Three New Methods for Analysis of Answer Changes

Educational and Psychological
Measurement

2016, Vol. 77(1) 54–81

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416632287

journals.sagepub.com/home/epm

Sandip Sinharay¹ and Matthew S. Johnson²

Abstract

In a pioneering research article, Wollack and colleagues suggested the “erasure detection index” (EDI) to detect test tampering. The EDI can be used with or without a continuity correction and is assumed to follow the standard normal distribution under the null hypothesis of no test tampering. When used without a continuity correction, the EDI often has inflated Type I error rates. When used with a continuity correction, the EDI has satisfactory Type I error rates, but smaller power compared with the EDI without a continuity correction. This article suggests three methods for detecting test tampering that do not rely on the assumption of a standard normal distribution under the null hypothesis. It is demonstrated in a detailed simulation study that the performance of each suggested method is slightly better than that of the EDI. The EDI and the suggested methods were applied to a real data set. The suggested methods, although more computation intensive than the EDI, seem to be promising in detecting test tampering.

Keywords

erasure analysis, generalized binomial model, Markov chain Monte Carlo, test fraud, test security

Cheating or other malfeasant behaviors during tests can reduce the validity of the interpretations of test scores and cause harm to other test takers, particularly in competitive situations in which test takers’ scores are compared (e.g., American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014, p. 132). Standards 6.6 and 8.11 of the

¹Pacific Metrics Corporation, Monterey, CA, USA

²Columbia University, New York, NY, USA

Corresponding Author:

Sandip Sinharay, Pacific Metrics Corporation, 1 Lower Ragsdale Drive, No. 150, Monterey, CA 93940, USA.
Email: ssinharay@pacificmetrics.com

Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014) include the recommendations 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
- testing programs may use technologies during scoring to detect possible irregularities such as computer analyses of erasure patterns.

Naturally, there is a growing interest in *erasure analysis*, which comprises analyses of erasure patterns in an attempt to detect test tampering. Erasures usually refer to erasing answers and choosing new answers (or, in other words, changing answers) on paper-and-pencil tests although examinees have been found to change answers in computer-based tests as well (e.g., Tiemann & Kingston, 2014). Henceforth, the term *analysis of answer changes* will be used instead of the term erasure analysis.

While research on answer changes (ACs) has often focused on groups of examinees (e.g., Maynes, 2013), new methods for analysis of ACs at the examinee level have recently been suggested by Belov (2015), van der Linden and Jeon (2012), van der Linden and Lewis (2015), and Wollack, Cohen, and Eckerly (2015). The focus in this article will be on the methodology of Wollack et al. (2015), which is based on an *erasure detection index* (EDI) and is quite simple conceptually and computationally; furthermore, the EDI was found to mostly have satisfactory Type I error rates and power in detailed simulations by Wollack et al. (2015). The EDI can be used with or without a continuity correction and is computed using item response theory (IRT).

Wollack et al. (2015) assumed that the EDI follows the standard normal distribution under the null hypothesis of no fraudulent ACs (or, equivalently, no test tampering). This assumption leads to the simplicity of the methodology, but may be violated, especially when the number of ACs is small. As a consequence, Wollack et al. (2015) found the Type I error rate of the EDI without a continuity correction (used with the assumption of a standard normal null distribution) to be inflated in several cases. The EDI with a continuity correction does not suffer from the problem of inflation of the Type I error rate of the EDI; however, the power of the EDI with a continuity correction is considerably smaller than that of the EDI without a continuity correction. Thus, there is a scope of further research on EDI and analysis of ACs in general.

This article suggests several methods for analysis of ACs with the goal of detecting test tampering. The suggested methods do not rely on the assumption of a standard normal null distribution. A detailed simulation study is performed to compare the Type I error rates and power of the suggested modified versions with those of the EDI.

The next section includes some background material including a review of the EDI. Three new methods for analysis of ACs (with the goal of detecting test tampering) are described in the Methods section. In the Simulation Study section, the Type I

error rates and power of the suggested methods are compared with those of the EDI. In the Application section, the EDI and the suggested methods are applied to a real data set. Conclusions and recommendations are provided in the last section.

As in Wollack et al. (2015), this article focuses only on dichotomous items and involves the assumption that the item parameters are known.

Background

Review of the Erasure Detection Index

Let us consider a test that consists of only dichotomous items whose parameters are assumed known (and are equal to the estimates computed from a previous calibration using a unidimensional IRT model). Let us consider one examinee who changed answers to a set $\mathcal{C}$ of items. Let $N_{\mathcal{C}}$ denote the number of items in $\mathcal{C}$. Let $\bar{\mathcal{C}}$ denote the set of items on which the examinee did not change answers.¹ Let W denote the raw score of the examinee on the items in $\mathcal{C}$. Note that W is also the *number of wrong-to-right* (WTR) ACs² and is often referred to as the *WTR score*. Let μ and σ , respectively, denote the expected value and standard deviation (SD) of W given the true ability parameter (θ) of the examinee. If θ is known, then μ can be computed as the sum of the probabilities of correct answers on the items in $\mathcal{C}$. For example, if the three-parameter logistic model (3PLM) is assumed to fit data from the test, then

$$\mu = \sum_{j=1}^{N_{\mathcal{C}}} P_j(\theta), \quad \text{where } P_j(\theta) = c_j + (1 - c_j) \frac{\exp[a_j(\theta - b_j)]}{1 + \exp[a_j(\theta - b_j)]}, \quad (1)$$

where a_j , b_j , and c_j respectively are the slope, difficulty, and guessing parameters of item j . Furthermore,

$$\sigma = \sqrt{\sum_{j=1}^{N_{\mathcal{C}}} P_j(\theta)[1 - P_j(\theta)]}. \quad (2)$$

However, θ is unknown for real data. Therefore, Wollack et al. (2015) recommended estimating θ based on the scores on the items in $\bar{\mathcal{C}}$. Let us denote this estimate as $\hat{\theta}_{\bar{\mathcal{C}}}$. The estimate $\hat{\theta}_{\bar{\mathcal{C}}}$ is robust to potentially fraudulent ACs.

The estimated mean and SD, denoted respectively by $\hat{\mu}$ and $\hat{\sigma}$, are obtained by replacing θ by $\hat{\theta}_{\bar{\mathcal{C}}}$ in Equations (1) and (2).

The EDI is then defined as

$$\text{EDI} = \frac{W - \hat{\mu} + c}{\hat{\sigma}}. \quad (3)$$

The quantity c , which represents a continuity correction, was assumed to be equal to 0 or -0.5 by Wollack et al. (2015) who assumed that the EDI follows the standard normal distribution under the null hypothesis of no fraudulent ACs. The null

hypothesis is rejected and an examinee is flagged for potentially fraudulent ACs if the examinee's EDI is a large positive number. For example, one would flag the examinees whose EDIs are larger than 2.33 if the significance level (or α level) of .01 is used.

The Choice of the Item Response Theory Model With the Erasure Detection Index

Wollack et al. (2015, p. 934) stated that any IRT model appropriate for the data may be used to compute the EDI, which makes their methodology quite flexible. In practice, it would be most convenient to use the EDI with the IRT model that is used in calibrating and scoring of the test in concern. If no IRT model is used in calibrating and scoring of the test, then the most convenient or appropriate IRT model (from the viewpoint of, e.g., software availability and IRT model fit) can be used to compute the EDI. In their simulation study, Wollack et al. (2015) used the nominal response model (NRM; Bock, 1972) with the EDI. Here, we use the 3PLM with the EDI in the simulation study that is described later.

A Further Look at the Erasure Detection Index

Wollack et al. (2015) found the EDI with $c=0$ (i.e., the one without a continuity correction) to have inflated Type I error rates in several cases, especially at small significance levels and for examinees with low ability levels. One explanation of this phenomenon is that the number of ACs is 4 or fewer for more than 97% of the examinees in the simulations of Wollack et al. (2015) whereas the standard normal null distribution of the EDI is expected to hold only when the number of ACs is large. Wollack et al. (2015) acknowledged that these results on the Type I error inflation of the EDI are troubling because test tampering is most likely to occur for examinees with low ability and is likely to be detected at small significance levels. They showed that when $c = -0.5$ (i.e., when a continuity correction is employed), the Type I error rates of the EDI are not inflated. However, when a continuity correction is employed, the EDI has substantially smaller power compared with that without a continuity correction.

In addition, the value of the EDI can be extreme for an examinee who changed answer only once; Qin and Cohen (2013) provided some examples of large absolute values of the EDI for those with one AC. Let us consider an examinee who changed answer only on item j and the changed answer is correct. For the examinee, $W=1$, $\hat{\mu} = P_j(\hat{\theta}_{\bar{c}})$, $\hat{\sigma} = \sqrt{P_j(\hat{\theta}_{\bar{c}})[1 - P_j(\hat{\theta}_{\bar{c}})]}$, and, from Equation (3), the EDI without a continuity correction becomes

$$\text{EDI} = \frac{1 - P_j(\hat{\theta}_{\bar{c}})}{\sqrt{P_j(\hat{\theta}_{\bar{c}})[1 - P_j(\hat{\theta}_{\bar{c}})]}} = \sqrt{\frac{1 - P_j(\hat{\theta}_{\bar{c}})}{P_j(\hat{\theta}_{\bar{c}})}} = \sqrt{\frac{1}{P_j(\hat{\theta}_{\bar{c}})} - 1}. \quad (4)$$

If a significance level of .01 is used, one would flag the examinee when his or her EDI is larger than 2.33, that is, from Equation (4), when $P_j(\hat{\theta}_{\bar{C}}) < 0.155$. For example, if $a_j = 1$, $b_j = 1$, and $c_j = 0.1$, then a flagging would occur whenever $\hat{\theta}_{\bar{C}}$ is -1.75 or smaller. Because an examinee, however weak, may answer a multiple-choice item just by random guessing, flagging based on only one AC seems too harsh.³ If a continuity correction is used with the EDI, then it is much less likely to flag an examinee based on one correct AC, but possible, for example, at significance level of .05 for an item with $a_j = 1$, $b_j = 1.0$, and $c_j = 0.05$ whenever $\hat{\theta}_{\bar{C}} < -1.7$.

Methods

Rationale Behind the Methods

Instead of assuming that the EDI follows the standard normal null distribution, it is possible to perform exact probability calculations to test the null hypothesis of no test tampering. The first two of the following methods involve exact probability calculations. Bayesian approaches are becoming increasingly popular in educational measurement (e.g., Sinharay, 2016). The third method involves a Bayesian approach. All the suggested methods, like the EDI, require only the final answers of the examinee (i.e., do not require the answers before the changes) and can be used with any IRT model that is appropriate for the data.

A Method Based on the Generalized Binomial Model

Wollack et al. (2015, p. 951) mentioned that instead of assuming that the EDI follows a normal distribution, one could use the generalized binomial model, as in van der Linden and Sotaridona (2006) and van der Linden and Jeon (2012). To apply the generalized binomial model, one observes that the WTR score of an examinee given $\hat{\theta}_{\bar{C}}$ follows the generalized binomial distribution. Then one computes the p -value for an examinee as the probability of a WTR score that is larger than or equal to the examinee's observed WTR score given that the examinee ability is equal to $\hat{\theta}_{\bar{C}}$, that is, computes

$$p\text{-value} = P(\text{WTR score} \geq W | \hat{\theta}_{\bar{C}}) = \sum_{k=W}^{N_C} P(\text{WTR score} = k | \hat{\theta}_{\bar{C}}), \quad (5)$$

using the generalized binomial distribution or compound binomial distribution (e.g., Lord, 1980). Thus, if an examinee changed answers on 5 items (i.e., $N_C = 5$) and has an observed WTR score of 4 ($W = 4$), the p -value under this method would be equal to the probability of a WTR score equal to 4 plus the probability of a WTR score equal to 5, where both probabilities are computed with the same 5-item set $\bar{C}$. The terms under the summation sign in the rightmost expression of Equation (5) are computed using the iterative approach of Lord and Wingersky (1984), which provides the probability, given θ , of each possible raw score on a set of binary items as a

function of the item parameters of those items. A small p -value leads to the rejection of the null hypothesis of no fraudulent ACs. This method is somewhat similar to the method of van der Linden and Jeon (2012) that also uses the generalized binomial distribution, but does not require the fitting of a second-stage IRT model as in the latter, and will be referred to as the *generalized binomial method* henceforth.

A Method Based on Exact Probabilities and Score Patterns

A more computation-intensive method involves the computation of the p -value for an examinee as the probability of a score pattern⁴ on the items in $\mathcal{C}$ that is as extreme or more extreme than the examinee's (observed) score pattern on those items conditional on $\hat{\theta}_{\bar{c}}$. That is,

$$p\text{-value} = P(\mathbf{x} \in \mathcal{R} | \hat{\theta}_{\bar{c}}), \quad (6)$$

where $\mathbf{x}$ is a score pattern on the items in $\mathcal{C}$ and $\mathcal{R}$ is the collection of the score patterns on the items in $\mathcal{C}$ that are as extreme or more extreme than the examinee's observed score pattern on those items. To apply this method, one requires a definition of when a score pattern on a set of items can be considered more extreme than another score pattern on the same set of items. Because the primary interest in methods for test tampering is to prevent examinees from obtaining a larger score than what they deserve, the ability estimate computed from a score pattern was used to determine how extreme the pattern is. That is, score pattern $\mathbf{x}_2$ is considered more extreme than score pattern $\mathbf{x}_1$ if the ability estimate computed from $\mathbf{x}_2$ is larger than that from $\mathbf{x}_1$. Then the p -value in Equation (6) is computed as the sum of the conditional probabilities (given $\hat{\theta}_{\bar{c}}$) of the score patterns in $\mathcal{R}$. The probability of a score pattern $\mathbf{x} = (x_1, x_2, \dots, x_{N_C})$ given $\hat{\theta}_{\bar{c}}$ is computed, as is customary for an IRT model, as

$$\prod_{j=1}^{N_C} P_j(\hat{\theta}_{\bar{c}})^{x_j} (1 - P_j(\hat{\theta}_{\bar{c}}))^{1-x_j}.$$

Let us consider an examinee who changed answers on five items and let us suppose that the examinee's observed score pattern on these items is (0, 1, 1, 1, 1), that is, the examinee answered the first of these five items incorrectly and the remaining correctly. Let us suppose that the ability estimate computed from the examinee's score pattern is 1.45. Let us also suppose that among the ability estimates computed from the $2^5 - 1 = 31$ other possible score patterns on these five items, the ability estimates computed from the all-correct score pattern and the score pattern (1, 1, 1, 1, 0) are larger than 1.45, and those computed from all the other score patterns are smaller than 1.45. Then, the p -value for this examinee would be the sum of the conditional probabilities (given $\hat{\theta}_{\bar{c}}$) of the three score patterns (0, 1, 1, 1, 1), (1, 1, 1, 1, 0), and (1, 1, 1, 1, 1). This method will be referred to as the *exact method* henceforth.

A Method Based on Posterior Predictive Model Checking

Instead of assuming that the null distribution of the EDI is the standard normal distribution, it is possible to approximate the null distribution of the EDI by its posterior predictive distribution (PPD; e.g., Gelman, Carlin, Stern, & Rubin, 2003). In this method, one simulates a large number of draws from the PPD of the EDI and obtains the p -value of the EDI as the proportion of these draws that are larger than the observed EDI. The resulting p -value is referred to as the posterior predictive p -value and this method is referred to as the posterior predictive model checking (PPMC) method (e.g., Gelman et al., 2003). A Markov chain Monte Carlo (MCMC) algorithm (e.g., Gelman et al., 2003) is used to simulate draws from the PPD of the EDI. To implement this method, one replicates the following steps a large number of times:

- Simulate a draw from the posterior distribution of the examinee ability based on the scores on the items in $\bar{C}$ using the MCMC algorithm
- Simulate scores on all items on the test using the above draw; these are referred to as the replicated scores
- Compute the replicated value of the EDI based on the replicated WTR score (where C for the examinee is the same as that for the original data) and an ability estimate based on the replicated scores on $\bar{C}$. This replicated value of the EDI is a draw from the PPD of the EDI

The posterior predictive p -value of the EDI is the proportion of replicated values of the EDI that are equal to or larger than the observed EDI. The PPMC method has been applied to assess the fit of IRT models by researchers such as Glas and Meijer (2003), Sinharay (2015), and Sinharay, Johnson, and Stern (2006)⁵ although its application to analysis of ACs is lacking. The PPMC method is usually slightly conservative compared to frequentist methods (e.g., Bayarri & Berger, 2000). Also, researchers such as Toribio and Albert (2011) found that the use of the PPMC method led to satisfactory Type I error rates of IRT model-fit statistics that have inflated Type I error rates under a frequentist approach. Therefore, this method may overcome the aforementioned problem of inflation of the Type I error rate of the EDI without a continuity correction. This method will be referred to as the *PPMC method* henceforth.

A Simulation Study

A detailed simulation study, somewhat similar to that in Wollack et al. (2015), was performed to compare the Type I error rate and power of the EDI to those of the suggested methods.

Design of the Simulation

The simulation study involved a long test comprising 50 dichotomous items, as in Wollack et al. (2015), and involved one large data set comprising 1,000,000 examinees.

While Wollack et al. (2015) used the NRM (Bock, 1972) in their simulations, we used the 3PLM as the IRT model because the logistic IRT models are more popular in operational testing than the NRM and neither our suggested methods nor the EDI require the use of the NRM. The item parameters were assumed known and equal to the estimates from a large-scale survey assessment (that employs the 3PLM) and not estimated during the analysis of ACs. The true abilities of the examinees were simulated from the standard normal distribution.

To compute the Type I error rates of the methods, score patterns that represent benign ACs were generated. Two types of benign ACs were generated: string-end ACs and random ACs. Each examinee was randomly assigned to either string-end AC or random AC with a probability of 0.5 for each. To compute the power of the methods, score patterns that represent fraudulent ACs were generated.

Because the item parameters are assumed known and each method is applied to one examinee at a time, the Type I error rate and power of the methods do not depend on the number of examinees in the data set or the proportion of examinees whose scores involved benign or fraudulent ACs—so these numbers were set large so as to be able to compute the Type I error rate and power for each type of AC precisely. A real data set would not include the large proportion of examinees with fraudulent ACs or benign ACs as in our simulation study—these proportions would probably be more like those used in the simulation study of Wollack et al. (2015).

String-end ACs occur when the examinees feel that they are running out of time and randomly guess on the remaining items, but later return to the previously guessed items and answer them on merit, changing some answers if needed. For each examinee assigned to string-end ACs, we simulated a random number Y from the binomial distribution with number of trials = 50 (i.e., equal to the number of items) and success probability = 0.25. Then, it is assumed that the examinee made string-end ACs on the last Y items of the test. Thus, on average, an examinee who is assigned to string-end ACs makes 12.5 such ACs.

Random ACs occur when an examinee accidentally chooses one answer, but, on reconsideration, changes it to another answer. Random ACs occur on very few items (e.g., Wollack et al., 2015). Therefore, for each examinee assigned to random ACs, we simulated a random number Y from the binomial distribution with number of trials = 50 and success probability = .02. Then, it is assumed that the examinee made random ACs on Y randomly chosen items on the test. Thus, on average, an examinee who is assigned to random ACs makes one such AC.

Wollack et al. (2015) considered another type of benign ACs that are referred to as misalignment ACs. However, misalignment ACs are similar to the string-end ACs and Wollack et al. (2015) found the Type I error rates for misalignment ACs and

string-end ACs to be very close (see their tables 3 and 5). Therefore, misalignment ACs were not considered here.

Fraudulent ACs were simulated for each examinee with a probability of .5. When an examinee was selected for fraudulent ACs, those ACs were simulated in addition to the benign ACs. Thus, in the data set, about 25% score patterns involved only string-end ACs, about 25% involved only random ACs, about 25% involved both random ACs and fraudulent ACs, and 25% involved both string-end ACs and fraudulent ACs. Among those assigned to fraudulent ACs, one-third were randomly assigned to fraudulent ACs on 5 items, one-third were randomly assigned to fraudulent ACs on 10 items,⁶ and the remaining one-third were assigned to fraudulent ACs on a variable number of items. For an examinee assigned to fraudulent ACs on a variable number of items, the number of fraudulent ACs is the integer nearest to the difference between 27.3 (that is the true raw score on the test corresponding to true θ of -0.126 , which is the 45th percentile of the standard normal distribution) and his or her true raw score. Wollack et al. (2015, p. 938) used this strategy because, on average, 45% of students across the United States are not proficient in mathematics. Thus, the fraudulent ACs on a variable number of items represents a situation where an administrator changed just enough items to help a student get from nonproficient to proficient.

Computation

For each examinee, the following steps were performed in the simulation study:

- Simulate the true ability from the standard normal distribution
- Simulate the scores on all the items under the 3PLM using the above true ability and the aforementioned true item parameters
- Simulate benign ACs; this step involves marking responses to Y items (where Y was simulated from a binomial distribution) as those involving ACs, but actually not changing the answers (in other words, retaining the scores generated in the above step). The marked responses are at the end of the test for string-end ACs, but are spread throughout the test for random ACs.
- Simulate 5, 10, or a variable number of fraudulent ACs; to simulate E fraudulent ACs, choose E incorrect answers among all incorrect answers for the examinee with equal probability and change them to correct answers. Note that for some examinees (especially those with large ability) assigned to fraudulent ACs, the number of incorrect answers may be D that is smaller than E and could even be equal to 0; in such cases, only D incorrect answers were changed to correct; later, this is reflected in the decrease in the power of the methods as examinee ability increases. If the number of ACs (combining both benign and fraudulent ACs) is 0, discard the data and go back to the third step above.

- Compute the p -values for the EDI, both with and without continuity correction, and for the suggested methods. This step requires computation of the ability estimate of the examinee for the items in $\bar{\mathcal{C}}$, the EDIs, running the MCMC algorithm required by the PPMC method, and the computation of ability estimates for all the score patterns in $\mathcal{C}$ (in the exact method).

A Fortran 90 computer program written by the authors was used for all the computations including the computation of the ability estimates and implementation of the MCMC algorithm. The prior distribution for the examinee ability was assumed to be the standard normal distribution. The weighted maximum likelihood (WMLE; Warm, 1989) of the examinee ability was used as the ability estimate because it was always finite whereas the maximum likelihood estimate of ability can be infinite, for example, for someone with all correct answers. The WMLEs were computed using the Newton–Raphson algorithm. The runtime was too long for the exact method whenever the number of ACs was larger than 20 (that happened for about 1.7% of the examinees in the simulated data set)—so the results reported for the method are based on examinees with up to 20 ACs; this should not be of much concern in practice because, for real data sets, the number of ACs per examinee is usually quite small (e.g., Primoli, Liassou, Bishop, & Nhouyvanisvong, 2011) and rarely exceeds 20; also, with ever-increasing power of computers, the runtime would become much smaller with faster computers in near future.

The Metropolis–Hastings algorithm (e.g., Gelman et al., 2003) was used as the MCMC algorithm in the PPMC method. The proposal/jumping distribution for the Metropolis–Hastings algorithm was chosen as a normal distribution with the previous draw as the mean. One chain of length 11,000 of the Metropolis–Hastings algorithm was run that ensures the convergence of the algorithm and the first 1,000 draws of the chain were discarded as burn-in, which leads to 10,000 draws from the posterior distribution of the ability for each examinee.

Distribution of the Erasure Detection Index Under the Null Hypothesis

The top left panel of Figure 1 shows the kernel-density estimates⁷ of the distributions of the values of the EDI both with and without a continuity correction (denoted as “EDI-CC” and “EDI-No CC”, respectively) for all the examinees whose responses involved only benign ACs. The standard normal distribution is also shown for comparison. The other three panels of the figure show similar plots for three subsets of those corresponding to the top left panel: those involving up to five ACs (top right), those with negative true ability (bottom left; investigators of ACs usually are more worried about score gains of low-ability examinees through ACs), and those involving more than five ACs (bottom right). The distribution of the EDI without a continuity correction is close to the standard normal distribution only for those with more than five ACs (bottom right panel); this is because, according to the central limit theorem for independent random variables (e.g., Rao, 1973, pp. 127-128), the null

Figure 1. Distribution of the erasure detection index.

Note. AC = answer change; CC = continuity correction.

distribution of the EDI is expected to be standard normal only for a large number of ACs. Given that the number of ACs is small in practice (e.g., Primoli et al., 2011), the figure provides some evidence against the use of the EDI. The distribution of the EDI with a continuity correction always lies to the left of that of the EDI without a continuity correction, which is expected given Equation (3). The distribution for those with negative true ability (bottom left panel) appears similar to that for all examinees (top left panel). The distributions of the EDIs have two or more modes in three panels because there are two modes in the distribution of the EDI for those who made exactly one benign AC—one mode for those whose changed answer was correct and the other for those whose changed answer was incorrect.⁸

Results on Type I Error Rates

As in Wollack et al. (2015), the examinees were divided into five quintile groups of equal size based on their true abilities (e.g., the examinees with the smallest 20% true abilities were in the first quintile group, etc.) and the Type I error rate and power of each method were computed for each quintile group.

Figure 2. Type I error rates of the methods for string-end ACs.

Note. AC = answer change; CC = continuity correction; EDI = erasure detection index; GBM = generalized binomial method; PPMC = posterior predictive model checking.

The bias and root mean squared error of the ability estimates were similar to those in tables 1 and 2 of Wollack et al. (2015) and are not repeated here.

Figure 2 shows the Type I error rates of the methods for significance levels of .05, .01, .001, and .0001 for those with only string-end ACs. In each panel, the Type I error rates for the five quintile groups for each method (shown using different point types as described in the top left panel; the exact method, PPMC method, and generalized binomial method are denoted as “Exact,” “PPMC,” and “GBM,” respectively) are joined using a dotted line. The significance level is denoted using a horizontal solid line in each panel—points above this line indicate inflated Type I error rates.

The standard errors associated with the Type I error rates shown in Figure 2 are roughly 0.001, 0.0004, 0.0001, and 0.00004, respectively, at levels of .05, .01, .001, and .0001. This is because about 250,000 examinees were assigned to only string-end ACs and, among them, 50,000 belonged to each quintile group, and, for example, when the level is .05, the corresponding standard error for a quintile group is roughly equal to $\sqrt{0.05 \times (1 - 0.05) / 50000} \approx 0.001$. In contrast, each Type I error rate for string-end ACs reported in table 3 of Wollack et al. (2015) was computed from 800 examinees and had corresponding standard errors of, for example, about 0.001 at level .001 and about 0.0001 at level .00001, making those reported rates less precise than ideal.⁹

The EDI without a continuity correction almost always has the largest Type I error rates among all the methods in Figure 2; also, the Type I error rates of the index are inflated at least for the first three quintile groups in each panel. This is the result that was referred to as *troubling* by Wollack et al. (2015). The Type I error rates of the exact method and the PPMC method occasionally exceed the nominal level in Figure 2, but are satisfactory according to Cochran's criterion for robustness (e.g., Wollack, Cohen, & Serlin, 2001, p. 394) that deems Type I error rates smaller than 0.06, 0.015, 0.0015, and 0.00015 to be satisfactory at levels .05, .01, .001, and .0001, respectively. The Type I error rate of the EDI with a continuity correction exceeds the nominal level in Figure 2 at level of .0001 for the third quintile group (and is unsatisfactory according to Cochran's criterion), but this is the only violation for that index. The Type I error rates of the generalized binomial method are always smaller than the nominal level. The Type I error rate of each method is mostly the largest for the first quintile group and the smallest for the fifth quintile group, which was also observed by Wollack et al. (2015).

Figure 3 shows the Type I error rates of the methods for those who only made random ACs. The EDI without a continuity correction has the largest Type I error rates among the methods in this figure as well, but its Type I error rates are not inflated (except for the first quintile group at level of .05) in this figure. The Type I error rates in Figure 3 are smaller compared with those in Figure 2, which was also observed in Wollack et al. (2015), mostly because random ACs occur on much fewer number of items on average (2) compared with string-end ACs (12.5) and a large value of EDI is quite unlikely for someone who made only a couple of ACs. Let us assume that an examinee in the first quintile group (and with true ability of only -2.0) made two ACs on items both of which have $a_j = 1$, $b_j = 0$, and $c_j = 0.2$. For this examinee, probability of a correct answer on each of these two items is about 0.30, $\hat{\mu}$ is about 0.59 and $\hat{\sigma}$ is about 0.65. Even if both these changed answers are correct, the EDI would be equal to about 2.21, which is smaller than the 99th percentile of the standard normal distribution—so the EDI of the examinee would not be significant at level of .01. We performed limited simulations where we simulated random ACs on 12.5 items on average and the corresponding Type I error rates rose to roughly the values observed in Figure 2.

Figure 3. Type I error rates of the methods for random ACs.
Note. AC = answer change; CC = continuity correction; EDI = erasure detection index; GBM = generalized binomial method; PPMC = posterior predictive model checking.

The phenomenon of the EDI (without a continuity correction) being significant for those with exactly one AC was mentioned earlier. In our simulations, roughly 20,000 examinees in each quintile group made exactly one benign AC and no fraudulent AC; Table 1 provides the number of statistically significant values of EDI without a continuity correction at four significance levels among those examinees. There are two significant values of EDI without a continuity correction even at level .001. While these two significant values lead to a Type I error rate of 0.0001 (among the 20,000 examinees in that quintile group), which is smaller than the significance level, our opinion is that it is problematic to have even one significant value based on only one benign AC at such a small significance level. The PPMC method is the only other method that led to any significant p -values for these roughly 20,000 examinees in

Table 1. The Number of Significant Values for Those With Exactly One Benign AC and No Fraudulent AC.

Quintile group	Level = 0.05		Level = 0.01		Level = 0.001		Level = 0.0001	
	EDI	PPMC	EDI	PPMC	EDI	PPMC	EDI	PPMC
1	1,056	512	40	33	2	0	0	0
2	359	16	3	0	0	0	0	0
3	66	0	0	0	0	0	0	0
4	2	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0

Note. AC = answer change; EDI = erasure detection index; PPMC = posterior predictive model checking.

each quintile group—Table 1 also provides the number of significant *p*-values for that method—these numbers are much smaller than those for EDI. Also, only 16 significant *p*-values for the PPMC method correspond to quintile groups 2 to 5. The EDI with a continuity correction, the generalized binomial method, and the exact method did not lead to a significant *p*-value for any of these examinees.

Results on Power

Figures 4 to 6 show the power of the methods for significance levels of .05, .01, .001, and .0001 for 5 fraudulent ACs, 10 fraudulent ACs, and a variable number of ACs, respectively. The standard error corresponding to any value of power shown in these figures is smaller than 0.003. The power of the EDI without a continuity correction is not shown in Figures 4 to 6 because of the inflated Type I error rates of the method in Figure 2. The three figures show that the EDI with a continuity correction has the smallest power except in the bottom two panels of Figures 5 and 6 where the PPMC method has the smallest power.¹⁰ The exact method has the largest power in Figure 5 (i.e., under severe test tampering) and the PPMC method has the largest power in Figure 4. The generalized binomial method has slightly larger power than the EDI with a continuity correction—this result agrees with the finding of Zopluoglu and Davenport (2012) that a method based on the generalized binomial distribution has slightly larger power compared with the ω index¹¹ (Wollack, 1997) in detecting answer-copying.

Discussion on the Comparative Performance of the Different Methods

Taking results on both Type I error rates and power into consideration, the exact method, which is based on exact probability calculations conditional on the ability estimate $\hat{\theta}_{\bar{C}}$, or the PPMC method are probably the best among all the methods considered in this article. The Type I error rates and power of both of these methods are satisfactory in comparison to the other methods. Among these two methods, the exact

Figure 4. Power of the methods for five fraudulent ACs.

Note. AC = answer change; CC = continuity correction; EDI = erasure detection index; GBM = generalized binomial method; PPMC = posterior predictive model checking.

method has slightly larger Type I error rates and slightly larger power compared to the PPMC method. Thus, if the control of the Type I error rate is an absolute priority, then one would probably choose the PPMC method as the most appropriate. However, if one is willing to accept slightly inflated Type I error rates in exchange of larger power or would like to avoid a significant p -value for an examinee who made exactly one AC, one would probably choose the exact method. The extent of inflation of Type I error rates is probably too large for EDI without a continuity correction for it to be used in a practical application, especially given the severe consequences of the flagging of an examinee for potential test tampering. While the EDI with a continuity correction does not suffer from the inflation of Type I

Figure 5. Power of the methods for 10 fraudulent ACs.
Note. AC = answer change; CC = continuity correction; EDI = erasure detection index; GBM = generalized binomial method; PPMC = posterior predictive model checking.

error rates, its power is the smallest among all the methods in most cases—so this index should probably be used only when computational simplicity is of prime importance.

In their simulation study, Wollack et al. (2015) used the NRM. Some results from a simulation study using the NRM are provided in the appendix; those results show that the comparative performance of the EDI and the suggested methods is very similar to those from the simulation study described above (except that the EDI with a continuity correction has inflated Type I error rates for quintile group 1 at level .01 or smaller), which demonstrates that the suggested methods improve over the EDI irrespective of the IRT model used.

Figure 6. Power of the methods for variable number of fraudulent ACs.

Note. AC = answer change; CC = continuity correction; EDI = erasure detection index; GBM = generalized binomial method; PPMC = posterior predictive model checking.

Application to Real Data

The methods were applied to a real data set that was used in van der Linden and Jeon (2012) and van der Linden, Jeon, and Ferrara (2011).

Data Set and Analyses

The data set consisted of the responses of 2,555 Grade 3 students to 65 mathematics items belonging to a large-scale assessment. On completion of the test, the students were encouraged to review their answers before turning in their answer sheets. The items were part of a larger set that had been calibrated under the 3PLM using

marginal maximum likelihood estimation with a standard normal prior on the examinee ability and showed excellent fit (van der Linden & Jeon, 2012). The answer sheets of the students were scanned using the option of erasure detection. The number of ACs is 0 for 217 examinees. The first quartile, median, third quartile, and maximum of the number of ACs were 2, 4, 9, and 39, respectively. Among those who made at least one AC, the first quartile, median, third quartile, and maximum of the WTR score were 1, 3, 6, and 36, respectively, and those of the proportion-correct WTR score (that is the WTR score divided by the number of ACs) were 0.5, 0.67, 0.91, and 1.00, respectively. Thus, most of the ACs resulted in correct answers, which agrees with table 2 of van der Linden et al. (2011) computed from this data set. The p -values for all the methods were computed for each examinee in the data set who made any AC. It took about 8 minutes to perform all the computations on a standard desktop computer.

Results

The proportion of significant p -values for the different methods for the five quintile groups (created on the basis on the abilities estimated from the items with no AC) among those with at least one AC at significance levels of .05, .01, .001, and .0001 are shown in Figure 7. The proportions of significant p -values are larger than the nominal level for each method at each level for all but the fifth quintile group. It is unknown which ACs are benign and which are fraudulent for these data, but the patterns of the figure mostly agree with the results from the simulation studies that the percentage of significant values is the largest for the exact method and the PPMC method compared with the other methods and the percentage decreases as ability estimate increases.

The top left panel of Figure 8 shows the values of the EDI with continuity correction (Y -axis) versus the number of ACs (X -axis). There is a positive association (correlation coefficient is 0.45) between these two variables. For each examinee with at least one AC, we computed the difference between the proportion-correct WTR score and the proportion-correct score on the items with no AC. One would expect that an examinee is more likely to be flagged when this difference is large. The top right panel of Figure 8 shows this difference in proportion correct along the X -axis and the EDI with continuity correction along the Y -axis for a random subsample of 500 examinees drawn from the sample.¹² The value of 2.33 of the EDI, which is the 99th percentile of the standard normal distribution, is shown using a dashed horizontal line. The bottom left panel shows the same difference in proportion correct scores along the X -axis and the p -value for the exact method along the Y -axis for the same subsample. The p -value of .01 is shown as a dashed horizontal line. As expected, as the difference in proportion correct increases, the EDI with continuity correction increases (the corresponding correlation coefficient is 0.80) and the p -value for the exact method decreases. The bottom right panel shows the p -values for the EDI with continuity correction versus the p -values for the exact method for the same

Figure 7. Percentage of significant p -values for the real data for the different methods.
Note. CC = continuity correction; EDI = erasure detection index; GBM = generalized binomial method; PPMC = posterior predictive model checking.

subsample, only showing p -values up to .10 for convenience of viewing, and includes a dashed diagonal line. More points below the diagonal line in the panel indicate that the p -values for the exact method are more often significant than those for the EDI with continuity correction, which agrees with Figure 7.

Figure 9, which is somewhat like figure 2 of van der Linden and Jeon (2012), shows the scores and patterns of ACs of four examinees. Each panel corresponds to one examinee. The titles of the panels indicate the number of ACs and the aforementioned difference in proportion corrects of the examinees. The bottommost panel corresponds to the examinee for whom the number of ACs was the largest (39). In each panel, the item position is shown along the X -axis and a horizontal solid line is drawn in the middle. For an item position, a vertical black bar is drawn connecting the horizontal solid line and the top of the panel if the item score is 1, a black dot is drawn on the horizontal solid line if the item score is 0, and a vertical gray bar is drawn connecting the horizontal solid line and the bottom of the panel if an AC occurred for the item. Thus, too many black bars on top of grey bars indicate many WTR ACs and would increase the chance of flagging of the examinee. The examinees represented in the top three panels, whose difference in proportion corrects is at least 0.28,

Figure 8. Some details about the real data.

Note. CC = continuity correction; EDI = erasure detection index.

answered almost all the items with ACs correctly and were flagged by all the methods at significance level of .01. The difference in proportion corrects for the examinee represented in the bottommost panel is negative (that is, the examinee did not benefit much from the ACs); thus, it is natural that the examinee was not flagged by any of the methods.

Conclusions

This article follows up on the research of Wollack et al. (2015) by suggesting several methods for analysis of ACs. The suggested methods seem to have superior combination of Type I error rates and power compared with the EDI with or without a continuity correction. Further insights (above and beyond Wollack et al., 2015) are provided on several aspects of the EDI such as (a) an explanation was provided on why its Type I error rate is smaller for random ACs compared with string-end ACs, (b) an explanation was provided of how the use of the EDI is problematic for examinees who made only one AC, (c) it was demonstrated that the null distribution of the EDI is close to the standard normal distribution for the examinees with a large number of ACs (Figure 1), (d) the Type I error rate and power of the EDI were

Figure 9. Details on four examinees for the real data.

computed with more precision compared with the study of Wollack et al. (2015), and (e) unlike in Wollack et al. (2015), a real data application of the EDI was included.

Also, this article examines the use of the EDI with the 3PLM, which is more popular in operational testing than the NRM used in the simulation of Wollack et al. (2015)—so this article would make the EDI more accessible to a wider audience.

Statistical indices for determination of test tampering are useful for providing confirming evidence of inappropriate behavior when evidence from other sources also exist, but the evidence provided by statistical indices is insufficient by itself. For example, Hanson, Harris, and Brennan (1994) commented that no statistical method on its own can provide conclusive proof that copying occurred (p. 25); the comment is true about fraudulent ACs as well. Researchers such as Tendeiro and Meijer

(2014, p. 257) recommended complementing statistical indices of detecting irregularities with other sources of information such as seating charts, video surveillance, or follow-up interviews. In addition, we think that if an index from analysis of ACs for an examinee is found to be statistically significant, then the examinee should be informed that his or her test could not be scored because of statistical irregularities and should be offered the opportunity to retake the test (van der Linden & Jeon, 2012) or should be given a timely opportunity to provide evidence that the score should not be cancelled (as mentioned in the Standard 8.11 of the Standards for Educational and Psychological Testing; American Educational Research Association et al., 2014).

There are several limitations of this paper and, consequently, several related topics can be further investigated. First, it is possible to compare other methods of analysis of ACs including those suggested by Belov (2015), van der Linden and Jeon (2012), and van der Linden and Lewis (2015) to the EDI and the suggested methods. Sinharay & Duong (2016) performed one such comparison. Second, while our simulation study was detailed, it is possible to perform more simulations, possibly with different true item parameters. Similarly, it is possible to consider applications of the EDI and the suggested modified versions to more real data examples. Third, the item parameters were treated as known in this article. It is possible to estimate item parameters and consider the effect of the estimation on the properties of the modified versions suggested here. However, Wollack et al. (2015, p. 951) noted that the estimation of item parameters would lead to negligible changes in the results on EDI and a similar result is expected to hold for the modified versions of EDI suggested here. Fourth, it is possible to consider additional ability estimates such as robust estimates (e.g., Mislevy & Bock, 1982) with EDI and its modified versions. Fifth, future research may investigate how the exact method can be implemented for a large number of ACs (note that the method was too time-consuming when the number of ACs was more than 20 in our simulations); Monte Carlo sampling from all possible response patterns is a possibility. Similarly, it is possible to explore the application of the randomized rule (e.g., Rao, 1973, p. 450) to the generalized binomial method for possible increase in power of the method. Finally, while Wollack and Eckerly (in press) have extended the EDI to detect test tampering at the class, school, or district level, it is possible to extend the suggested modified versions to those cases as well.

Appendix

A Simulation Study Using the Nominal Response Model

We performed another simulation study using the NRM (Bock, 1972). The design of the study was very similar to that of Wollack et al. (2015) with the exceptions that the sample size was 1,000,000, as in the simulation study with the 3PLM, and the proportions of examinees with different types of erasures are the same as those in the simulation study described earlier. A subset of the item parameters used in Wollack (1997) were used as the true item parameters. The runtime was too long for the exact

Figure A1. Type I error rates of the methods for string-end ACs under the NRM.

Note. AC = answer change; CC = continuity correction; EDI = erasure detection index; GBM = generalized binomial method; NRM = nominal response model; PPMC = posterior predictive model checking.

method whenever the number of ACs was larger than 8 (this number is smaller compared to that for the 3PLM because, for each item, the number of possible responses is 2 under the 3PLM, but is 5 under the NRM)—so the results reported for the exact method are based on examinees with up to 8 ACs.

Figure A1 shows the Type I error rates of the methods for those with string-end ACs and Figure A2 shows the power of the methods for those with ten fraudulent ACs. The Type I error rates of the methods in Figure A1 are similar to those in Figure 2 except that the EDI with a continuity correction has inflated Type I error rates for quintile group 1 at levels .01 or smaller and the extent of inflation at levels .001 and .0001 is quite severe even according to Cochran's criteria of

Figure A2. Power of the methods for 10 fraudulent ACs under the NRM.
Note. AC = answer change; CC = continuity correction; EDI = erasure detection index; GBM = generalized binomial method; NRM = nominal response model; PPMC = posterior predictive model checking.

robustness (e.g., Wollack et al., 2001, p. 394). Note that Wollack et al. (2015) attributed some instances of inflated Type I error rates of the EDI with a continuity correction to small sample sizes in their simulations. Figure A1, where the Type I error rates are estimated much more precisely than in Wollack et al. (2015), shows that the EDI with a continuity correction indeed has inflated Type I error rates (that cannot be explained by small sample sizes) in some instances. The power of the methods in Figure A2 are similar to those in Figure 5. Thus, the suggested modified versions seem to improve over the EDI irrespective of the IRT model used.

Authors' Note

Any opinions expressed in this publication are those of the authors and not necessarily of Pacific Metrics Corporation or Columbia University.

Acknowledgments

The authors would like to thank the editor George Marcoulides and the two anonymous reviewers for several helpful comments that led to a significant improvement of the article. The authors would also like to thank Minjeong Jeon for sharing a data set that was used in this article. Finally, the authors are grateful to James Wollack for his positive and helpful comments on an earlier version of this paper and for sharing a set of item parameters from his dissertation.

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. $\mathcal{C}$ and $\bar{\mathcal{C}}$ are nonoverlapping and their union is the set of all items administered to the examinee.
2. This is because a right-to-right AC is impossible for regular dichotomously scored multiple-choice items that involve only one correct answer option.
3. In practice, the EDI is likely to be used with a smaller significance level. However, given that there is no widely accepted guidance on how to use such indices and given that Wollack et al. (2015) has provided results at significance level of .05 and smaller, one cannot rule out the possibility of the use of the EDI at a level of .01.
4. A score pattern on a set of items is a combination of scores on the items of the set.
5. With the first two of these focusing on person fit, which may be used to detect cheating among other testing irregularities.
6. Wollack et al. (2015) also considered 15 fraudulent ACs, but found the power of the EDI for 15 items only slightly larger than that for 10 items; we found similar results for all the methods in limited simulations and do not consider 15 fraudulent ACs in our simulation.
7. Created using the function “density” in the R software (R Core Team, 2015).
8. To verify this, we repeated the following steps 5,000 times: (a) generate a θ from the standard normal distribution, (b) draw an item randomly from the 50 aforementioned items, (c) assume that the examinee changed answer only for this item and compute the EDI for this item with the above draw of the θ , once for a correct answer and once for an incorrect answer. A histogram of the resulting 10,000 EDIs was bimodal.

9. Wollack et al. (2015, p. 943) attributed the inflated Type I error rates of the EDI with a continuity correction in their study for a few cases (such as in Table 5) to the small sample sizes for those cases.
10. Some limited simulations shows that this could partially be due to the use of only 10,000 draws from the posterior distribution; the use of a larger number of draws could lead to larger power for small levels.
11. That is conceptually similar to the EDI.
12. Plotting of all the examinees led to an overcrowded plot—that is why a subsample was chosen.

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.
- Bayarri, M. J., & Berger, J. O. (2000). P-values for composite null models. *Journal of the American Statistical Association*, 95, 1127-1142.
- Belov, D. I. (2015). Robust detection of examinees with aberrant answer changes. *Journal of Educational Measurement*, 52, 437-456.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. New York, NY: Chapman & Hall.
- Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, 27, 217-233.
- Hanson, B. A., Harris, D. J., & Brennan, R. L. (1994). *A comparison of several statistical methods for examining allegations of copying* (ACT Research Rep. Series no. 87-15). Iowa City, IA: American College Testing.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equating”. *Applied Psychological Measurement*, 8, 453-461.
- Maynes, D. (2013). Educator cheating and the statistical detection of group-based test security threats. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 173-199). New York, NY: Routledge.
- Mislevy, R. J., & Bock, R. D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement*, 42, 725-737.
- Primoli, V., Liassou, D., Bishop, N. S., & Nhouyvanisvong, A. (2011, April). *Erasure descriptive statistics and covariates*. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.
- Qin, S., & Cohen, A. S. (2013, July). *An empirical analysis for detection of possible test tampering*. Paper presented at the annual meeting of the Psychometric Society, Arnhem, The Netherlands.
- 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: Wiley.

- Sinharay, S. (2015). Assessment of person fit for mixed-format tests. *Journal of Educational and Behavioral Statistics*, 40, 343-365.
- Sinharay, S. (2016). Bayesian model fit and model comparison. In W. van der Linden (Ed.), *Handbook of item response theory: Vol. 2. Statistical tools*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Sinharay, S., & Duong, M. (2016). *Detection of aberrant answer changes: An alternative to the indices based on Kullback-Leibler divergence*. Manuscript submitted for publication.
- Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30, 298-321.
- 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.
- Tiemann, G. C., & Kingston, N. M. (2014). An exploration of answer changing behavior on a computer-based high-stakes achievement test. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 158-171). New York, NY: Routledge.
- Toribio, S. G., & Albert, J. H. (2011). Discrepancy measures for item fit analysis in item response theory. *Journal of Statistical Computation and Simulation*, 81, 1345-1360.
- van der Linden, W. J., & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics*, 37, 180-199.
- van der Linden, W. J., Jeon, M., & Ferrara, S. (2011). A paradox in the study of the benefits of test-item review. *Journal of Educational Measurement*, 48, 380-398.
- van der Linden, W. J., & Lewis, C. (2015). Bayesian checks on cheating on tests. *Psychometrika*, 80, 689-706.
- van der Linden, W. J., & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, 31, 283-304.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427-450.
- Wollack, J., & Eckerly, C. (in press). Detecting test tampering at the group level. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests*. Washington, DC: Routledge.
- Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, 21, 307-320.
- Wollack, J. A., Cohen, A. S., & Eckerly, C. A. (2015). Detecting test tampering using item response theory. *Educational and Psychological Measurement*, 75, 931-953.
- Wollack, J. A., Cohen, A. S., & Serlin, R. C. (2001). Defining error rates and power for detecting answer copying. *Applied Psychological Measurement*, 25, 385-404.
- Zopluoglu, C., & Davenport, E. C. (2012). The empirical power and Type I error rates of the GBT and ω indices in detecting answer copying on multiple-choice tests. *Educational and Psychological Measurement*, 72, 975-1000.