The Lack of Robustness of a Statistic Based on the Neyman-Pearson Lemma to Violation of its Underlying Assumptions

Sandip Sinharay, Educational Testing Service

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

The citation for the article is: Sinharay, S. (in press). The lack of robustness of a statistic based on the Neyman-Pearson lemma to violations of its underlying assumptions. Applied Psychological Measurement.

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

The Lack of Robustness of a Statistic Based on the Neyman-Pearson Lemma to Violations of its Underlying Assumptions

Sandip Sinharay, Educational Testing Service

February 16, 2021

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

The Lack of Robustness of a Statistic Based on the Neyman-Pearson Lemma to Violations of its Underlying Assumptions

Abstract

Drasgow, Levine, and Zickar (1996) suggested a statistic based on the Neyman-Pearson lemma (e.g., Lehmann & Romano, 2005, p. 60) for detecting preknowledge on a known set of items. The statistic is a special case of the optimal appropriateness indices of Levine and Drasgow (1988) and is the most powerful statistic for detecting item preknowledge when the assumptions underlying the statistic hold for the data (e.g., Belov, 2016; Drasgow et al., 1996). This paper demonstrated using real data analysis that one assumption underlying the statistic of Drasgow et al. (1996) is often likely to be violated in practice. This paper also demonstrated, using simulated data, that the statistic is not robust to realistic violations of its underlying assumptions. Together, the results from the real data and the simulations demonstrate that the statistic of Drasgow et al. (1996) may not always be the optimum statistic in practice and occasionally has smaller power than another statistic for detecting preknowledge on a known set of items, especially when the assumptions underlying the former statistic do not hold. The findings of this paper demonstrate the importance of keeping in mind the assumptions underlying and the limitations of any statistic or method.

Key words: Item preknowledge; optimal appropriateness index; signed likelihood ratio test.

Acknowledgements

The author would like to thank John Donoghue, the editor, Brian Habing, the Associate Editor, the two anonymous reviewers, Hongwen Guo, Bingchen Liu, Matthew Johnson, and Rebecca Zwick for several helpful comments that led to a significant improvement of the manuscript. The author would also like to express his gratitude to Carol Eckerly, Sarah Toton, and James Wollack for sharing with him data or data summaries that were used in the research that led to this paper. The research was supported by the Institute of Education Sciences, US Department of Education, through Grant R305D170026.

1. Introduction

Item preknowledge refers to some examinees having prior access to test items and/or answers before taking the test. The items that the examinees have prior access to are referred to as *compromised*. Levine and Drasgow (1988) suggested a set of test statistics based on the *Neyman-Pearson lemma* (NPL; e.g., Lehmann & Romano, 2005, p. 60) to detect examinees whose response patterns are *aberrant* due to cheating, language issues etc.—these statistics are referred to as the *optimal appropriateness indices* (OAIs). A statistic that is a special case of the OAIs and can be applied to detect preknowledge on a known set of compromised items was suggested by Drasgow et al. (1996)—the statistic will be referred to as the *Drasgow-Levine-Zickar statistic* (DLZS). Belov (2016) used a statistic very similar to the DLZS to detect preknowledge on a known set of items (e.g., Sinharay, 2017a). Because of the similarity of the statistic of Belov (2016) to the DLZS, these two statistics will be treated to be the same in this paper.

The DLZS is based on two assumptions, one regarding the true ability distributions of the examinees and the other regarding the success probability of the examinees with preknowledge on the compromised items. Researchers such as Belov (2016) and Drasgow et al. (1996) stated that the DLZS is the most powerful statistic for detecting preknowledge on a known set of items under the two aforementioned assumptions. The DLZS has been employed to detect preknowledge by Drasgow et al. (1996), Belov (2016), Sinharay (2017a), and Sinharay (2017b) among others. However, there is a lack of knowledge on the extent to which the two assumptions underlying the DLZS hold for real data and on the robustness of the DLZS to violations of its underlying assumptions. This paper aims to fill those gaps. The goals of this paper are to (a) demonstrate using real data that one of the two assumptions underlying the DLZS is often likely to be violated in practice, (b) demonstrate using simulated data that the DLZS is not robust to realistic violations of its underlying assumptions. The lack of the robustness is demonstrated by showing that the DLZS is often less powerful than the signed likelihood ratio (SLR) statistic (Sinharay, 2017c), which is another statistic for detecting preknowledge on a known set of compromised items, under realistic violations of the underlying assumptions. The overarching goal of this paper is

to demonstrate that the DLZS may not always be the optimum statistic for detecting preknowledge on a known set of compromised items.

The next section includes descriptions of the DLZS and the SLR statistic. In the following section, evidence from real data analyses and literature review are brought to bear on the two assumptions under which the DLZS is the most powerful statistic. The Simulations section includes a simulation study that examines the extent of robustness of the DLZS to realistic violations of its underlying assumptions. The Real Data section includes an example where the DLZS is found to flag fewer cheaters than the SLR statistic for two real data sets that included some known compromised items and known cheaters. Conclusions and recommendations are provided in the last section.

2. Background

2.1 Notation

Let x_i denote the score of an examinee whose true ability is represented as θ on item $i, i = 1, 2, \cdots, I$, of a test. This paper only considers dichotomous items that are scored as correct (1) or incorrect (0). Let $P_i(\theta)$ denote the probability of a correct answer on item i by the examinee under an item response theory (IRT) model. For example, if the 2-parameter logistic model (2PLM) is used for modeling the data from the test, then

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

where a_i and b_i respectively are the slope and difficulty parameters of item i.

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_I)$ denote the scores of the examinee on all the items. This paper focuses on the case when a set $\mathcal{C}$ of items, which is known to the investigator, has been compromised and the investigator intends to detect the examinees who may have benefitted from preknowledge of the items in $\mathcal{C}$. Examples of preknowledge on a known set of compromised items for high-stakes tests can be found in Cizek and Wollack (2017, p. 14) and Eckerly, Smith, and Lee (2018). Therefore, the problem of detecting preknowledge on a known set of items is important in practice, notwithstanding the fact that the set of compromised items is often unknown.

Let the vector $\boldsymbol{x}_{\mathcal{C}}$ denote the scores of the examinee on the compromised items.[1] Let $\mathcal{U}$ denote the set of uncompromised items, where an *uncompromised* item is one that is not known to have been compromised, and let the vector $\boldsymbol{x}_{\mathcal{U}}$ denote the scores of the examinee on the uncompromised items. Thus, for example, $\boldsymbol{x}_{\mathcal{C}} = (x_1, x_2, \ldots, x_{10})'$ and $\boldsymbol{x}_{\mathcal{U}} = (x_{11}, x_{12}, \ldots, x_{30})'$ for a 30-item test whose first 10 items were compromised.

Suppose that for the examinee, n_{c1} denotes the number of correct answers to the compromised items and n_{c0} denotes the number of incorrect answers to the compromised items. That is, $\sum_{i \in \mathcal{C}} x_i = n_{c1}$ and $n_{c0} = I - n_{c1}$. It is assumed in this paper that there are no missing item scores.

2.2 The OAI and the DLZS

In the approach of Levine and Drasgow (1988) for detecting examinees whose responses/scores are aberrant due to factors such as cheating and carelessness, one assumes a probability model $P_{\text{Aberrant}}(\boldsymbol{X} = \boldsymbol{x})$ to describe the item scores under *aberrant* responding. One also assumes another model $P_{\text{Normal}}(\boldsymbol{X} = \boldsymbol{x})$ to describe the item scores under *normal* or non-aberrant responding. Then, one computes the *likelihood ratio* λ for each examinee as

$$\lambda = \frac{P_{\text{Aberrant}}(\boldsymbol{X} = \boldsymbol{x})}{P_{\text{Normal}}(\boldsymbol{X} = \boldsymbol{x})}, \tag{2}$$

and flags an examinee as having a significant extent of aberrant responses if $\lambda > k$ for the examinee. The critical value k can be computed by simulating data of several normal examinees. Levine and Drasgow (1988) referred to the λ of Equation 2 as the OAI. Typically, $P_{\text{Normal}}(\boldsymbol{X} = \boldsymbol{x})$ is computed as the marginal probability of $\boldsymbol{x}$ under an IRT

[1] This paper considers the simple case when all examinees benefitting from preknowledge were administered all items in $\mathcal{C}$. The methods and findings of this paper apply in a straightforward manner to the case when those benefitting from preknowledge were administered some (but not all) items in $\mathcal{C}$, which could happen on an adaptive test.

model. For example, Levine and Drasgow (1988, p. 170) stated that

$$P_{\text{Normal}}(\boldsymbol{X}=\boldsymbol{x})=\int_{\theta} f(\theta)\prod_{i=1}^{I} P_i(\theta)^{x_i}[1-P_i(\theta)]^{(1-x_i)}d\theta, \tag{3}$$

where $f(\theta)$, the examinee ability distribution, is assumed to be the standard normal distribution, that is,

$$f(\theta)=\frac{1}{\sqrt{2\pi}}e^{-\frac{\theta^2}{2}}.$$

The choice of the model for $P_{\text{Aberrant}}(\boldsymbol{X}=\boldsymbol{x})$ is more complicated and depends on the type of aberrance that is of interest.

Drasgow et al. (1996) suggested a special case of the OAI to detect cheating, like preknowledge, on a known set of items, which is the problem of interest in this paper. To compute $P_{\text{Aberrant}}(\boldsymbol{X}=\boldsymbol{x})$ for this special case, it is assumed that the cheaters correctly answer the compromised items with a large probability (p) and answer the uncompromised items in the same manner as the non-cheaters. These assumptions allow one to compute $P_{\text{Aberrant}}(\boldsymbol{X}=\boldsymbol{x})$ for this case as

$$P_{\text{Aberrant}}(\boldsymbol{X}=\boldsymbol{x})=p^{n_{c1}}(1-p)^{n_{c0}}\int_{\theta} f(\theta)\prod_{i\in\mathcal{U}} P_i(\theta)^{x_i}[1-P_i(\theta)]^{(1-x_i)}d\theta. \tag{4}$$

The probability of normal responding is computed as in Equation 3. Then, using Equations 2-4, the likelihood ratio for detecting cheating on a known set of compromised items for an examinee, referred to here as the DLZS, is obtained as

$$\text{DLZS}=\frac{P_{\text{Aberrant}}(\boldsymbol{X}=\boldsymbol{x})}{P_{\text{Normal}}(\boldsymbol{X}=\boldsymbol{x})}=\frac{p^{n_{c1}}(1-p)^{n_{c0}}\int_{\theta} f(\theta)\prod_{i\in\mathcal{U}} P_i(\theta)^{x_i}[1-P_i(\theta)]^{(1-x_i)}d\theta}{\int_{\theta} f(\theta)\prod_{i=1}^{I} P_i(\theta)^{x_i}[1-P_i(\theta)]^{(1-x_i)}d\theta}. \tag{5}$$

Thus, the DLZS is a special case of the OAI that is provided in Equation 2. Drasgow et al. (1996, p. 50) suggested using p close to 1 depending on the investigator's opinion about the application and recommended using numerical integration to approximate the integrals in the numerator and denominator of Equation 5.

Belov (2016) suggested a statistic to detect preknowledge on a known set of compromised items. Sinharay (2017a) proved that the statistic of Belov (2016) is essentially the same as the DLZS except that Belov (2016) suggested using $p=0.95$ and replacing the

integrals in the numerator and denominator of Equation 5 by a summation over a grid of 101 equispaced values $\theta_1 = -5, \theta_2 = -4.9, \cdots, \theta_{100} = 4.9, \theta_{101} = 5$. Specifically, the version of DLZS used by Belov (2016) is given by

$$\text{DLZS} = \frac{0.95^{n_{c1}}(1-0.95)^{n_{c0}} \sum_{j=1}^{101} f(\theta_j) \prod_{i \in \mathcal{U}} P_i(\theta_j)^{x_i}[1-P_i(\theta_j)]^{(1-x_i)}}{\sum_{j=1}^{101} f(\theta_j) \prod_{i=1}^{I} P_i(\theta_j)^{x_i}[1-P_i(\theta_j)]^{(1-x_i)}}. \tag{6}$$

In this paper, the DLZS was approximated using Equation 6.[2] The distribution of DLZS under the null hypothesis of no item preknowledge is not known yet, even asymptotically. Drasgow et al. (1996) recommended computing the critical/cutoff values for the DLZS using simulations. Note that Drasgow et al. (1996) considered other types of aberrance such as cheating on an unknown subset of items and faking on personality tests that involve the computation of the λ in manners different from that in Equations 5 and 6.

2.3 The DLZS, the NPL, and the Underlying Assumptions

The OAI and, consequently, the DLZS, is based on the NPL that is briefly described in Appendix A of this paper. For example, Levine and Drasgow (1988, p. 165) stated that

> *The Neyman-Pearson Lemma asserts that a likelihood ratio test is optimal in the sense that if a statistical test for aberrance has the same probability of incorrectly classifying a normal examinee as a likelihood ratio test, then the likelihood ratio test has equal or greater probability of correctly classifying aberrant examinees. There may be other tests that classify as well as a likelihood ratio test, but none can be better.*

As discussed in Appendix A, the NPL can be used to obtain most powerful tests only for the case when a simple null hypothesis is tested against a simple alternative hypothesis, where a *simple hypothesis* is a hypothesis under which the distribution of the random variable of interest is completely known, that is, does not depend on any unknown parameters.

[2]Other numerical integration approaches and a grid of more than 101 values lead to values very close to those obtained from Equation 6.

However, the problem of detecting item preknowledge on a known set of items, which is the problem of interest in this paper, is not inherently equivalent to the test of a simple null hypothesis versus a simple alternative hypothesis. For example, if an examinee has preknowledge of an item, the probability of a correct answer is unknown and could be equal to 1 or 0.95 or a value even smaller—so the probability distribution of the item scores under preknowledge is not completely known. Drasgow et al. (1996) re-framed/converted this problem of detecting preknowledge on a known set of items to a problem involving the testing of a simple null hypothesis versus a simple alternative hypothesis. Their re-framed (simple) null hypothesis is that the distribution of the examinee's item scores is provided by Equation 3 and the re-framed (simple) alternative hypothesis is that the distribution of the examinee's item scores is provided by Equation 4. This re-framing ensured the applicability of the NPL to the problem and ensured that the DLZS is the most powerful test of the simple null versus the simple alternative hypotheses assumed by Drasgow et al. (1996). However, the re-framing of Drasgow et al. (1996) involved the following two crucial assumptions: (a) the ability distribution of the examinees (both under the null and alternative hypotheses) is the standard normal distribution—this assumption allowed the integrations in Equations 3 and 4, and (b) the probability of a correct answer by an examinee with preknowledge on a compromised item is a known number (like 0.95) that is very close to 1.0.

As stated by, for example, Drasgow et al. (1996) and Belov (2016), the DLZS given by Equation 6 is the most powerful statistic for detecting preknowledge on a known set of compromised items only when the two aforementioned assumptions are satisfied for the data at hand. However, there is a lack of any examination or evidence on how often these two assumptions, which helped them to convert the complex null and alternative hypotheses to simple ones, hold for real data. There is a similar lack of research on the robustness of the DLZS to these assumptions, that is, on whether the DLZS is the most powerful statistic under violations of one or more of these two assumptions.

The DLZS is also based on the assumptions that the IRT model holds and the item parameters are precisely known, but these assumptions are not examined here and are

assumed to hold. That is primarily because IRT model misfit and/or imprecise parameter estimates would lead to all reported scores being unfair—so it is assumed that the test administrators ensured adequate IRT model fit and precise parameters estimates. It may be worthwhile to examine the relationship between IRT model misfit and the DLZS in future research.

2.4 The Signed Likelihood Ratio Statistic

Sinharay (2017c) suggested the signed likelihood ratio (SLR) statistic for the detection of preknowledge on a known set of compromised items. For an examinee, let us define the maximum likelihood estimate (MLE) of the examinee ability from the scores on the compromised items as $\hat{\theta}_{\mathcal{C}}$, that from the scores on the uncompromised items as $\hat{\theta}_{\mathcal{U}}$, and that from the scores on all the items as $\hat{\theta}$. Sinharay (2017c) argued that the problem of detection of preknowledge on a known set of items ($\mathcal{C}$) is essentially the same as that of testing the null hypothesis $H_0 : \theta_{\mathcal{C}} = \theta_{\mathcal{U}}$ versus the alternative hypothesis $H_0 : \theta_{\mathcal{C}} > \theta_{\mathcal{U}}$, where $\theta_{\mathcal{C}}$ and $\theta_{\mathcal{U}}$ respectively are the true values corresponding to $\hat{\theta}_{\mathcal{C}}$ and $\hat{\theta}_{\mathcal{U}}$.

The likelihood ratio test (LRT) statistic (e.g., Guo & Drasgow, 2010) for testing $H_0 : \theta_{\mathcal{C}} = \theta_{\mathcal{U}}$ versus the alternative hypothesis $H_0 : \theta_{\mathcal{C}} \neq \theta_{\mathcal{U}}$ is given by

$$\Lambda = 2[\ell(\hat{\theta}_{\mathcal{C}}; \boldsymbol{x}_{\mathcal{C}}) + \ell(\hat{\theta}_{\mathcal{U}}; \boldsymbol{x}_{\mathcal{U}}) - \ell(\hat{\theta}; \boldsymbol{x})], \tag{7}$$

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

$\ell(\hat{\theta}_{\mathcal{U}}; \boldsymbol{x}_{\mathcal{U}})$ = log-likelihood of the scores on the uncompromised items at $\hat{\theta}_{\mathcal{U}}$,

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

To test the null hypothesis $H_0 : \theta_{\mathcal{C}} = \theta_{\mathcal{U}}$ versus the alternative hypothesis $H_0 : \theta_{\mathcal{C}} > \theta_{\mathcal{U}}$, Sinharay (2017c) suggested using the signed likelihood ratio (SLR) statistic, which is a function of the statistic Λ, and is given by

$$L_s = \begin{cases} \sqrt{\Lambda} \text{ if } \hat{\theta}_{\mathcal{C}} \geq \hat{\theta}_{\mathcal{U}}, \\ -\sqrt{\Lambda} \text{ if } \hat{\theta}_{\mathcal{C}} < \hat{\theta}_{\mathcal{U}}. \end{cases}$$

A large value of L_s leads to the rejection of the null hypothesis of no item preknowledge. The statistic L_s was proved to have an asymptotic standard normal distribution under the null hypothesis by Sinharay (2017c).

Because the null and alternative hypotheses underlying the SLR statistic are not simple hypotheses, the SLR statistic cannot be expected to be the most powerful statistic for detecting preknowledge on a known set of compromised items, even though the statistic is based on a likelihood ratio. However, Sinharay (2017c), Sinharay (2017d), Sinharay and Jensen (2019), and Wang, Liu, Robin, and Guo (2019) found the SLR statistic to be as powerful as or more powerful than the existing statistics in detecting item preknowledge and found the Type I error rate of the statistic to be very close to the nominal level. Specifically, Sinharay (2017d) proved the performance of the SLR statistic to be very similar to that of the posterior shift statistic that was found to be the most powerful among eight preknowledge-detection statistics by Belov (2016). However, there is a lack of comparison of the performance of the SLR statistic to that of the DLZS. In this paper, the performance of the SLR statistic is compared to that of the DLZS in the Simulation and Real Data sections.

3. Taking a Deeper Look at the Assumptions Behind the DLZS

This section focuses on two of the assumptions underlying the DLZS. The assumption of a standard normal ability distribution is discussed first. Next, an examination is performed of the assumption on the probability of a correct answer of an examinee who benefited from item preknowledge.

3.1 The Assumption of Standard Normal Ability Distributions

In the computation of the DLZS, the true population/ability distributions for those with item preknowledge and not with item preknowledge are assumed to be the standard normal distribution. However, researchers such as Li and Cai (2017) and Woods and Thissen (2006) asserted that given a specific IRT model, the examinee ability distribution

may deviate from the normal distribution. Li and Cai (2017) further stated that if multiple sub-populations are grouped together, the combined ability distribution may be multimodal or another type of non-normal distribution. In the context of item preknowledge, those with preknowledge and without preknowledge could constitute the multiple subpopulations mentioned by Li and Cai (2017) and could lead to the violation of the assumption that the population distribution is standard normal. Unfortunately, this assumption is not easy to check in practice due to the lack of a decent-sized sample of examinees who are known to have preknowledge on a known set of items. Note that the normality or otherwise of the estimated examinee-ability distribution from the whole sample does not provide any useful information about the violation of the assumption. Appendix B shows examples of cases when (a) the normality assumption holds and the estimated ability distribution is non-normal and (b) the normality assumption does not hold and the estimated ability distribution is close to normal.

3.2 The Assumption Regarding the Probability of a Correct Answer

The expression of the DLZS given by Equation 6 was obtained by setting the probability of a correct answer on a compromised item by an examinee with preknowledge (p of Equation 5) equal to 0.95. However, there is a lack of evidence in favor of the assumption that p is equal to 0.95 (or any other value like 0.90 or 0.99). In practice, compromised items could be accompanied with correct answer keys, incorrect answer keys, or, no answer keys. For example, Eckerly et al. (2018) reported that all the items, along with answer keys, on a test form were found on a website, but the answer keys provided on the website were correct for 24 items and incorrect for 36 items. One may expect p to depend on whether the answer key was correct in such a case. In general, the value of p will depend on several factors such as (a) additional information (like answers) that is available with the item and the accuracy of that information, (b) the examinee's ability level (stronger examinees are more likely to find the correct answers if the answer keys are unavailable), (c) the length of the time the examinees had between their availability of the items and the test administration, and (d) the resources that are available to the examinee (for example,

does the examinee know someone who can help him/her find the correct answers to the compromised items if the answer keys are not available?). Therefore, a blanket assumption of $p = 0.95$ is probably not justified. In addition, the probability of a correct answer on an item for which, for example, $a_i = 1$ and $b_i = -2$ in Equation 1, is larger than 0.95 without preknowledge for all examinees with θ larger than about 1.0—the assumption of $p = 0.95$ under preknowledge for such an examinee is equivalent to the assumption that preknowledge leads to worse-than-expected performance for the examinee and contradicts the common knowledge that preknowledge leads to better performance (e.g., Smith & Davis-Becker, 2011). Three data sets are analyzed below in an attempt to obtain evidence regarding the success probability of those with item preknowledge on compromised items.

3.2.1 Two Licensure Test Data Sets

Let us consider item-response data from two forms of a non-adaptive licensure test. The data sets were analyzed in several chapters of Cizek and Wollack (2017) and also in Sinharay (2017c). Both forms include 170 operational items that are dichotomously scored. The sample sizes were 1,636 for Form 1 and 1,644 for Form 2. The licensure organization who provided the data identified as compromised 63 and 61 items on Forms 1 and 2, respectively and 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 may be treated as true cheaters for all practical purposes. The exact type of cheating that the 94 flagged examinees may have been involved with is unknown. Consequently, the number of examinees flagged specifically for item preknowledge is unknown. While researchers such as Sinharay (2017c) and Boughton, Smith, and Ren (2017) found evidence of several examinees benefiting from item preknowledge for the data set, other researchers such as Zopluoglu (2017) found evidence of several examinees benefiting from answer-copying for the data set. The proportion-correct scores of all examinees were computed on the items that were known to have been compromised. Thus, for example, the proportion correct score on the compromised items for an examinee on

the first form represents the proportion of the 63 compromised items that the examinee answered correctly. Histograms of the proportion-correct scores on the compromised items of only the flagged examinees for Forms 1 (left panel) and 2 (right panel) are provided in Figure 1. A vertical dashed line is shown at the value of 0.95 in each panel. The figure

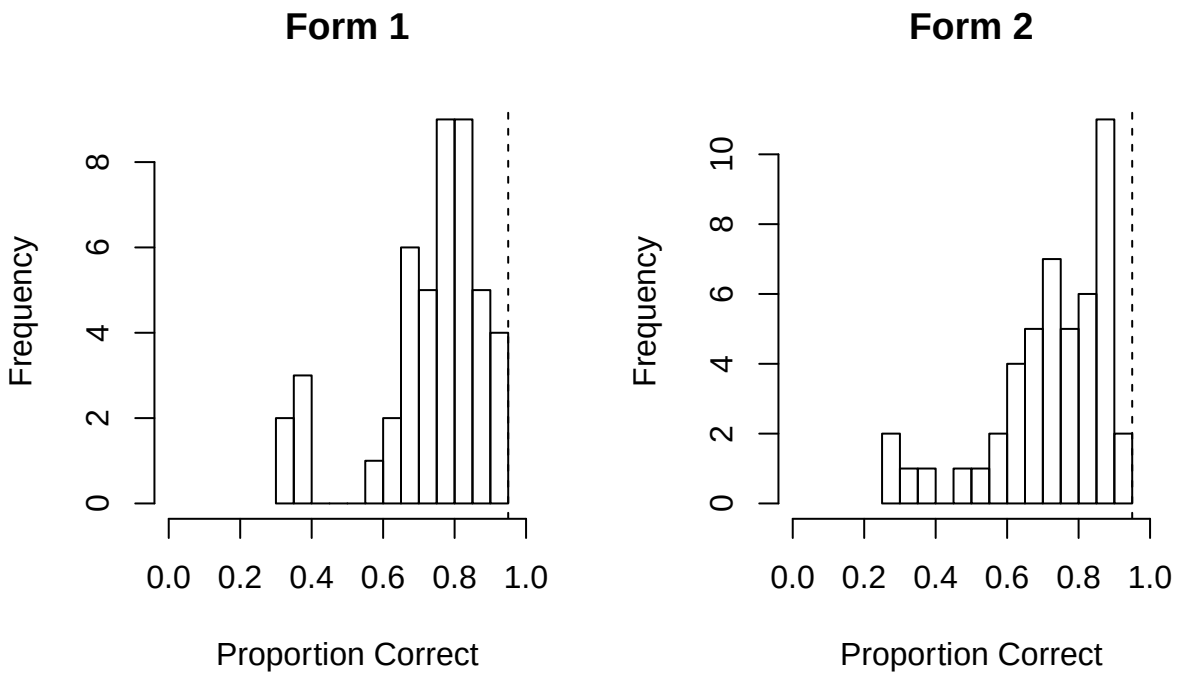

Figure 1. Histograms of the proportion-correct scores on the compromised items of the flagged examinees on the two forms of the licensure test.

shows that the proportion-correct is 0.95 or smaller for all flagged examinees—the mean proportion correct on the compromised items for the flagged examinees for the two forms are 0.74 and 0.73, respectively. Combining over the two test forms, the null hypothesis of $p = 0.95$ is rejected at 5% level for 85 examinees[3] among the 94 flagged examinees. For the 23 examinees for whom the DLZS was statistically significant at 5% level (see the real data section later for details on the computation of DLZS for these data), the proportion-correct

[3]The proportion-correct scores of all of these 85 examinees were smaller than 0.89.

scores range between 0.85 and 0.95, their mean is 0.90, and the null hypothesis of $p = 0.95$ is rejected at 5% level for more than half of these examinees.[4] Thus, the results for the licensure data sets do not provide much evidence in favor of the assumption that the p of Equation 5 is equal to 0.95.

3.2.2 A Data Set Involving Artificially Created Item Preknowledge

Toton and Maynes (2019) and Belov and Toton (2020) analyzed a data set from a study in which item preknowledge was artificially created for a sample of 93 undergraduate students who were administered a 25-item test that resembles the GRE® Quantitative Reasoning test. Preknowledge on a known set of 12 items was artificially created by sharing/disclosing 12 out of the 25 items with some of the 93 examinees 20 minutes before the test; 30 examinees received only the 12 compromised/disclosed items while 30 others received the 12 items and their answers. The average proportion correct scores of these two groups of examinees on the 12 disclosed items are provided in Table 1. For comparison purposes, the table also includes the average proportion correct scores on the disclosed items of the 33 examinees without preknowledge and the average proportion correct scores on the 13 undisclosed items for all the three groups of examinees. The table shows that the average proportion correct score for the "Both Items & Answers" group (that is, the 30 examinees who received both the items and their answers) on the disclosed items was 0.94, which is quite close to the value of 0.95. However, the average for the "Only Items, No Answers" group on the disclosed items was 0.72, which, while being somewhat larger than the corresponding value for those without preknowledge, is well below the value of 0.95 assumed in Belov (2016). The null hypothesis p=0.95 on the disclosed items is rejected at 5% significance level by an exact test of binomial proportions (e.g., Lehmann & Romano,

[4]It is expected that these 23 examinees include some false positives and hence cannot be considered as the only cheaters. Also, it is expected that some actual cheaters do not feature in these 23 examinees due to Type II error that is common in hypothesis testing. So these numbers cannot be used as evidence for setting p=0.90 either.

2005, pp. 68-69) for most examinees under the "Only Items, No Answers" condition (this is clear from the mean of 0.72 for the condition in Table 1) and for three examinees (that is, for 10% of the 30) under the "Both Items & Answers" condition.

Table 1. Average Proportion Correct Scores for Three Groups of Undergraduate Students on the Disclosed and Undisclosed Items.

Examinee Group	Sample Size	Average Proportion Correct	
		Undisclosed items	Disclosed items
No Preknowledge	33	0.58	0.64
Only Items, No Answers	30	0.59	0.72
Both Items & Answers	30	0.62	0.94

3.2.3 Discussion on Proportion Correct Scores on Compromised Items

The above real data examples demonstrate that the assumption that the probability of a correct answer by a cheater on a compromised item is equal to 0.95 or any other fixed number, which is a crucial assumption underlying the DLZS, is often likely to be violated in practice. However, such a violation does not necessarily mean that the DLZS would not be the most powerful statistic. That is because some statistical approaches are known to be robust to certain assumptions and not robust to some other assumptions. For example, analysis of variance and related procedures such as multiple comparison are robust to the underlying normality assumption, but not robust to the presence of severe outliers (e.g.; Montgomery, 2013, pp. 81-82). Therefore, a simulation study was performed to examine the robustness of the DLZS to violations of the assumptions underlying the statistic. This examination of robustness is performed by comparing the performances of the DLZS and the SLR statistic (whose Type I error rate has been found to be very close to the nominal level) in the context of detection of preknowledge on a known set of compromised items. It would be examined whether the DLZS is more powerful among the two statistics when the assumptions underlying the former are not satisfied.

4. Simulations: A Comparison of the DLZS and the SLR Statistic

To examine the robustness of the DLZS to realistic violations of its underlying assumptions, its performance was compared using simulated data sets to that of the SLR statistic (Sinharay, 2017c) for the case of known compromised items. The simulated data include data that satisfy the above-mentioned two assumptions (under which the DLZS should be the most powerful statistic) and data that do not satisfy the assumptions (under which it is unknown whether the DLZS is the most powerful statistic). In doing so, this paper is the first to report a detailed comparison of the performances of the DLZS and the SLR statistic.

4.1 Design of the Simulations

All simulations involved a non-adaptive 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.[5] The true abilities of those who did not benefit from item preknowledge (non-cheaters) were simulated from a standard normal distribution. The true abilities of those who benefited from item preknowledge (cheaters) were simulated from one of the following three distributions: (a) a standard normal or $\mathcal{N}(0,1)$ distribution, (b) a uniform distribution between -3 and 0, or $\mathcal{U}(-3,0)$ distribution, (c) a mixture of the $\mathcal{N}(0,1)$ and $\mathcal{U}(-3,0)$ distributions.[6] The first two of the three ability distributions were used to simulate examinee abilities of cheaters by Belov (2016) and Sinharay (2017d). The third distribution is in between the first two distributions. The first distribution represents the case when the cheaters have the same ability on average as the non-cheaters and the second and third represent the case when the cheaters have smaller ability on average and an ability distribution of a different shape than the non-cheaters.

[5]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 author).

[6]To simulate a random draw from this mixture, one first simulates a random number r from the $\mathcal{U}(0,1)$ distribution and then simulates a draw from the $\mathcal{N}(0,1)$ or the $\mathcal{U}(-3,0)$ distribution depending on whether r is smaller than 0.5 or not.

Note that the DLZS is computed under the assumption of the standard normal ability distribution of cheaters and non-cheaters—so the first of the three ability distributions favors the DLZS and the other two do not. The set of the compromised items was assumed to be a subset of size 10, 20, or 30 of the 100 items on the test; the set of uncompromised items included the remaining 90, 80, or 70 items of the test. The number of cheaters was assumed to be 5%, 10%, or 20% of the number of non-cheaters.

The item scores of the non-cheaters on all items and of the cheaters on the uncompromised items were simulated from the 2PLM. In three sets of simulations, the item scores of the cheaters on the compromised items were simulated in three different ways. It was assumed in the first set of simulations that the probability of a correct answer of a cheater on the compromised items is 0.95 (this assumption is favorable to the DLZS)—the corresponding item scores were simulated from a Bernoulli distribution with success probability of 0.95. This set of simulations will be referred to as those under the "fixed success probability" condition. In the second set of simulations, the item scores of a cheater on a compromised item were simulated from a Bernoulli distribution with success probability that is randomly generated from a $\mathcal{U}(0.8, 1)$ distribution. This condition will be referred to as the "less variable success probability" condition. In the third set of simulations, the item scores of a 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 examinee, or, by shifting the ability to the right by 2.0. Item response data under aberrant responding has been simulated after shifting the examinee ability (or a "θ-shift") by researchers such as Drasgow et al. (1996), Glas and Dagohoy (2007), and Zickar and Drasgow (1996). The simulations with shifted examinee abilities recreate the scenario that item preknowledge leads to a boost in the ability so that the success probability of a cheater on a compromised item is not a fixed value and is larger than what is expected under no preknowledge. This condition will be referred to as the "more variable success probability" condition. When the ability distribution of the cheaters was standard normal, their success probabilities on the compromised items varied between 0.63 and 1.00 (a range that is wider compared to that between 0.8 and 1.0 assumed under the

"less variable success probability" condition) under the "more variable success probability" condition, with the average value being equal to 0.91. Note that because of a θ-shift, item preknowledge always leads to better performance under the "more variable success probability" condition, unlike under the "fixed success probability" and "less variable success probability" conditions.

The four simulation factors were crossed with each other. Thus, 81 simulation conditions (involving all combinations of three ability distributions of the cheaters, three ways to compute the success probability of the cheaters on the compromised items, three sizes of the set of compromised items, and three values of the percent of cheaters) were considered. For each simulation condition, 100 data sets were simulated; the number of non-cheaters 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 set of compromised items was the same for all examinees with preknowledge in each iteration/replication, but varied over the 100 iterations for each simulation condition.

4.2 Computations

For each simulation condition, the following computational steps were performed 100 times to simulate and analyze 100 data sets:

1. Simulate scores of 2,000 non-cheaters to the 100 items from the 2PLM. Simulate scores of 100, 200, or 400 cheaters (depending on the simulation condition) on the non-compromised items from the 2PLM and on the compromised items after computing their success probabilities on the compromised items in one among the three aforementioned manners.

2. Compute the estimated item parameters, using the marginal maximum likelihood estimation procedure, from the data set that included the cheaters and non-cheaters.

3. For each examinee, compute the SLR statistic and the DLZS. The MLE of ability, restricted to the range -4.0 and 4.0, was used to compute the SLR statistic. The item parameter estimates obtained in the previous step were used in these calculations.

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

The Type I error rates and power of the SLR statistic were computed at 1% and 5% significance levels using cutoffs of 2.33 and 1.64 (that are the corresponding normal percentiles), respectively. The computation of the Type I error rates and power of the DLZS required the computation of appropriate cutoffs using simulated data, as recommended by Drasgow et al. (1996), because the null distribution of the statistic is unknown; the cutoff for a data set was computed as the 99th (at 1% level) or 95th (at 5% level) percentile of the distribution of the DLZS among the true non-cheaters in the data set.

The comparison of the power of statistics for detecting aberrant examinees has been performed using receiver operating characteristics (ROC) curves by, for example, Drasgow, Levine, and Williams (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 non-aberrant examinees are known, a 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 non-aberrant examinee is larger than y
- 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 successive 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., Hanley & McNeil, 1982) of a statistic is a measure of how powerful the statistic is. In the context of detecting aberrant examinees, researchers such as Belov (2016) used *truncated ROC areas*, or areas under the ROC curves truncated between 0 and 0.1 and divided by 0.10—that is because false positive rates larger than 0.10 are hardly employed in the context of detecting aberrant examinees (e.g.,

Wollack, Cohen, & Eckerly, 2015). The truncated ROC area of a very powerful statistic is expected to be close to 1 and a larger area corresponds to a more powerful statistic. The truncated ROC areas of the SLR statistic and the DLZS were computed for all the 81 simulation conditions.

4.3 Results

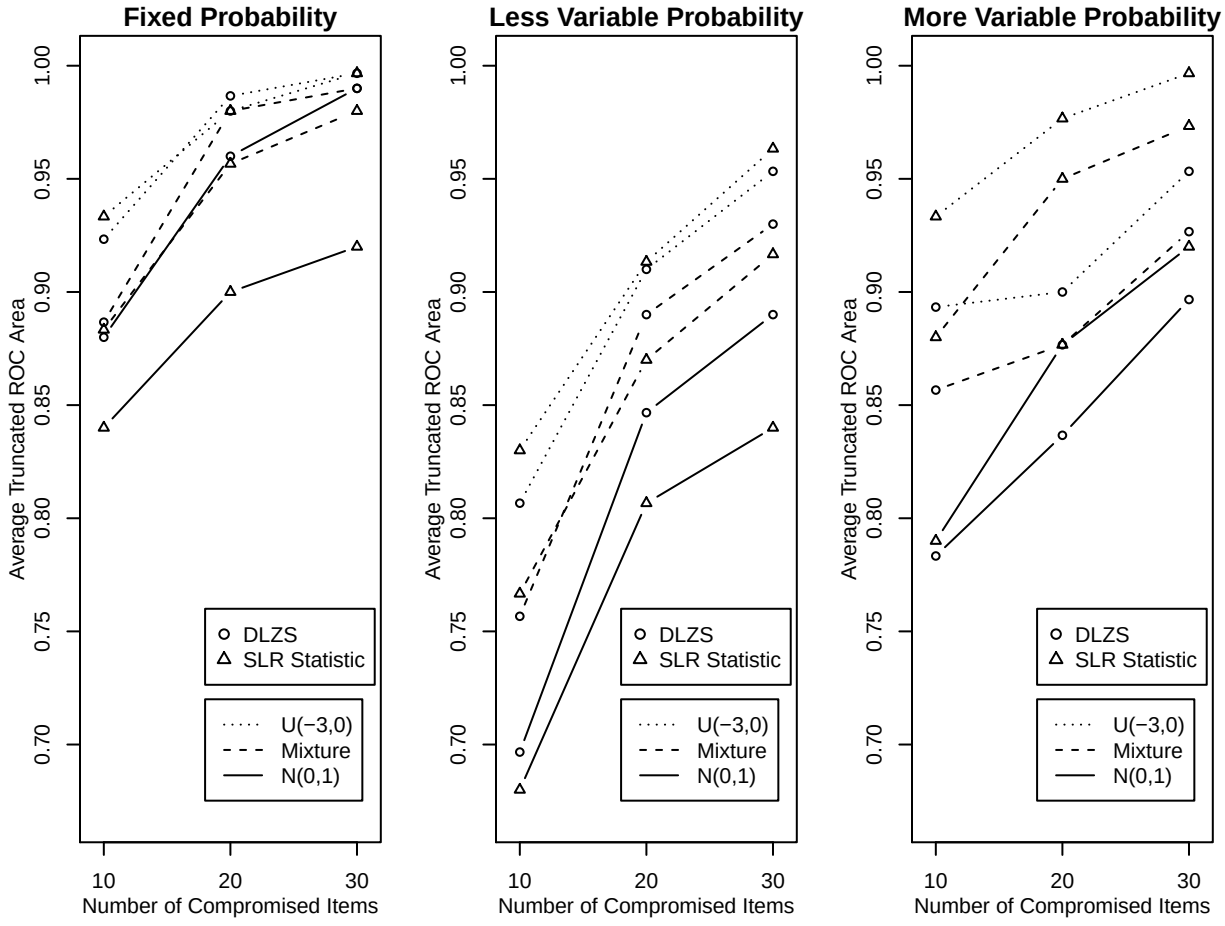

Figure 2. The average truncated ROC areas for the two statistics for the simulated data.

When the other three simulation factors were fixed, the truncated ROC area of neither statistic was affected by the percent of examinees benefiting from preknowledge—so the truncated ROC areas were averaged over the three levels of this percent. The average truncated ROC areas of the SLR statistic and the DLZS for the various levels of the other

three simulation factors are shown in Figure 2. The three panels of the figure respectively show the average ROC areas for the "fixed success probability", "less variable success probability", and "more variable success probability" conditions. In each panel, two dotted lines, two solid lines, and two dashed lines respectively show the average ROC areas for the simulation conditions in which the $\mathcal{U}(-3, 0)$ distribution, the $\mathcal{N}(0, 1)$ distribution, or their mixture was used as the true ability distribution for the cheaters. The hollow circles and hollow triangles respectively denote the average ROC areas of the DLZS and the SLR statistic.

Several findings are common for the two statistics. The average truncated ROC areas for the statistics increase when the number of compromised items increase. For a given way to compute the success probability of cheaters on the compromised items, the areas for the statistics are the largest under the $\mathcal{U}(-3, 0)$ distribution and smallest under the $\mathcal{N}(0, 1)$ distribution of the cheaters, a finding that is similar to one in Belov (2016) and Sinharay (2017d). With respect to the comparative performance of the two statistics, the two solid lines on the leftmost panel of Figure 2 indicate that the DLZS is more powerful than the SLR statistic when the two above-mentioned assumptions underlying the former statistic are satisfied, that is, under "fixed success probability" and standard normal ability distribution. The two dotted lines on the leftmost panel indicate that when the assumption of fixed success probability holds, but the assumption of standard normal ability distribution does not, the DLZS is slightly more powerful than the SLR statistic for 20 compromised items, but not for 10 or 30 compromised items. The rightmost panel of the figure (that corresponding to the "more variable success probability" condition) shows that when the assumption of fixed success probability is severely violated, the SLR statistic has substantially larger power compared to the DLZS irrespective of the ability distribution. The middle panel of the figure indicates that the comparative performance of the two statistics is somewhere in between their comparative performance in the other two panels and there is no clear winner under moderate violations of the assumption of fixed success probability. That it, under the "less variable success probability" condition, the DLZS is more powerful than the SLR statistic for the standard normal ability

distribution (as indicated by the two solid lines), less powerful for the $U(-3, 0)$ ability distribution (two dotted lines), and more powerful for 20 or 30 compromised items for the mixture distribution (two dashed lines).

The Type I error rates of the SLR statistic and the DLZS (not shown here) were very close to the nominal level at both 1% and 5% significance levels. The comparative performance of the SLR statistic and the DLZS with respect to power (not shown here) was very similar to their comparative performance with respect to the average truncated ROC area, that is, the DLZS is more powerful than the SLR statistic under the conditions with fixed success probability and standard normal ability distribution, but not necessarily for the other conditions.

Additional simulations, which were similar to the above simulations but involved no estimation of item parameters, were also performed. In these simulations, the item parameters were assumed known and equal to their true values—both Belov (2016) and Sinharay (2017d) made this assumption in their simulation studies. The average truncated ROC areas of both the statistics were larger than those in Figure 2 in these additional simulations, but the comparative performance of the DLZS and the SLR statistic was the same as above; that is, these simulations also showed that the DLZS statistic is more powerful than the SLR statistic only when the two assumptions underlying the former statistic hold. Results of these additional simulations are not included here and can be obtained upon request from the authors.

Overall, the results from the simulations indicate that while the DLZS is more powerful than the SLR statistic when the two assumptions underlying the former statistic hold, it is often less powerful than the SLR statistic when the assumptions do not hold. In fact, the DLZS can be considerably less powerful than the SLR statistic under violations of its assumptions, as is clear from the rightmost panel of Figure 2. Thus, keeping in mind the fact that the Type I error rate of the SLR statistic has been found to be very close to the nominal level[7] by researchers such as Sinharay (2017c), Sinharay (2017d), and Wang

[7]It should be noted that one can use the NPL to compare a statistic to other statistics whose Type I error rates are not larger than the nominal level.

et al. (2019), the DLZS cannot be considered entirely robust to realistic violations of its underlying assumptions.

5. Comparison of DLZS and SLR statistics for Real Data

The values of the DLZS and the SLR statistic were computed for data from the two above-mentioned licensure test forms after fitting the 2PLM to them. The values of the S-χ^2 item-fit statistic (Orlando & Thissen, 2000) and a statistic for detecting local dependence (Chen & Thissen, 1997) for the two forms indicate that the 2PLM fits the data sets adequately. The sets of compromised items (of sizes 63 and 61, respectively) that were identified by the licensure organization were used as the set of compromised items ($\mathcal{C}$) in the analysis. The MLE of ability, restricted to the range -4.0 and 4.0, was used to compute the SLR statistic. Critical values for the SLR statistic were appropriate percentiles from the standard normal distribution. Critical values for the DLZS were computed using simulations, as recommended by Drasgow et al. (1996), by simulating data from the 2PLM using the item-parameter estimates from the original data sets, computing the values of the DLZS for the simulated examinees, and setting the critical values equal to the appropriate percentiles of the values of DLZS for the simulated examinees.

Table 2. Percent of Significant Values of the Statistics for the Two Licensure Test Data Sets.

Statistic	Percent Significant				Truncated ROC Area	
	Form 1		Form 2		Form	Form
	1%	5%	1%	5%	1	2
DLZS	15.2	30.4	22.9	29.2	0.57	0.59
SLR Statistic	19.6	39.1	25.0	29.2	0.61	0.60

Table 2 shows the percent of statistically significant values of the DLZS (first row of numbers) and the SLR statistic (second row) at 1% and 5% significance levels among the examinees who were flagged as cheaters by the licensure organization (46 and 48 examinees, respectively, for the two forms). Columns 2-3 and 4-5 of Table 2 respectively show the percentages for Forms 1 and 2. The percent-significants do not differ between the two

statistics for the second form, but those for the SLR statistic are considerably larger than those for the DLZS for the first form. It is possible to compute the truncated ROC areas for the two statistics for the data sets by treating the flagged and non-flagged examinees as true cheaters and non-cheaters, respectively. These areas are shown in Columns 6 and 7 of Table 2. The areas for the SLR statistic are slightly larger than those for the DLZS for both the data sets.

Given that (a) the 2PLM appears to adequately fit the data sets, (b) the critical values for the DLZS were found using simulations so that its Type I error rate is close to the nominal level, (c) the Type I error rate of the SLR statistic has been found close to the nominal level by researchers such as Sinharay (2017c), Sinharay (2017d), and Wang et al. (2019), and (d) the investigative procedure used by the licensure organization was quite rigorous so that the flagged examinees can be considered as true cheaters for all practical purposes, Table 2 appears to demonstrate that the SLR may be more useful than the DLZS for some real data sets. The reason for the larger number of significant values of the SLR statistic may be that the assumption of $p = 0.95$ is unlikely to be true for these data sets, as was demonstrated in Figure 1 and the surrounding discussion, causing the DLZS to not be the most powerful statistic (among those whose Type I error rates are not larger than the nominal level) for the data.

6. Conclusions

Researchers such as Belov (2016) and Drasgow et al. (1996) stated that the DLZS suggested by Drasgow et al. (1996) is the most powerful statistic for detecting preknowledge on a known set of (compromised) items under the assumptions that the ability distribution is standard normal and the probability of a correct answer on a compromised item by a cheater is equal to a large value such as 0.95. While one can expect the DLZS to be most powerful only when these assumptions are satisfied for the data at hand, there is a lack of studies on finding how often these assumptions hold for real data and on the robustness of the DLZS to violations of these assumptions. This paper demonstrated using real data that the second assumption may often not hold in practice and demonstrated using simulated

data that another statistic (the SLR statistic suggested by Sinharay, 2017c) may be more powerful than the DLZS, especially when the above-mentioned two assumptions are not appropriate. Thus, this paper shows that the DLZS is not entirely robust to realistic violations of its underlying assumptions. Drasgow et al. (1996) re-framed the problem of detecting preknowledge on a known set of items as the test of a simple null versus simple alternative hypotheses and suggested the DLZS to solve the problem. However, the problem is inherently not a test of a simple null versus a simple alternative hypothesis[8] and the simple alternative hypothesis assumed by Drasgow et al. (1996) may not always reflect the reality in real cases of item preknowledge—the DLZS may not be the most powerful statistic in such cases. Thus, this paper has the important practical implication that one should look beyond the DLZS in investigations of preknowledge on a known set of items, especially when evidence justifying the two assumptions underlying the DLZS is lacking, and should consider other statistics such as the SLR statistic that make milder assumptions. Note that Bayesian approaches for detecting preknowledge (e.g., Wang, Liu, & Hambleton, 2017; Sinharay & Johnson, 2020b) were not considered here and those approaches may also perform as well as or better than the DLZS in practice.

Three more limitations of the DLZS are the following:

- The DLZS is yet to be extended to polytomous items. Such an extension would require the assumption of fixed probabilities of various scores for the cheaters on the compromised polytomous items and it is very difficult to obtain such fixed probabilities. On the other hand, the SLR statistic has been extended to polytomous items by Sinharay (2017c).

- The asymptotic distribution of the DLZS under the null hypothesis is yet to be derived. While Drasgow et al. (1996, p. 63) stated that simulations can be used to compute critical values for the DLZS, they also admitted that critical values computed using

[8]Thus, a most powerful test may not exist for this problem in general. Discussions in, for example, Lehmann and Romano (2005, p. 65) imply that most powerful tests often do not exist for hypotheses that are not simple. The search for a most powerful statistic for this problem is a potential area of future research.

simulations may not be accurate. In addition, researchers such as Box (1979) recommended using test statistics with known null distributions in practice and the DLZS does not satisfy this recommendation.

- The DLZS may have low power when the set of compromised items is not precisely known, that is, when the assumption of known compromised items is violated. Under the violation of this assumption, Belov (2016) found the posterior shift statistic to be more powerful than the DLZS and Sinharay (2017c) found the SLR statistic to be about as powerful as the posterior shift statistic—so the SLR statistic is expected to be more powerful than the DLZS under violations of this assumption. In additional simulations, the SLR statistic was found much more powerful than the DLZS under such a violation even when the two assumptions on the true ability distribution and the success probability of the cheaters on the compromised items are satisfied; appendix D includes some details from one such simulation.

Given the abundance of assumptions in the models and methods in our field, the results of this paper suggest that researchers and practitioners should carefully evaluate whether their data are likely to support the assumptions underlying the model or method that they plan to use, examine how the model or method is likely to perform under realistic violations of the assumptions, and consider other models or methods if necessary. Specifically, before using what they believe is the most powerful test, researchers and practitioners should ask questions such as "Most powerful under what conditions?", "Do those conditions hold for the data at hand?", and "What test should be used if the assumptions do not hold for the data?"

This paper has several limitations, and, consequently, it is possible to perform future research on several related areas. First, while this paper demonstrated that the DLZS is not the most powerful test in general for detecting item preknowledge on a known set of compromised items, it is possible to perform future research on finding a statistic that is in some sense the optimum statistic (and is more powerful than other statistics in most realistic conditions) for the problem. Second, because the DLZS is a special case of the

OAI, it is possible to extend this study to examine the robustness of the OAI to violations of its underlying assumptions. Third, it is possible to compare the DLZS to Bayesian statistics (that are not covered by the NPL) for detecting item preknowledge (e.g., those suggested by Sinharay & Johnson, 2020a, 2020b; Wang et al., 2017) in a future study.

References

Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, *40*, 83–97.

Belov, D. I., & Toton, S. L. (2020, September). *Clique-based detection of examinees with pre-knowledge on real, marked data.* Paper presented at the Virtual Annual meeting of the National Council of Measurement in Education.

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

Box, G. E. P. (1979). Some problems of statistics and everyday life. *Journal of the American Statistical Association*, *74*, 1–4.

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

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

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

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.

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

Eckerly, C., Smith, R., & Lee, Y. (2018, October). *An introduction to item preknowledge*

detection with real data applications. Paper presented at the Conference on Test Security, Park City, UT.

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.

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

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.

Li, Z., & Cai, L. (2017). Summed score likelihood–based indices for testing latent variable distribution fit in item response theory. *Educational and Psychological Measurement*, *78*, 857–886.

Montgomery, D. C. (2013). *Design and analysis of experiments* (8th ed.). New York, NY: Wiley.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50–64.

Sinharay, S. (2017a). 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.

Sinharay, S. (2017b). Are the nonparametric person-fit statistics more powerful than their parametric counterparts? Revisiting the simulations in Karabatsos (2003). *Applied Measurement in Education*, *30*, 314–328.

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

Sinharay, S. (2017d). 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., & 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. (2020a). Detecting test fraud using Bayes factors. *Behaviormetrika*, *47*, 339–354.

Sinharay, S., & Johnson, M. S. (2020b). The use of the posterior probability in score differencing. *Journal of Educational and Behavioral Statistics*. (Advance online publication. doi:10.3102/1076998620957423)

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

Toton, S. L., & Maynes, D. D. (2019). Detecting examinees with pre-knowledge in experimental data using conditional scaling of response times. *Frontiers in Education*, *4*, 1–18.

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.

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

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*, 281–301.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement*, *20*, 71–87.

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

Appendix A: The Neyman-Pearson Lemma

Consider the case when an investigator has a sample of observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and the joint probability density function (pdf) or the joint probability mass function (pmf) of the sample observations under two competing hypotheses H_0 and H_1 is given by $f_0(\boldsymbol{x})$ and $f_1(\boldsymbol{x})$, respectively. The NPL states that

- Any test that is given by
  - Reject H_0 if $\frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} > k$
  - Do not reject H_0 if $\frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} < k$

  for some $k \geq 0$ is the most powerful for its size for testing H_0 versus H_1, and
- Given $0 \leq \alpha \leq 1$, there exists a level-α test of the above form.

Note that $\frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})}$, which is the ratio of the likelihoods under H_1 and H_0, is often referred to as the *likelihood ratio*, and plays a vital role in rejecting the null hypothesis according to the NPL. Also note that the lemma applies only to the case of simple null and alternative hypotheses, that is, to the case when both $f_0(\boldsymbol{x})$ and $f_1(\boldsymbol{x})$ are completely known, that is, do not involve any unknown parameters.

Appendix B: The Assumptions on the Ability Distribution and the Estimated Ability Distribution

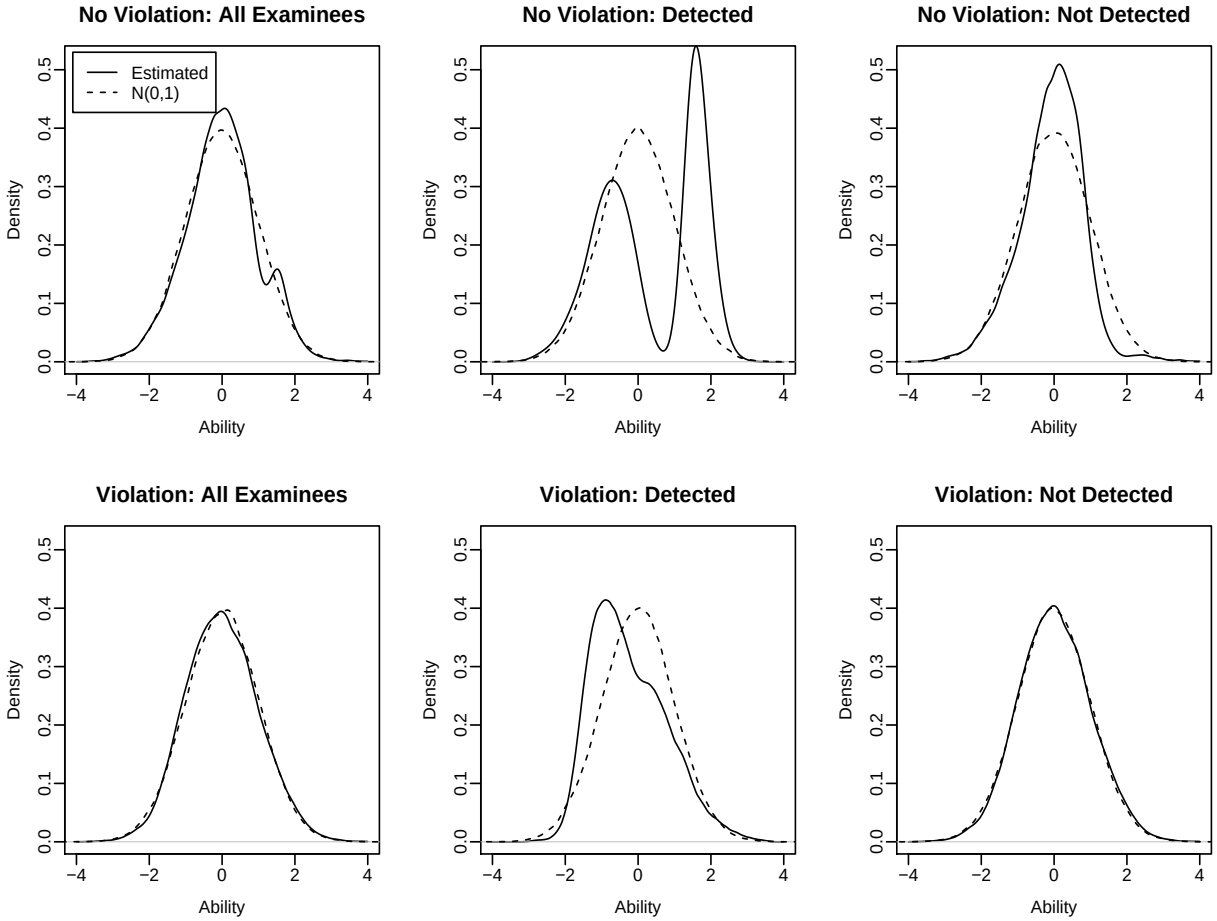

Figure B1. The estimated ability distributions when the assumption on the ability distribution made by the DLZS is violated (top row) and not violated (bottom row).

Figure B1 shows the estimated ability distributions for the full sample of examinees (two leftmost panels), the examinees with statistically significant values of the DLZS ("Detected"; two middle panels), and the examinees with non-significant values of the DLZS ("Not Detected"; two rightmost panels) for two simulation conditions under item preknowledge on a 100-item test on which 30 items were compromised. The R package "mirt" (Chalmers, 2012) was used to fit the 2PLM to the data and then to simulate five

plausible values for each examinee—each estimated ability distribution is essentially a density plot using all of these plausible values for the appropriate group of examinees. The density of the standard normal distribution is shown using a dashed line in all panels of Figure B1.

The top three panels correspond to a case when the true ability distribution is standard normal for both the non-cheaters and cheaters and yet the estimated ability distribution (solid line) is non-normal. The items were simulated to be difficult in general for the examinee sample and the second mode in the top left panel for $\theta \approx 1.6$ corresponds to the cheaters' superior performance on several difficult items that are compromised.

The three bottom panels correspond to a case where the true ability distribution is $\mathcal{N}(0, 1)$ for non-cheaters and $\mathcal{U}(-3, 0)$ for cheaters, the same assumptions that were made to simulate some of the data in the Simulations section of this paper, and yet the estimated ability distribution for the full sample is very close to standard normal.

Even though the top three panels correspond to the case of no violation of the assumption on the ability distributions while the bottom three panels correspond to a violation, the estimated ability distributions in all of the three panels in the top row of the figure appear non-normal while the estimated ability distribution in only the middle panel of the bottom row appears non-normal. Thus, Figure B1 shows that the normality or otherwise of the estimated ability distribution may not provide adequate evidence regarding the violation of the assumption of normality of the true ability distributions.

Appendix C: The ROC Curve for One Simulation Condition

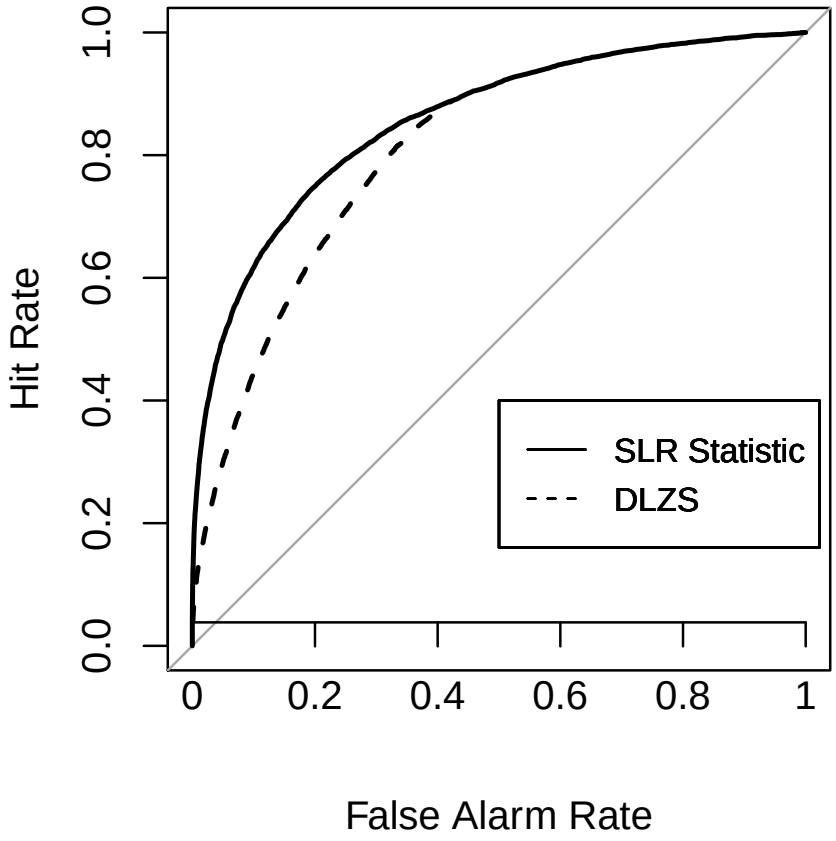

Figure C1. The ROC curve for the DLZS and the SLR statistic for one simulation condition.

Figure C1 shows the ROC curves for the SLR statistic (solid line) and the DLZS (dotted line) for the case of 20 compromised items and 20% aberrant examinees under the standard normal ability distribution and the "more variable success probability" condition. A diagonal line is shown for convenience. The curve for the SLR statistic is above that of the DLZS, which indicates that the former is more powerful than the latter irrespective of the level of significance for this simulation case.

Appendix D: Results for Compromised Items that Are not Precisely Known

One hundred data sets involving 100 items and 2,000 non-aberrant examinees were simulated as in the Simulations section of this paper under the condition of 20% cheaters and 30 compromised items. However, while simulating the data, it was assumed that among the 20% cheaters in a data set, half had preknowledge of 15 of the 30 compromised items and the other half had preknowledge of the other 15 compromised items. Such a situation may arise when, for example, two sets of common/anchor items[9] appeared on two different websites and were potentially exposed to two different groups of examinees. In these simulations, the true ability distributions of both the cheaters and non-cheaters were assumed to be the standard normal distribution and it was assumed that those with preknowledge of a compromised item had a probability of 0.95 of correctly answering the item (note that under these conditions, the DLZS was more powerful than the SLR statistic—see Figure 2).

While analyzing the data, the combined set of 30 compromised items was treated as $\mathcal{C}$ and the set of the remaining 70 items was treated as $\mathcal{U}$. The truncated ROC areas of the SLR statistic and the DLZS for this simulation condition were 0.72 and 0.62, respectively. Thus, the SLR statistic is much more powerful than the DLZS under this condition. The application of the DLZS to a data set under this condition involves the assumption that the success probability of a cheater on each of the 30 compromised items is 0.95, whereas, the true success probability is 0.95 for only 15 of those items. Thus, the DLZS is testing against the wrong alternative hypothesis (that makes an assumption on the cheaters that is stronger than what is appropriate) and hence cannot be expected to be very powerful. In contrast, in the application of the SLR statistic to a data set under this condition, the alternative hypothesis is that the cheaters performed comparatively better on the set of 30 compromised items than on the 70 uncompromised items, which is correct even when the cheaters have preknowledge of only 15 out of the 30 compromised items.

[9] *Common items* are items that are used for equating the scores on a new form of a test to those on an old form, are often reused, and are occasionally compromised, as mentioned in Drasgow et al. (1996).