinees for Form 1 and 1,644 examinees for Form 2. The licensure organization who provided the data identified as compromised 63 and 61 items on Forms 1 and 2, respectively (though several of these items were common between the forms so that the combined number of compromised items was 64). The organization also flagged 46 and 48 individuals on Forms 1 and 2, respectively, as possible cheaters from a variety of statistical analysis and a rigorous investigative process that brought in other information; given the rigor of the investigative process, these examinees will be treated as truly aberrant.

The values of PSS and L_s were computed from the two data sets. The Rasch model is operationally used in the test—the difficulty parameters were estimated using the marginal maximum likelihood estimation procedure from the data sets and used in the computation of the statistics.

Table 3. The Proportion of Examinees for Which PSS and L_s Were Significant for the Two Data Sets.

Group of examinees	Statistic	Form 1			Form 2		
		.001	.01	.05	.001	.01	.05
Not flagged	PSS	0.00	0.02	0.07	0.00	0.05	0.11
Not flagged	L_s	0.00	0.02	0.06	0.01	0.06	0.11
Flagged	PSS	0.15	0.24	0.33	0.17	0.27	0.33
Flagged	L_s	0.16	0.25	0.35	0.17	0.27	0.33

Note. PSS = posterior shift statistic.

The WLEs of abilities, computed using the Newton–Raphson algorithm, were used to compute L_s . The ability levels 5.0, 4.9, . . . , −4.9, −5.0 were used to compute the PSS.

The proportions of examinees for which the two statistics were significant at significance levels of .001, .01, and .05 are provided in Table 3. The first two lines of Table 3 include the proportions significant among the examinees who were not flagged by the licensure organization. The last two lines include the proportions significant only among the (46 or 48) examinees who were flagged by the licensure organization; thus, for example, the proportion 0.15 for the PSS at the .001 level for Form 1 in the third row of numbers denotes that among the 46 examinees flagged by the licensure organization, the PSS was significant at the .001 level for seven examinees (note that $7 / 46 \approx 0.15$). The value of L_s of an examinee was considered significant when it was larger than the corresponding quantile of the $\mathcal{N}(0, 1)$ distribution. The value of PSS of an examinee was considered significant when it was larger than 0.077, 0.091, and 0.097, respectively, for Form 1, and 0.078, 0.092, and 0.098, respectively, for Form 2 at levels of .05, .01, and .001, respectively. To compute the critical values of the PSS for a data set, (a) scores on each item were simulated under the assumption of no item preknowledge for each examinee in the data set using the estimated item and ability parameters, (b) the PSS was computed from these simulated scores, (c) the appropriate quantiles of these PSSs were used as critical values. Table 3 shows that the proportions of significant values are close for the two statistics in all cases. However, although the proportions were obtained without any simulations (and using percentiles of the standard normal distribution) for L_s , a simulation had to be performed to obtain the proportions for the PSS. Table 3 also shows that the proportions significant for each statistic are much larger among the examinees flagged by the licensure organization (bottom two rows of the table) than among those not flagged (top two rows 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 truly aberrant. Note that item compromise was not the only reason of flagging by the licensure organization—so a proportion smaller than 1.0 in the bottom two rows is not a severe limitation of the statistics.

Conclusion

Standard 6.6 of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014) includes the recommendation that when the test results have important consequences, score integrity should be supported through active efforts to prevent, detect, and correct scores obtained by fraudulent or deceptive means. As a consequence, there is a growing interest in detection of test fraud.

This article compares the performances of two recently suggested statistics for detecting item preknowledge, which is a major type of test fraud, using simulated and real data involving both nonadaptive and adaptive tests. The performances of the two statistics were found very similar and, as the R codes in the appendix show, the computation of both the statistics is simple.

Among the two statistics, L_s has a known asymptotic distribution under the null hypothesis, the result holding even for a small number of compromised items (see, for example, Table 2); in contrast, the PSS does not have a known distribution under the null hypothesis. Therefore, the following two types of arguments, the first from a psychometric viewpoint and the second from a statistical viewpoint (also, the first from a practical viewpoint and the second from a theoretical viewpoint), imply that practitioners will find L_s to be more attractive than the PSS:

- The same reasoning (discussed under the first limitation of the PSS earlier in this article) that Belov (2015) used to recommend D_1 to D_3 (whose null distribution is known) over D (whose null distribution is unknown) implies that for real data, simulation-based critical values may adversely affect the performance of the PSS and L_s should be used instead of PSS.
- Box (1979) recommended that when model criticism “is done formally, using tests of goodness of fit, it must, I think, employ sampling theory for its justification.” (p. 2)¹⁰. The statistic L_s , which uses sampling theory for its justification, follows the recommendation of Box (1979) more closely than does the PSS.

Statistical indices for detecting item preknowledge are useful for providing confirming evidence of inappropriate behavior when evidence from other sources also exists, but the evidence provided by one statistical index is insufficient by itself. For example, Hanson, Harris, and Brennan (1994) commented that no statistical index on its own can provide conclusive proof that test fraud occurred; however, several different types of statistical indices may provide strong evidence in combination. For example, as stated on the web page of *FindLaw* (<http://caselaw.findlaw.com/us-5th-circuit/1374633.html>), a testing company withheld the SAT I score of an examinee because he made too many erasures, his score gain was unusually large, and the answer options chosen by him had an unusual correspondence with another examinee who was seated diagonally in front of him; the court agreed with the testing company. Researchers such as Tendeiro and Meijer (2014) recommended complementing statistical indices of detecting irregularities with other sources of information such as seating charts, video surveillance, or follow-up interviews.

There are several limitations of this article and, consequently, several related topics can be further investigated. First, although the simulation study was detailed, it is possible to perform more simulations, possibly with different true item parameters, different test lengths, and different IRT models. Similarly, it is possible to consider applications of the statistics to more real data examples. Second, the item parameters were treated as known in the simulations in this article. This assumption is reasonable especially in the context of CATs where item parameters are assumed known and the ability parameters are estimated. However, it is possible to estimate item parameters and consider the effect of the estimation on the properties of the indices considered here, especially for nonadaptive tests. Third, this article only involves tests with dichotomous items, and it is possible in future research to consider tests with polytomous items or tests with a mix of dichotomous and polytomous items. Finally, it would be interesting to combine the statistics here with the methods of van der Linden and van Krimpen-Stoop (2003) and Qian et al. (2016) that use response times to detect item preknowledge.

Appendix

R Code to Compute the L_s and Posterior Shift Statistic (PSS) for Two Examinees

```
# Log-likelihood on one item under the Rasch model with difficulty=0
prob = function(t) {return(1 - 1/(1+exp(t)))}
logpr=function(y,t) {p=prob(t)
                    return(ifelse(y==1,log(p),log(1-p)))}

nitem=20
indc=c(rep(0,10),rep(1,10)) # Indicators of compromised items
nc=sum(indc)
nu=nitem-nc
L=rep(0,2)
Lc=rep(0,2)
Lu=rep(0,2)
y=matrix(0,2,nitem)
y[1,]=rep(c(1,0,0,1,1,1,0,0,1,0),2) #Scores for non-aberrant examinee
y[1,20]=1
y[2,]=c(rep(0,10),rep(1,10)) #Scores for aberrant examinee
y[2,1:2]=1
y[2,19:20]=0
yc=y[,indc==1]
yu=y[,indc==0]
# Compute the MLEs on all items, compromised items, and non-compromised items
library(irtos) #The R package 'irtos' will be used to estimate ability
t=mlebm(y,cbind(rep(1,nitem),rep(0,nitem),rep(0,nitem)))
tc=mlebm(yc,cbind(rep(1,nc),rep(0,nc),rep(0,nc)))
tu=mlebm(yu,cbind(rep(1,nu),rep(0,nu),rep(0,nu)))
# Compute log-likelihoods on all items, compromised items, and non-compromised
items
for (m in 1:2)
  {L[m]=sum(logpr(y[m,],t[m]))
  Lu[m]=sum(logpr(yu[m,],tu[m]))
  Lc[m]=sum(logpr(yc[m,],tc[m]))}
cat('\Theta estimates (compromised, non-compromised, overall) for non-
aberrant examinee:
',round(c(tc[1],tu[1],t[1]),2),'\\n')
# Ls for non-aberrant examinee
cat('\Ls for non-aberrant examinee=',sign(tc[1]-tu[1])*sqrt(2*(Lc[1]+
Lu[1]-L[1])),'\n')
cat('\Theta estimates (compromised, non-compromised, overall) for aberrant
examinee:
',round(c(tc[2],tu[2],t[2]),2),'\\n')
# Ls for aberrant examinee
cat('\Ls for aberrant examinee=',sign(tc[2]-tu[2])*sqrt(2*(Lc[2]+Lu[2]-
L[2])),'\n')
# Computation of PSS
ng=101 #Number of ability levels
theta=seq(-5,5,,ng) #The ability levels
Fu=matrix(1,2,ng)
Fc=Fu
for (m in 1:2) {for (j in 1:ng) #Computation of Fc and Fu
{for (i in 1:nu) {Fu[m,j]=Fu[m,j]*exp(logpr(yu[m,i],theta[j]))}
for (i in 1:nc) {Fc[m,j]=Fc[m,j]*exp(logpr(yc[m,i],theta[j]))}
```



```

Fu[m,]=Fu[m,]/sum(Fu[m,])
Fc[m,]=Fc[m,]/sum(Fc[m,])
width=theta[2]-theta[1]
pss=rep(0,2)
for (m in 1:2) { k=0
  for (j in ng:1) {if (Fc[m,j]>= Fu[m,j])
    {k=j}
  else
    {break}}
pss[m]=0
if (k>0) {for (j in (k:ng))
  {pss[m] = pss[m] + width*(Fc[m,j]-Fu[m,j])}}
cat('PSS for non-aberrant and aberrant examinees: ',pss,'\n')}

```

Author's Note

The author is currently employed at Educational Testing Service, Princeton, New Jersey.

Acknowledgment

The author thanks the editor Hua Hua Chang and the anonymous reviewers for several helpful comments that led to a significant improvement of the article. The author also thanks James Wollack for generously sharing two data sets that were used in the research that led to this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Two other statistics based on the Wald test and the score test, which were also suggested by Sinharay (2017a), performed as well as (in general) or slightly worse than (for a few compromised items) that based on likelihood ratio test (LRT) in general—so results on those would not be reported here.
2. Note that Equation 4 of Belov (2016), which provides the equation for the posterior shift statistic (PSS), includes a typographical error and hence appears different from Equation 2 here.
3. That is not unrealistic; for example, Belov (2016) chose the compromised items in his simulations on CAT as those with the largest exposures where the items were selected using the maximum information criterion, and typically, the items with the largest discrimination parameters suffer from the largest exposure under the maximum information criterion. The average discrimination was 1.18 for c but 0.73 for u in the simulation study (whose design is very similar to the simulation study in Belov, 2016) later in this article.
4. The suggestion of a generous reviewer led to this suggestion of an effect size.
5. Thus, for example, the true b_i s were simulated from a normal distribution with mean = 0.49 and variance = 1.27.
6. The rest of the c_m is randomly drawn from the noncompromised items.
7. This difference is most likely an outcome of the use of true item parameters that are different from those of Belov (2016). It could be an outcome of different ability ranges or yet another factor.

8. Also, by using the variability between the 10 replications, the differences in the areas between the two statistics were found statistically significant for this case.
9. Created using the function “density” in the R software (R Core Team, 2015).
10. Note that the PSS, L_s , and other statistics for detecting item preknowledge all essentially involve a test of goodness of fit between the item scores of an examinee and the item response theory (IRT) model under no item preknowledge.

References

- American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Belov, D. I. (2015). Robust detection of examinees with aberrant answer changes. *Journal of Educational Measurement*, 52, 437-456.
- Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, 40, 83-97.
- Box, G. E. P. (1979). Some problems of statistics and everyday life. *Journal of the American Statistical Association*, 74, 1-4.
- Chang, H. H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, 58, 37-52.
- Cizek, G. J., & Wollack, J. A. (2017). *Handbook of detecting cheating on tests*. Washington, DC: Routledge.
- Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London, England: Chapman & Hall.
- Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education*, 9, 47-64.
- Eckerly, C. A., Babcock, B. G., & Wollack, J. A. (2015, April). *Preknowledge detection using a scale-purified deterministic gated IRT model*. Paper presented at the Annual meeting of the National Council of Measurement in Education, Chicago, IL.
- Hanson, B. A., Harris, D. J., & Brennan, R. L. (1994). *A comparison of several statistical methods for examining allegations of copying* (ACT Research Report Series No. 87-15). Iowa City, IA: American College Testing.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16, 277-298.
- Kyle, T. (2002). *Cheating scandal rocks GRE, ETS*. Retrieved from <http://thedartmouth.com/2002/08/09/cheating-scandal-rocks-gre-ets/>
- Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). New York, NY: Springer-Verlag.
- Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness index. *Psychometrika*, 53, 161-176.
- McLeod, L. D., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, 27, 121-137.
- Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice*, 35, 38-47.
- R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: John Wiley.
- Segall, D. O. (2002). An item response model for characterizing test compromise. *Journal of Educational and Behavioral Statistics*, 27, 163-179.
- Sinharay, S. (2016). Asymptotic corrections of standardized extended caution indices. *Applied Psychological Measurement*, 40, 418-433.
- Sinharay, S. (2017a). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, 42, 46-68. doi:10.3102/1076998616673872

- Sinharay, S. (2017b). On the equivalence of a likelihood ratio of Drasgow, Levine, and Zickar (1996) and the statistic based on the Neyman-Pearson lemma of Belov (2016). *Applied Psychological Measurement, 41*, 145-149. doi:10.1177/0146621616673597
- Sinharay, S., Duong, M. Q., & Wood, S. W. (in press). A new statistic for detection of aberrant answer changes. *Journal of Educational Measurement*.
- 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. (2006). A lognormal model for response times on test forms. *Journal of Educational and Behavioral Statistics, 31*, 181-204.
- van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika, 68*, 251-265.
- Veerkamp, W. J. J., & Glas, C. A. W. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics, 25*, 373-389.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.