Detecting Fraudulent Erasures At An Aggregate Level

Sandip Sinharay, Educational Testing Service

An Updated Version of this document appeared in the Journal of Educational and Behavioral Statistics. The website for the article is http://journals.sagepub.com/doi/abs/10.3102/1076998617739626

The citation for the article is: Sinharay, S. (2017). Detecting fraudulent erasures at an aggregate level. Journal of Educational and Behavioral Statistics. Advance Online Publication. doi: 10.3102/1076998617739626.

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

Detecting Fraudulent Erasures At An Aggregate Level

Abstract

Wollack, Cohen, and Eckerly (2015) suggested the "erasure detection index" (EDI) to detect fradulent erasures for individual examinees. Wollack and Eckerly (2017) extended the EDI to detect fradulent erasures at the group level. The EDI at the group level was found to be slightly conservative. This paper suggests two modifications of the EDI for the group level. The asymptotic null distribution of the two modified indices is proved to be the standard normal distribution. In a simulation study, the modified indices are shown to have Type I error rates close to the nominal level and larger power than the index of Wollack and Eckerly (2017). A real data example is also included.

Key words: data forensics; erasure analysis; test fraud; test security

Acknowledgements

The author would like to thank the editor Li Cai and the two anonymous reviewers for several helpful comments that led to a significant improvement of the paper. The author would also like to thank Shelby Haberman, Dan McCaffrey, and J. R. Lockwood for their helpful comments. Finally, the author is grateful to James Wollack for his useful comments and for sharing a data set that was used in this paper. 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.

There is a growing interest in *erasure analysis*, which comprises analyses of erasure patterns in an attempt to detect *test tampering* that leads to *fraudulent* or *aberrant* erasures. Standard 8.11 of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014) include the recommendation that testing programs may use technologies such as computer analyses of erasure patterns in the answer-sheets to detect possible irregularities.

Erasures on paper-and-pencil tests have received most attention. However, erasures essentially mean answer changes (ACs) and computer-based tests (CBTs) may also suffer from fraudulent ACs. Tiemann and Kingston (2014) and Sinharay, Duong, and Wood (2017) provided examples of CBTs in which ACs are allowed—fraudulent ACs can definitely occur for such tests.

Wollack et al. (2015) suggested the *erasure detection index* (EDI) to detect fraudulent erasures for individual examinees. The EDI is based on item response theory (IRT). Wollack and Eckerly (2017) extended the EDI to detect fraudulent erasures at the group level, where a group could be a class, school, or district that the examinees belong to. The EDI at the group level would be denoted as EDI_g henceforth. Note that the groups in applications of the group-level EDI are known in advance, that is, the groups do not have to be identified using a statistical/psychometric method.

A continuity correction is used with EDI_g. Wollack and Eckerly (2017) found EDI_g to be slightly conservative and attributed the conservativeness to the continuity correction. The purpose of this paper is to demonstrate, first using theory of large-sample statistical inference and then using a simulation, that this continuity correction is not required and it unnecessarily reduces the power of EDI_g. It is demonstrated that two modified versions of EDI_g that involve no continuity correction have Type I error rates closer to the nominal level and larger power compared to EDI_g.

The next section includes some background material including a review of the EDI at the individual level (Wollack et al., 2015) and at the group level (EDI_g; Wollack & Eckerly, 2017). The modified versions of EDI_g are discussed in the Methods section. In

the Simulation Study section, the Type I error rates and power of the modified versions of EDI_g are compared with those of EDI_g. Conclusions and recommendations are provided in the last section.

As in Wollack et al. (2015) and Wollack and Eckerly (2017), this paper focuses only on dichotomous items and involves the assumption that the item parameters are known. Note that to apply any of the analysis discussed in this paper, the investigator has to know, for each examinee, the items on which he/she produced an erasure. As discussed in Cizek and Wollack (2017, p. 15), several states use scanners to collect such information on erasures.

Background

Erasure Analysis in Practice and Research

Erasure analysis was brought into prominence during the widespread allegation of educator cheating in Atlanta public schools on the Georgia Criterion-referenced competency tests in 2009 (e.g., Kingston, 2013; Maynes, 2013; Wollack et al., 2015). A special investigation by the state of Georgia identified 178 educators within Atlanta public schools as being involved in cheating (e.g., Maynes, 2013, p. 173). Since then, erasure analysis has been performed in several state tests. A survey conducted by USA Today in September 2011 of State Education Agencies found that 20 states and Washington, DC, conducted some type of erasure analysis (e.g., McClintock, 2015). In a report for the Council of Chief State School Officers, Fremer and Olson (2015) mentioned that erasure analysis and analysis of gain scores are used more often to investigate testing irregularities than other types of analyses because they are "so readily performed and because they have proven their value in practice".

The average wrong-to-right (WTR) erasure count is operationally used in several states to detect fraudulent erasure at the school level or class level (e.g., Bishop & Egan, 2017; McClintock, 2015; Wollack & Eckerly, 2017). Typically, the average ($\bar{x}$) and SD (s_x) of WTR count is computed over all the examinees (for example, of a state) who took the test; then, as described in, for example, Bishop and Egan (2017, pp. 204-205), one flags a group (e.g., a class or a school) with n_g examinees if the average WTR count for the group

is outside a confidence bound $(\bar{x} - Qs_x/\sqrt{n_g}, \bar{x} + Qs_x/\sqrt{n_g})$, where Q is an appropriate quantile of the standard normal distribution. The basis of this flagging is the assumption that

$$\text{WTR}_{std} = \frac{\bar{x}_g - \bar{x}}{s_x/\sqrt{n_g}}, \tag{1}$$

which is a standardized version of the average WTR count for the group, follows a standard normal distribution under the null hypothesis of no fraudulent erasures, where $\bar{x}_g$ is the average WTR count for the group. The performance of WTR_{std} will be examined later in this paper.

To address the increasing interest in practice on erasure analysis, there has been an upswing in research on the topic. Recently, researchers such as Belov (2015), Sinharay et al. (2017), Sinharay and Johnson (2017), van der Linden and Jeon (2012), van der Linden and Lewis (2015), Wollack et al. (2015), and Wollack and Eckerly (2017) presented new statistics for individual-level or group-level erasure analysis. Sinharay et al. (2017) performed a comprehensive comparison of several of these statistics at the individual level—they found the EDI (Wollack et al., 2015) and their suggested statistic L-index, which is based on the likelihood ratio statistic, to have performed the best.

Erasure Detection Index at the Individual Level

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 an item response theory (IRT) model. Let us consider examinee j; let E_j denote the set of items on which erasures were found for the examinee. Note that the erasures could have been produced by the examinee and/or an educator and some erasures could be *benign*, that is, not fraudulent. Let N_j denote the number of items in E_j. Let $\overline{E}_j$ denote the set of items on which no erasures were found for examinee j. [1] Let X_j denote the raw score of the examinee on the items in E_j. Note that X_j is also the number of WTR erasures [2] and

[1] E_j and $\overline{E}_j$ are non-overlapping and their union is the set of all items administered to the examinee.

[2] This is because a right-to-right erasure is impossible for regular dichotomously-scored multiple-choice items that involve only one correct answer option.

is often referred to as the *WTR score*. Let μ_j and σ_j respectively denote the expected value and standard deviation (SD) of X_j given the true ability parameter (θ_j) of the examinee. For example, μ_j can be computed as the sum of the probabilities of correct answers on the items in E_j.

Wollack et al. (2015) and Wollack and Eckerly (2017) used in their erasure analysis the nominal response model (NRM; Bock, 1972) under which $p_{ik}(\theta_j)$, the probability that an examinee of ability θ_j chooses the response option k on item i, is given by

$$p_{ik}(\theta_j) = \frac{\exp[\zeta_{ik} + \lambda_{ik}\theta_j]}{\sum_m \exp[\zeta_{im} + \lambda_{im}\theta_j]},$$

where ζ_{im} and λ_{im} respectively are the intercept and slope parameters for response option m of item i.

Let $P_i(\theta_j)$ denote the probability of a correct answer on item i by examinee j whose ability is equal to θ_j. For the NRM, $P_i(\theta_j) = p_{ik_i}(\theta_j)$, where the alternative k_i represents the correct answer option for item i. One then obtains

$$\mu_j = \sum_{i \in E_j} P_i(\theta_j) \text{ and } \sigma_j = \sqrt{\sum_{i \in E_j} P_i(\theta_j)[1 - P_i(\theta_j)]}. \tag{2}$$

The ability θ_j is unknown for real data. Wollack et al. (2015) recommended estimating θ_j from the responses on the items in $\overline{E}_j$. Let us denote this estimate as $\hat{\theta}_j$. The estimate $\hat{\theta}_j$ is robust to potentially aberrant erasures, and, because $\overline{E}_j$ is usually a large part of the whole test, typically has a small standard error and hence can be considered close to θ_j.

The estimated mean and SD, denoted respectively by $\hat{\mu}_j$ and $\hat{\sigma}_j$, are obtained by replacing θ_j by $\hat{\theta}_j$ in Equation 2.

The EDI at the examinee level is then defined as

$$\text{EDI} = \frac{X_j - \hat{\mu}_j + c}{\hat{\sigma}_j}. \tag{3}$$

The quantity c, which represents a continuity correction, was assumed to be equal to -0.5 by Wollack et al. (2015) who assumed that the EDI approximately follows the standard normal distribution under the null hypothesis of no fraudulent erasures. The null hypothesis is rejected and an examinee is flagged for potentially fraudulent erasures 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 0.01 is used.

The Extension of the Erasure Detection Index to the Group Level

Consider a group of examinees, where a group could refer to a class, school, or district. Suppose that at least one erasure was found for J examinees in the group. Wollack and Eckerly (2017) defined the group-level EDI, or, EDI_g, as

$$\text{EDI}_g = \frac{\sum_{j=1}^{J}(X_j - \hat{\mu}_j) - 0.5}{\sqrt{\sum_{j=1}^{J}\hat{\sigma_j^2}}}. \tag{4}$$

Because each statistic is defined for one examinee group at a time in this paper, no subscript for the group is used in the notations. The subtraction of 0.5 in the numerator of the right-hand side of the above equation denotes a continuity correction of -0.5 for EDI_g. Wollack and Eckerly (2017) commented that the continuity correction is small at the group level because it represents a small fraction of the expected number of erasures and its impact on power should be minimal (p. 219). Wollack and Eckerly (2017) also noted that EDI_g essentially treats the entire group of examinees as if it were a single student taking a very long test and computes the index over all erasures in the group.

Wollack and Eckerly (2017) assumed that EDI_g appoximately follows the standard normal distribution under the null hypothesis of no fraudulent erasures. The null hypothesis is rejected and the examinee group is flagged for potentially fraudulent erasures if the group's EDI_g is a large positive number.

Wollack and Eckerly (2017) found, in a detailed simulation study, that EDI_g, either at the class or school level, was slightly conservative, that is, its Type I error rate was slightly smaller than the nominal level. For example, in their Table 11.2, the Type I error rate of EDI_g for classes, aggregated over all of their simulation conditions was 0.005 at level 0.01 and 0.029 at level 0.05. Wollack and Eckerly (2017) noted that a possible reason of this conservativeness is the continuity correction. However, they did not provide any results on the Type I error rate or the power of EDI_g without a continuity correction.

Methods: Two Modified Versions of EDI_g

The Continuity Correction Involved in EDI_g

The continuity correction involved in the EDI (at the examinee level) given by Equation 3 was introduced to reduce the Type I error rate of the index; without the correction, the EDI often led to inflated Type I error rates, especially when E_j includes only a few items. Sinharay and Johnson (2017) showed in a simulation study that the null distribution of the EDI without a continuity correction is quite different from the standard normal distribution when N_j is 5 or smaller, but is close to the standard normal distribution when N_j is larger than 5. Primoli, Liassou, Bishop, and Nhouyvanisvong (2011) found that erasures are found on 2% items per examinee on average; therefore, on average, the number of erasures per examinee is 2 on a 50-item test. So, the EDI without a continuity correction at the examinee level will often not follow a standard normal distribution and be larger on average than a standard normal random variable and the continuity correction suggested by Wollack et al. (2015) is one way to control the Type I error rate of the EDI.

Further, a continuity correction is often used when the distribution of a test statistic consisting of discrete observations is approximated by a continuous random variable. Yates' continuity correction (Yates, 1934) of the Pearson's χ^2 statistic, in which 0.5 is subtracted from the absolute difference of the observed and expected frequency in the numerator, is a prime example of a continuity correction.

However, the normality assumption is more likely to be satisfied for EDI_g without any continuity correction than for the EDI without a continuity correction at the individual level. Given the erasure rate of 2% items per examinee (e.g., Primoli et al., 2011), the number of erasures on a test by an examinee roughly follows a binomial distribution with N=the number of items and success probability=0.02 (e.g., Wollack et al., 2015). Then, the probability of finding at least one erasure for any given examinee on a 50-item test is 0.64, which means that the expected number of examinees with at least one erasure on such a test is about 13 in a class with 20 examinees. Further, for such a class and a test, 20 erasures would be found on average and the chance of finding more than a total

of 5 erasures is 1.00 up to 2 decimal places. Then, in practice, EDI_g without a continuity correction would most often follow a standard normal distribution given that EDI_g treats the entire group of examinees as if it were a single student taking a very long test and computes the index over all erasures in the group (Wollack & Eckerly, 2017) and the null distribution of the EDI without a continuity correction is very close to the standard normal distribution when the number of erasures is more than 5 (Sinharay & Johnson, 2017). Also note that several researchers (e.g., Furr, 2010) noted that the Yates' correction leads to conservative tests and is not needed except for very small sample sizes.

The First Modified Version and its Asymptotic Null Distribution

A modified group-level EDI, or, EDI_g^N, is defined as

$$\text{EDI}_g^N = \frac{\sum_{j=1}^J (X_j - \hat{\mu}_j)}{\sqrt{\sum_{j=1}^J \hat{\sigma_j^2}}}. \tag{5}$$

The EDI_g^N is similar to EDI_g with the only difference that the former does not involve a continuity correction. The superscript N in the symbol EDI_g^N denotes "No" continuity correction.

Under the null hypothesis of no fraudulent erasures,

- $\frac{\sum_{j=1}^J (X_j - \mu_j)}{\sqrt{\sum_{j=1}^J \sigma_j^2}} \xrightarrow{d} \mathcal{N}(0,1)$, where the symbol $\xrightarrow{d}$ denotes "converges in distribution" and $\mathcal{N}(0,1)$ denotes the standard normal distribution, by the central limit theorem (CLT; e.g., Rao, 1973, pp. 127-128).

- As $\hat{\theta}_j \to \theta_j$ (e.g., Chang & Stout, 1993), $\sum_{j=1}^J \hat{\mu}_j \to \sum_{j=1}^J \mu_j$ and $\sum_{j=1}^J \hat{\sigma}_j^2 \to \sum_{j=1}^J \sigma_j^2$.

- $\frac{\sum_{j=1}^J (X_j - \mu_j)}{\sqrt{\sum_{j=1}^J \hat{\sigma}_j^2}} = \frac{\sum_{j=1}^J (X_j - \mu_j)}{\sqrt{\sum_{j=1}^J \sigma_j^2}} \times \frac{\sqrt{\sum_{j=1}^J \sigma_j^2}}{\sqrt{\sum_{j=1}^J \hat{\sigma}_j^2}} \xrightarrow{d} \mathcal{N}(0,1)$ by the Slutsky's theorem (e.g., Casella & Berger, 2002, pp. 239-240) and the standard normality of $\frac{\sum_{j=1}^J (X_j - \mu_j)}{\sqrt{\sum_{j=1}^J \sigma_j^2}}$.

- $$\frac{\sum_{j=1}^J (X_j - \hat{\mu}_j)}{\sqrt{\sum_{j=1}^J \hat{\sigma}_j^2}} = \frac{\sum_{j=1}^J (X_j - \mu_j)}{\sqrt{\sum_{j=1}^J \hat{\sigma}_j^2}} + \frac{\sum_{j=1}^J \mu_j - \sum_{j=1}^J \hat{\mu}_j}{\sqrt{\sum_{j=1}^J \hat{\sigma}_j^2}} \xrightarrow{d} \mathcal{N}(0,1) \tag{6}$$

by the Slutsky's theorem and the standard normality of $\frac{\sum_{j=1}^{J}(X_j-\mu_j)}{\sqrt{\sum_{j=1}^{J}\hat{\sigma}_j^2}}$.

Thus, EDI_g^N has an asymptotic standard normal distribution under the null hypothesis. Further, EDI_g^N, because of no continuity correction, will always be larger than EDI_g.

The Second Modified Version and its Asymptotic Null Distribution

From Equations 2 and 5, EDI_g^N can be expressed as

$$\text{EDI}_g^N = \frac{\sum_{j=1}^{J}(X_j - \sum_{i\in E_j} P_i(\hat{\theta}_j))}{\sqrt{\sum_{j=1}^{J}\sum_{i\in E_j} P_i(\hat{\theta}_j)[1-P_i(\hat{\theta}_j)]}}, \tag{7}$$

where the denominator is supposed to be the estimated standard deviation of the numerator. However, the above formula was obtained by assuming that the examinee abilities are known and then by replacing the abilities by their estimates. However, researchers have found that when the examinee abilities are replaced by their estimates in a statistic, the resulting statistic often does not follow the theorized null distribution. For example, the popular person-fit statistic l_z (Drasgow, Levine, & Williams, 1985), which is obtained by replacing the examinee ability by its estimate in an expression somewhat similar to the right-hand side of Equation 7 (similar in the sense of being the standardized version of another statistic), has been shown to not follow its theorized standard normal null distribution even for long tests. Snijders (2001) and Sinharay (2016) suggested an adjusted statistic l_z^* that has a standard normal null distribution asymptotically. The adjustment of Snijders (2001) and Sinharay (2016) is based on the Taylor-series expansion (e.g., Casella & Berger, 2002, p. 240). A similar Taylor-series expansion is applied here on EDI_g^N in the following derivation.

The variance of the numerator in Equation 7 is equal to the sum of the variances of $[X_j - \sum_{i\in E_j} P_i(\hat{\theta}_j)]$ over j because of the independence of the examinees in a group under the null hypothesis of no fraudulent erasures. Further,

$$\text{Var}\left(X_j - \sum_{i\in E_j} P_i(\hat{\theta}_j)\right) = \text{Var}(X_j) + \text{Var}\left(\sum_{i\in E_j} P_i(\hat{\theta}_j)\right), \tag{8}$$

because, conditional on the examinee abilities, [3] X_j and $\sum_{i \in E_j} P_i(\hat{\theta}_j)$ are independent by the local independence assumption of IRT given that X_j is based on the item-scores on E_j whereas $\hat{\theta}_j$ is based on $\overline{E}_j$. Then,

$$\mathrm{Var}(X_j) = \sum_{i \in E_j} P_i(\theta_j)[1 - P_i(\theta_j)]. \tag{9}$$

Further, by the Taylor-series expansion of the first order (e.g., Casella & Berger, 2002, p. 240),

$$\sum_{i \in E_j} P_i(\hat{\theta}_j) \approx \sum_{i \in E_j} P_i(\theta_j) + (\hat{\theta}_j - \theta_j) \sum_{i \in E_j} P_i'(\theta_j), \tag{10}$$

where $P_i'(\theta_j)$ is the first derivative of $P_i(\theta_j)$ with respect to θ_j. For the NRM (Bock, 1972), $P_i'(\theta_j)$ is equal to $P_i(\theta_j)[\lambda_{ik_i} - \sum_m \lambda_{im} p_{im}(\theta_j)]$ as shown in, e.g., Baker and Kim (2004, p. 252). Taking variances of both sides of Equation 10 and noting that the first term of the right-hand side of Equation 10 is a constant,

$$\mathrm{Var}\left(\sum_{i \in E_j} P_i(\hat{\theta}_j)\right) = \mathrm{Var}(\hat{\theta}_j) \left[\sum_{i \in E_j} P_i'(\theta_j)\right]^2 \cdot \tag{11}$$

The above expression of variance can also be obtained by the delta method (e.g., Casella & Berger, 2002, p. 240-245). Thus, by Equations 8, 9, and 11,

$$\mathrm{Var}\left(\sum_{j=1}^{J}\left[X_j - \sum_{i \in E_j} P_i(\hat{\theta}_j)\right]\right) = \sum_{j=1}^{J}\sum_{i \in E_j} P_i(\theta_j)[1 - P_i(\theta_j)] + \sum_{j=1}^{J} \mathrm{Var}(\hat{\theta}_j) \left[\sum_{i \in E_j} P_i'(\theta_j)\right]^2 \cdot \tag{12}$$

An estimate of the quantities in the right-hand side of the above equation can be obtained by replacing the θ_j by $\hat{\theta}_j$ for all j. Then, using Equation 12, another modified version of EDI_g can be defined as the ratio of $\sum_{j=1}^{J}(X_j - \sum_{i \in E_j} P_i(\hat{\theta}_j))$ and its estimated standard deviation [4], that is, as

$$\mathrm{EDI}_g^A = \frac{\sum_{j=1}^{J}(X_j - \sum_{i \in E_j} P_i(\hat{\theta}_j))}{\sqrt{\sum_{j=1}^{J}\sum_{i \in E_j} P_i(\hat{\theta}_j)[1 - P_i(\hat{\theta}_j)] + \sum_{j=1}^{J} \widehat{\mathrm{Var}}(\hat{\theta}_j) \left[\sum_{i \in E_j} P_i'(\hat{\theta}_j)\right]^2}}, \tag{13}$$

[3] the variances in Equation 8 and elsewhere are conditional on the true ability and item parameters. For convenience, the notations do not reflect the conditioning.

[4] Note here that the asymptotic mean of $\sum_{j=1}^{J}(X_j - \sum_{i \in E_j} P_i(\hat{\theta}_j))$ is 0.

where $\widehat{\mathrm{Var}}(\hat{\theta}_j)$ can be computed as the reciprocal of the estimated information on the ability for student j based on $\overline{E}_j$. If $\hat{\theta}_j$ is computed using the Newton-Raphson algorithm, then $\widehat{\mathrm{Var}}(\hat{\theta}_j)$ can be obtained from the same computer program. [5] The superscript A in the symbol EDI_g^A denotes "adjusted". Thus, EDI_g^A may be considered to be an adjusted version of EDI_g whose denominator has been adjusted to reflect the correct variance of the numerator. The asymptotic null distribution of EDI_g^A is standard normal by the CLT for independent random variables [6] (e.g., Rao, 1973, pp. 127-128) and the Slutsky's theorem (e.g., Casella & Berger, 2002, pp. 239-240). A comparison of Equations 7 and 13 shows that the numerator of EDI_g^N and EDI_g^A are the same, but the denominator of the latter is larger than that of the former by the (non-negative) term $\sum_{j=1}^J \widehat{\mathrm{Var}}(\hat{\theta}_j)\left[\sum_{i\in E_j} P_i'(\hat{\theta}_j)\right]^2$. Thus, EDI_g^A will always be smaller than or equal to EDI_g^N in absolute value . It is difficult to prove such a relationship between EDI_g^A and EDI_g in general. The numerator of EDI_g^A is larger than that of EDI_g by 0.5, but the denominator of EDI_g^A is larger than that of EDI_g by $\sum_{j=1}^J \widehat{\mathrm{Var}}(\hat{\theta}_j)\left[\sum_{i\in E_j} P_i'(\hat{\theta}_j)\right]^2$. It was found in the simulations and the real data example (described later) that EDI_g^A is most often larger than EDI_g. For the schools/districts that have large values of these statistics, however, EDI_g was larger than EDI_g^A; this is somewhat expected; for a school with a large value of EDI_g, the numerator of the right-hand side of Equation 4 is much larger than its denominator and an addition of $\sum_{j=1}^J \widehat{\mathrm{Var}}(\hat{\theta}_j)\left[\sum_{i\in E_j} P_i'(\hat{\theta}_j)\right]^2$ to the denominator, especially for a large school (for which J would be large), will have a comparatively larger effect than the addition of 0.5 to the numerator—that would lead to EDI_g being larger than EDI_g^A.

The Role of Independence

Because the examinee group is known in the computation of EDI_g, EDI_g^N, or EDI_g^A, statistical inference on these indices can be performed conditional on the true abilities of the group of examinees—this conditional inference allows the use of the local independence

[5] Each step of the Newton-Raphson algorithm involves the estimated information at the current ability estimate (e.g., Casella & Berger, 2002, p. 66-67). So, the reciprocal of the estimated information after the algorithm has converged can be used as $\widehat{\mathrm{Var}}(\hat{\theta}_j)$.

[6] where the variables are $X_j - \sum_{i\in E_j} P_i(\hat{\theta}_j), j = 1, 2, \cdots, J$.

assumption of IRT (that implies that conditional on the examinee ability, the scores on two different parts of the test, E_j and $\bar{E}_j$, are independent) in determining the distribution of these indices under the null hypothesis of no fraudulent erasures. For example, as described earlier, local independence leads to the independence of X_j and $\sum_{i \in E_j} P_i(\hat{\theta}_j)$ given the true ability. Further, under the null hypothesis of no fraudulent erasures, the scores (or ability estimates) of the different examinees in a group are independent of each other—this independence allows the denominators of Equations 4, 5, 13 etc. to be a simple sum over the examinees and makes the null distribution of these indices relatively simple.

A Simulation Study

A detailed simulation study, similar to that in Wollack and Eckerly (2017), was performed to compare the Type I error rate and power of EDI_g to those of EDI_g^N and EDI_g^A.

Design of the Simulation

The design of the simulation study was exactly as in Wollack and Eckerly (2017) except that while 1,000 schools were used in Wollack and Eckerly (2017) used, 10,000 schools were used here for each simulation condition to estimate the Type I error rate and power more precisely. [7] A 50-item test and the NRM (Bock, 1972) was used. The following factors were varied:

- the number of classes within a school (1, 3, or 6)
- the number of students within a class (15, 25, or 35)
- the proportion of tampered classes within a school (0, 0.33, 0.67, or 1)
- the number of erasure victims in a class (1, 3, 5, or 10)
- the number of erasures per erasure victim (3, 5, or 10)

The data for a simulation condition were simulated using the following steps:

[7] For example, at level=0.001, the standard error of the Type I error rate for schools is 0.001 if 1000 schools are used in the simulations, but 0.0003 if 10000 schools are used.

- Complete and untampered data for the number of classes and students stipulated by the simulation condition on a 50-item, 5-alternative test were simulated under the NRM (Bock, 1972) using item parameters from the college-level test of English language used in Wollack et al. (2015) and Wollack and Eckerly (2017). Schools were generated to be of different quality by sampling the mean school ability (θ_S) from a normal distribution with mean 0 and standard deviation (SD) of 0.5. Within each school, item-scores were simulated for all examinees. All classes with a school were assumed to be of the same average ability, [8] that is, the ability of students in all classes of a school were simulated from a normal distribution with mean θ_S and SD of 1.

- Benign erasures, which include both misalignment erasures and random erasures, were simulated. Misalignment erasures (or shift errors) occur when an examinee accidentally bubbles in the answer to item i in the space on the answer sheet reserved for item i+1 (or i-1) and continues to mark answers for a string of consecutive items in the wrong fields. The erasure comes about when the examinee finally realizes the mistake, changes the answers to the misaligned items, and marks those same answers again, this time in the correct fields on the answer sheet. Random erasures occur when an examinee either accidentally bubbles in the wrong answer on the answer sheet, identified it immediately, and changes it to the intended answer, or initially answers an item one way, but on reconsideration, changes that answer. Within each school, reflecting what is observed in reality, 2% students were randomly selected as candidates to produce misalignment erasures and the remaining 98% students were candidates to produce random erasures. For each candidate of misalignment erasure, the number of misaligned items was sampled from a binomial distribution with 50 trials and success probability of 0.25. Then, the starting point of the misalignment was determined by randomly selecting an item between Item 1 and 50-k+1 where k is the number of misaligned items. The initial answer was determined by shifting the final answers one

[8] Limited simulations show that making the classes within a school to have different average abilities does not alter the conclusions from the simulation.

spot. If the initial and final answers were different, it was recorded as an erasure. For candidates of random erasures, the number of randomly erased items was sampled from a binomial distribution with 50 trials and success probability of 0.02. The specific items that were erased were selected at random from all items.

- Fraudulent erasures were simulated. Within each school, the specific classes for which tampering occurred and specific items on which tampering occurred was determined randomly. All tampered items were assumed to result in WTR erasures. To generate, for example, 5 fraudulent erasures for an examinee, 5 incorrect answers were randomly chosen among all incorrect answers of the examinee and were changed to correct answers. In the event that a student's raw score was too high to produce the number of WTR erasures stipulated by the simulation condition, the student was given a perfect score.

Computation

The maximum likelihood estimate (MLE) of the examinee ability was used as the ability estimate and the MLE was restricted between -4.5 and 4.5. Note that for each examinee, the MLE was computed from the items without erasures. Because the number of erasures were 3, 5, or 10 and the expected number of benign erasure was 1 for 98% of the examinees, the MLE was computed from between 39 and 46 untampered items for a majority of examinees—so the MLEs can be considered stable, that is, they had small standard errors. The MLEs were computed using the Newton-Raphson algorithm.

Distribution of the Indices Under the Null Hypothesis

The left panel of Figure 1 shows the kernel-density estimates [9] of the distributions of the values of EDI_g, EDI_g^N and EDI_g^A for a random subset of 2,000 classes under the condition of 1 class per school, 15 students per class, and proportion of tampered classes within a school=0; thus, this condition is associated with no fraudulent erasures and hence the distribution of EDI_g^N and EDI_g^A should be close to the standard normal distribution

[9] created using the function "density" in the R software (R Core Team, 2017).

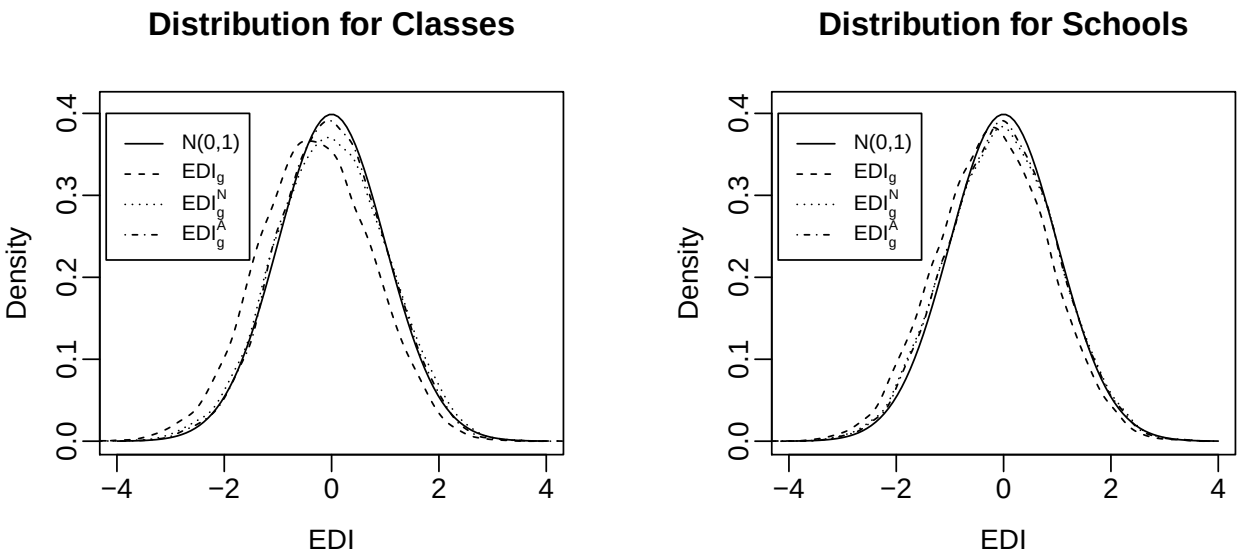

Figure 1. The null distribution of the EDI and the modified versions for classes and schools.

according to the theoretical results included earlier. The standard normal distribution is shown in the figure using a solid line for comparison. Table 1 provides the first four moments (mean, SD, skewness, and kurtosis [10]) and five percentiles (25th, median, 75th, 95th, and 99th) for the distributions shown in the left panel of Figure 1. The right panel of Figure 1 shows a plot for the distributions of the values of EDI_g, EDI_g^N, and EDI_g^A of

[10]Note that 3 has been subtracted from the formula of kurtosis so that the kurtosis of the standard normal distribution is 0 according to the formula used in this paper.

Table 1. Summaries of the Distributions of EDI_g, EDI_g^N and EDI_g^A for the Class Level.

Index	Moments				Percentiles				
	Mean	SD	Skewness	Kurtosis	25	50	75	95	99
$\mathcal{N}(0,1)$	0.00	1.00	0.00	0.00	-0.67	0.00	0.67	1.64	2.33
EDI_g	-0.31	1.05	-0.04	-0.03	-1.03	-0.31	0.41	1.42	2.04
EDI_g^A	-0.01	1.00	-0.04	-0.02	-0.69	-0.02	0.67	1.64	2.21
EDI_g^N	-0.01	1.05	-0.04	-0.03	-0.73	-0.02	0.71	1.72	2.35

2,000 schools for the simulation condition of 3 classes per school, 15 students per class, and proportion of tampered classes within a school=0; thus, this condition is also associated with no fraudulent erasures and hence the distribution of EDI_g^N and EDI_g^A should be close to the standard normal distribution. The distributions of EDI_g^N and EDI_g^A are much closer than that of EDI_g to the standard normal distribution in both panels. The distributions of EDI_g^N and EDI_g^A appear indistinguishable in the right panel (presumably because the values of EDI_g^N and EDI_g^A are based on information from 45 students each); however, in the left panel (where the values of EDI_g^N and EDI_g^A are based on information from only 15 students each), the distribution of EDI_g^A is slightly closer to the standard normal distribution compared to that of EDI_g^N, especially at the right tail of the normal distribution where the rejection decisions are made; further, Table 1 shows that the values for EDI_g^A are closer than those for EDI_g^N to the standard normal distribution for classes.

So, it seems that EDI_g^A follows the standard normal distribution slightly more closely than does EDI_g^N under the null hypothesis, especially for classes. A χ^2 test [11] (e.g., Cochran, 1952) rejected the null hypothesis that EDI_g^N for the classes follows the standard normal distribution, but did not reject the same hypothesis for EDI_g^A.

Results on Type I Error Rates

Table 2 of the current paper, like Table 11.2 of Wollack and Eckerly (2017), shows the Type I error rates of EDI_g, EDI_g^N, and EDI_g^A for classes and schools collapsed over the levels of the different factors from the simulation conditions that did not involve any fraudulent erasures. Four significance (α) levels were considered: 0.0001, 0.001, 0.01, and 0.05. The rates for EDI_g are very close to those in Table 11.2 of Wollack and Eckerly (2017) and support the conclusion of Wollack and Eckerly (2017) that EDI_g is slightly conservative in all conditions. The rates for EDI_g^N are the closest, in comparison to the other two indices, to the nominal level; they are slightly larger than the nominal level in some cases (for example,

[11] where the values of each of the indices were grouped into one of 10 roughly equal-size groups and then the observed and expected numbers in the groups were used to compute a χ^2 statistic whole null distribution is the χ^2 distribution with 9 degrees of freedom.

Table 2. The Overall Type I Error Rates for EDI_g, EDI_g^N and EDI_g^A.

Level of Aggregation	Index	α=0.0001	α=0.001	α=0.01	α=0.05
Class	EDI_g	0.00005	0.0005	0.005	0.031
Class	EDI_g^A	0.00007	0.0008	0.008	0.046
Class	EDI_g^N	0.00012	0.0011	0.011	0.052
School	EDI_g	0.00006	0.0006	0.007	0.035
School	EDI_g^A	0.00010	0.0009	0.009	0.048
School	EDI_g^N	0.00013	0.0011	0.010	0.052

the average Type I error rate for level=0.05 is 0.052), but are satisfactory according to Cochran's criterion for robustness (e.g., Cochran, 1952; Wollack, Cohen, & Serlin, 2001) [12].

The Type I error rates for EDI_g^A, while further from the nominal level compared to EDI_g^N, are closer to the nominal level compared to EDI_g and are always slightly smaller than or equal to the nominal level. Keeping in mind the important consequences of a false alarm in the context of erasure analyses, some practitioners would probably prefer EDI_g^A, whose Type I error rate is smaller than the nominal level (and yet quite close to the nominal level), over EDI_g^N whose Type I error rate is closest to the nominal level but can occasionally be slightly larger than the nominal level.

The only factor (among those manipulated here) that influenced the Type I error rates of the indices for the classes is the class size. This is expected given that the assumption of a standard normal null distribution of EDI_g, EDI_g^A, and EDI_g^N would be satisfied to a greater extent as class size increases. Figure 2 shows the Type I error rates of the indices for different class sizes for significance levels 0.05, 0.01, 0.001, and 0.0001. Each panel, which corresponds to a significance level, shows three dashed lines connecting one among three types of points that denote the Type I error rates for different class sizes. Each point type corresponds to an index. For example, the point type for EDI_g is a hollow circle. In each

[12] According to Cochran's criterion for robustness, estimated Type I error rates smaller than 0.06, 0.015, 0.0015, and 0.00015 are satisfactory at levels 0.05, 0.01, 0.001, and 0.0001, respectively

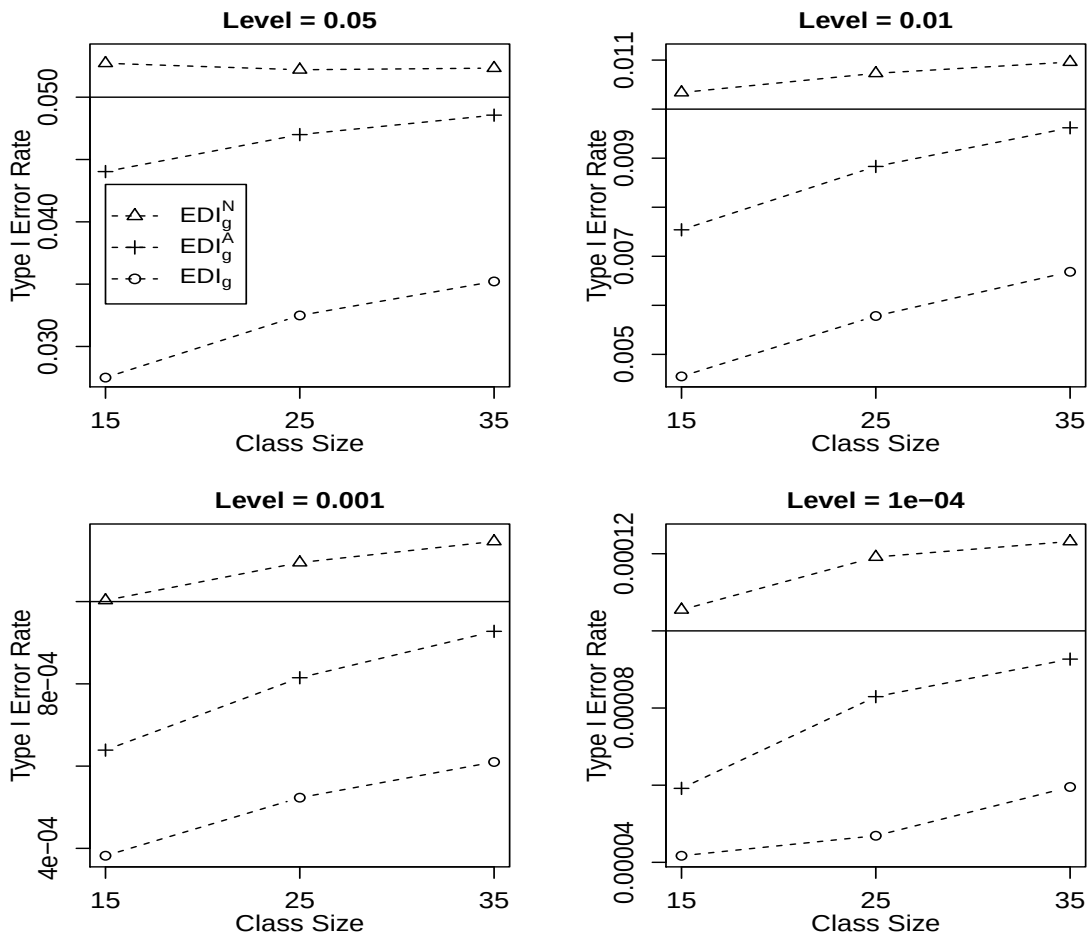

Figure 2. Type I Error Rates of the indices for different class sizes.

panel, the significance level is denoted by a solid horizontal line. The figure shows that as the class size increases, the Type I error rate of each index increases. For EDI_g and EDI_g^A, the increase is desirable because their Type I error rates are smaller than the nominal level. For EDI_g^N, the increase is not desirable because its Type I error rate is slightly inflated. However, even with the increase, the Type I error rate of EDI_g^N is satisfactory according to Cochran's robustness criterion (Cochran, 1952) for the largest class size.

Also note that the increase of the Type I error rate with an increase in the class size in Figure 2 does not mean that the Type I error rate of one or more of these indices will keep increasing or will be severely inflated for much larger groups of examinees (of, say, size 100). The Type I error rates at the school-level, which are shown in Table 2, are computed

using somewhere between 15 and 210 examinees (that is because a school includes 1, 3, or 6 classes with 15, 25, or 35 students each) and they are quite close to the nominal level. To further investigate this issue, some limited simulations were performed with an additional class size of 100. Figure 3 shows the Type I error rates for class sizes between 15 and 100 at levels of 0.05 and 0.01. For class size of 100, the Type I error rates of both EDI_g^N and EDI_g^A are very close to the nominal level while that of EDI_g is closer to the nominal level compared to a class size of 35, but is still considerably smaller than the nominal level.

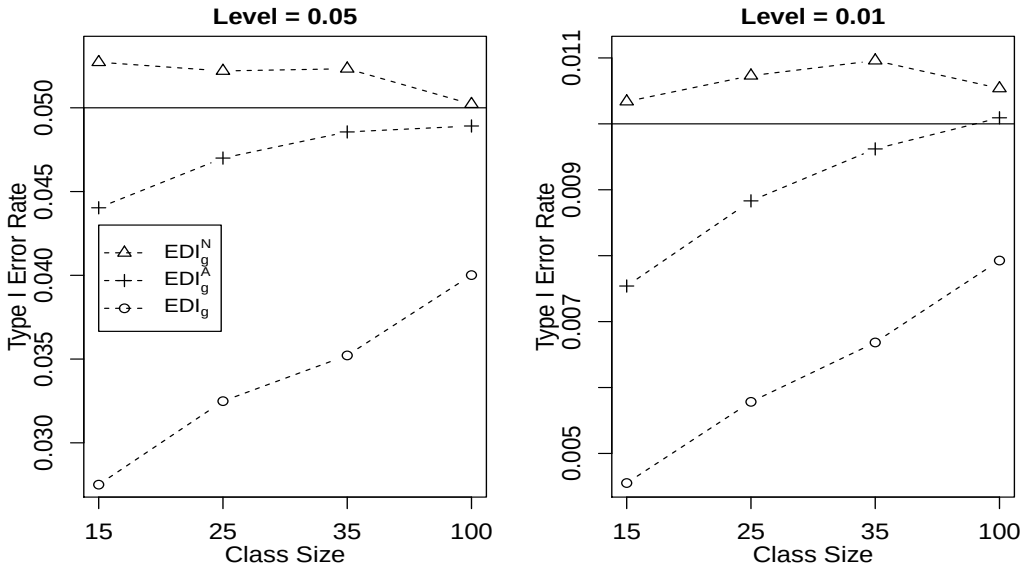

Figure 3. Type I Error Rates of the indices for class sizes of 15 to 100.

Results on Power

Table 3 shows the power (at α levels 0.0001, 0.001, 0.01, and 0.05) of EDI_g, EDI_g^N, and EDI_g^A for classes and schools collapsed over the levels of the different factors from the conditions of the simulation that involved some fraudulent erasures. The values of power for EDI_g are always slightly smaller than those of either of EDI_g^A and EDI_g^N, with the difference being smaller for schools than classes. The values of power of EDI_g^A and EDI_g^N are the same up to two decimal places for schools at three out of the four significance levels. The power of EDI_g^A for classes is either equal to or smaller by 0.01 than that of EDI_g^N up to

Table 3. The Overall Power of the indices.

Level of Aggregation	Index	α=0.0001	α=0.001	α=0.01	α=0.05
Class	EDI_g	0.39	0.47	0.58	0.69
Class	EDI_g^A	0.40	0.49	0.60	0.72
Class	EDI_g^N	0.41	0.50	0.61	0.72
School	EDI_g	0.38	0.44	0.52	0.60
School	EDI_g^A	0.39	0.45	0.53	0.62
School	EDI_g^N	0.39	0.45	0.54	0.62

two decimal places.

Figure 4, whose left and right panels are like Figures 11.1 and 11.2, respectively, of Wollack and Eckerly (2017), shows the power of EDI_g, EDI_g^N, and EDI_g^A for significance level 0.001 to detect classes for different number of erasures (left panel) or different class sizes (right panel) and different number of erasure victims in a class. For each line type, the power of EDI_g^N, EDI_g^A, and EDI_g for different number of erasure victims in a class is shown using a line of that type joining hollow triangles, plus signs, and hollow circles, respectively. Each line type corresponds to a value of the number of erasures per erasure victim (left panel) or class size (right panel). For example, in the left panel, a solid line joining hollow circles denotes the power for EDI_g for 1, 3, 5, or 10 erasure victims per class where each victim made 5 erasures.

Figures 5 and 6, which are like Figures 11.3 and 11.4, respectively, of Wollack and Eckerly (2017), show the power of EDI_g, EDI_g^A and EDI_g^N at significance level of 0.001 to detect schools. In these figures, "#T. Class" in the legends denotes the number of tampered classes. Figure 5 shows power for different number of erasures and Figure 6 shows power for different class sizes.

In Figures 4-6, the power of each index follows patterns that are very similar to those in Wollack and Eckerly (2017); for example, in both Figures 5 and 6, the power of each index increases as the number of tampered classes increases and as the number of erasure victims per class increases. Figures 4-6 also show that EDI_g^A and EDI_g^N are slightly more powerful

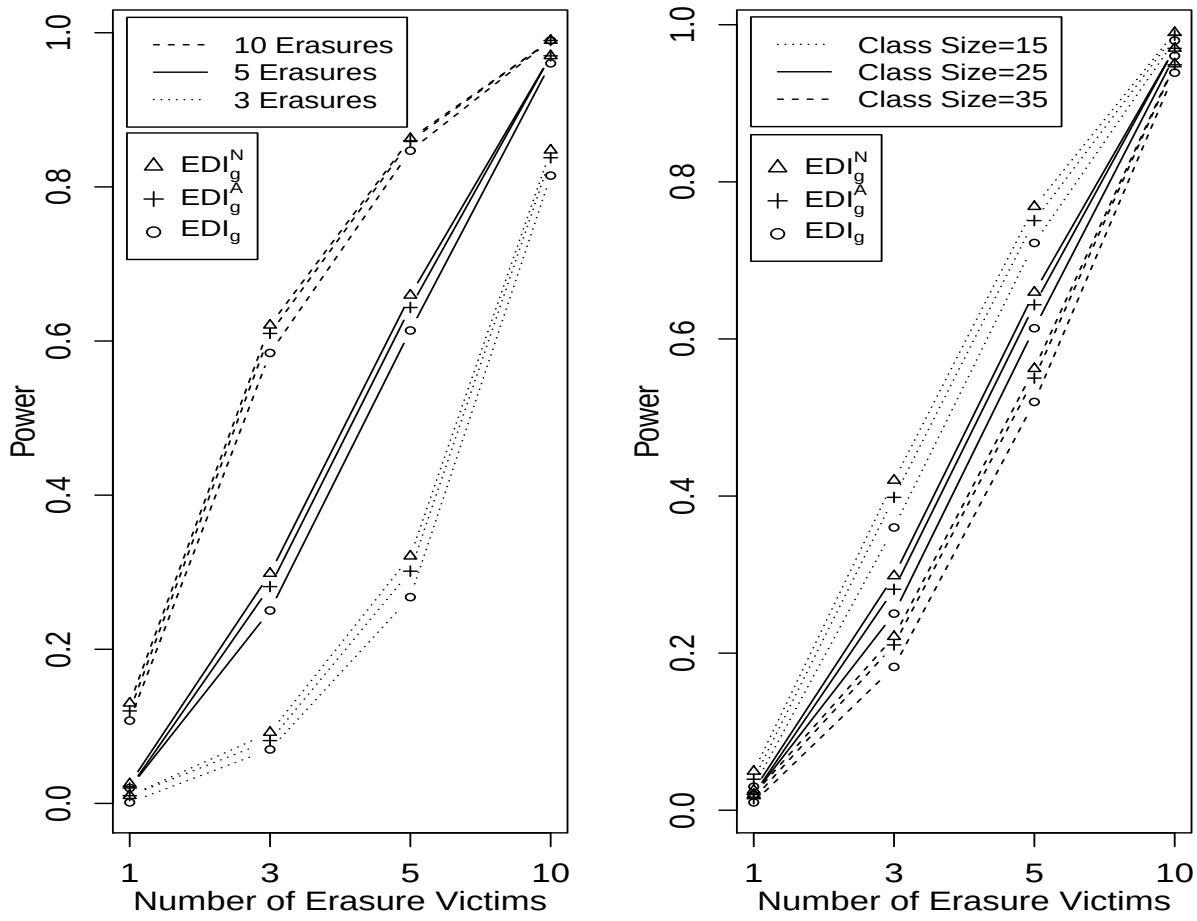

Figure 4. Power of the indices at level=0.001 to detect classes for different number of erasure victims and erasures (left panel) and different number of erasure victims and class size (right panel).

than EDI_g under all simulation conditions for classes and, to a lesser extent, for schools. The gain in power for EDI_g^N over EDI_g in these figures is sometimes up to 0.05 (especially in Figure 4). Thus, even though Wollack and Eckerly (2017) stated that the impact of the continuity correction involved in EDI_g on power should be minimal, these figures show that the impact may not be minimal under some circumstances. Between EDI_g^A and EDI_g^N, the latter has slightly larger or equal power than the former in all simulation cases.

A casual look at Table 3 and Figures 4-6 may provide the impression that the power of the statistics decreases with an increase in sample size; for example,

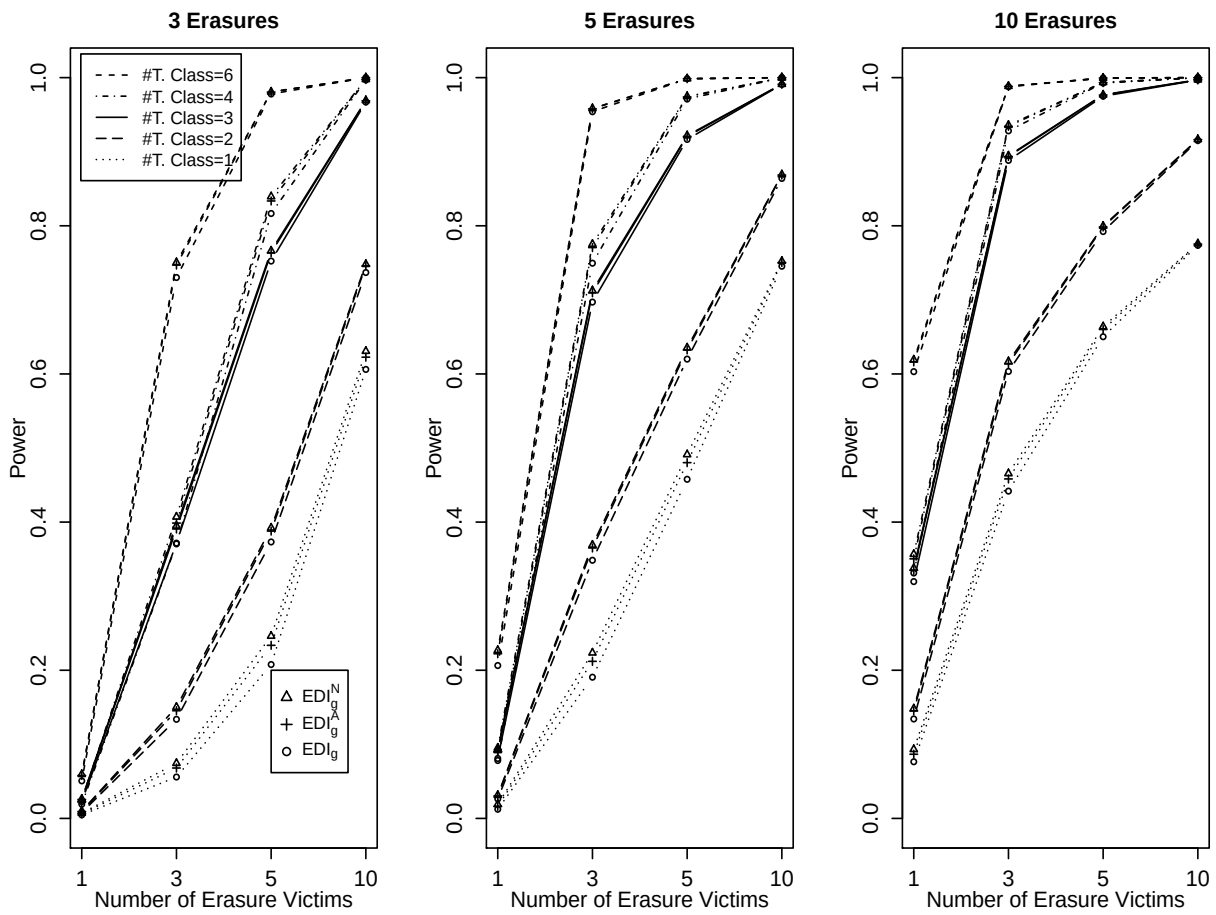

Figure 5. Power of the indices at level=0.001 to detect schools for different numbers of tampered classes, erasure victims and erased items.

- In Table 3, at any significance level, the overall power for schools is smaller than that for classes even though the schools include more students than classes
- In the right panel of Figure 4, the power for class size 15 is larger than that for class size 35.

However, one should be careful about comparing the numbers in Table 3 and Figures 4-6 and a careful comparison shows that Table 3 and Figures 4-6 do not defy the principle of power increasing with sample size (e.g., Rao, 1973, p. 464). For example,

- The smaller overall power for schools than classes in Table 3 is partially explained by

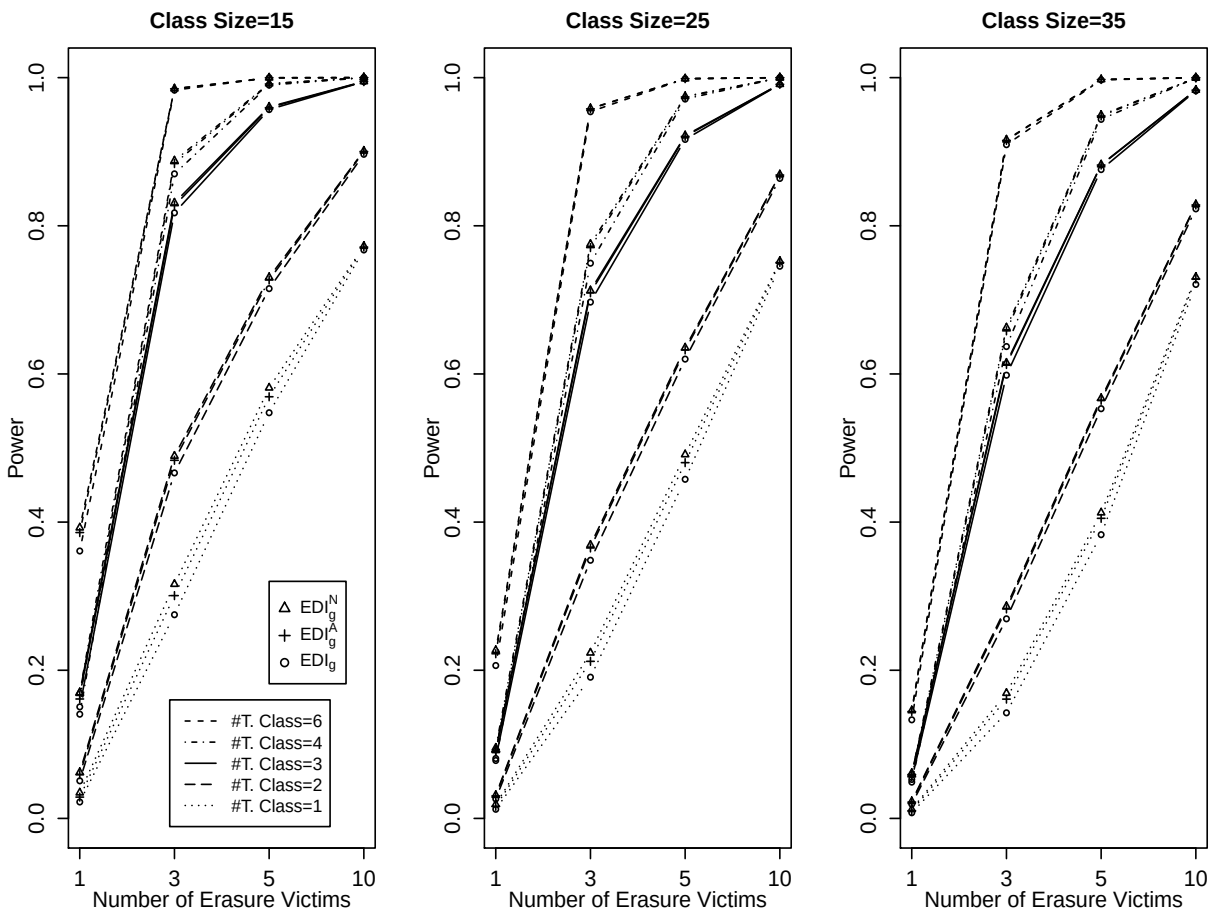

Figure 6. Power of the indices at level=0.001 to detect schools for different numbers of tampered classes, erasure victims and class size.

the fact that the computation of the overall power of schools involved many classes with no fraudulent erasures whereas the overall power of schools was computed only using classes with some fraudulent erasures

- In Figure 4, given a number of erasure victims, the proportion of erasure victims increases as the class size decreases, which causes the increase in power as one goes from, say, class size of 35 to class size of 15, both for 5 erasure victims. If one keeps the proportion of erasure victims constant, however, the power increases with an increase in class size as one would expect; for example, in Figure 4, the power of EDI_g^A is

about 0.40 for class size=15 and number of erasure victims=3, but is about 0.65 for class size=25 and number of erasure victims=5 (the proportion of erasure victims in a class=0.20 in both of these cases).

Figure 7 shows the average power of the three statistics for schools when the proportion of erasure victims in a class=0.20 for two class sizes and three levels of number of erasures. The figure shows that other factors remaining the same, the power of any index increases

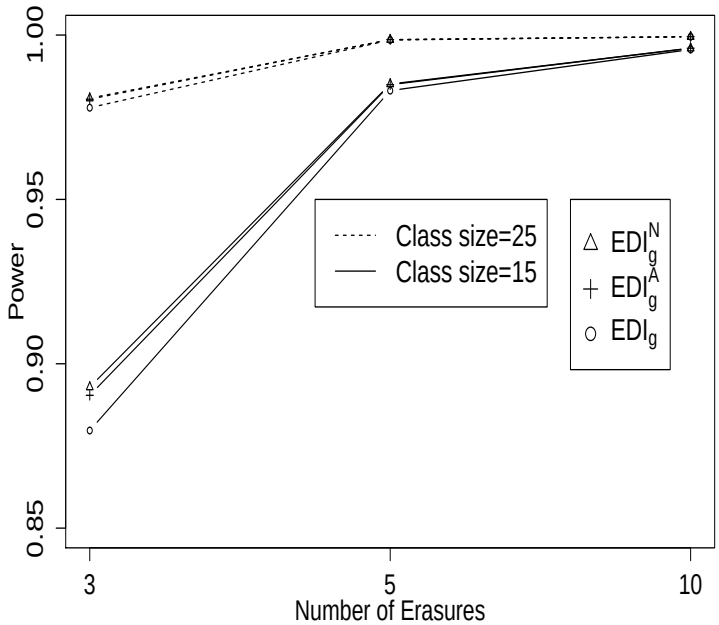

Figure 7. Power of the indices at level=0.001 to detect schools for class sizes 15 and 25 for 20% eraser victims.

with an increase in the class size, which is expected. For example, for 3 erasures per examinee, the power of EDI_g^A is 0.89 for class size of 15, but 0.98 for class size of 25.

Discussion on the Comparative Performance of the Indices

The values of Type I error rates and power from the simulations demonstrate that EDI_g^A has a better balance of Type I error rates and power compared to EDI_g; the Type I error

rates of EDI_g^A are smaller than or equal to the nominal level on average and the power of EDI_g^A is larger than that of EDI_g. Thus, the practitioners should seriously consider applying EDI_g^A to detect fraudulent erasures at group level. The results from the simulations also show that EDI_g^N may be preferred by some practitioners; EDI_g^N is computationally simpler than EDI_g^A (and computationally as easy as EDI_g) and is more powerful than EDI_g^A and EDI_g and the Type I error rates of EDI_g^N are closest to the nominal level on average among these three indices; however, one limitation of EDI_g^N, keeping in mind the severe consequences of a false alarm in the context of erasure analysis, is that its Type I error rate may sometimes be slightly larger than the nominal level.

The Type I error rate and power of the WTR count, which is operationally used in several states, were also examined in the simulation study. Specifically, the WTR_{std} statistic provided by Equation 1 was computed for each class and school. Overall, WTR_{std} did not have satisfactory Type I error rate or power; a part of it can be attributed to the simulation design; for example, in the case when the proportion of tampered classes is 1, the level of tampering is the same in each school [13] and hence the power of WTR_{std} would be close to the Type I error rate. However, even in the simulation conditions most favorable to WTR_{std}, the statistic was less powerful than each of EDI_g, EDI_g^A and EDI_g^N. For example, while the average power of EDI_g, EDI_g^A and EDI_g^N to detect classes were 0.53, 0.56, and 0.55 at level=0.01 for the simulation cases where the proportion of tampered classes is 0.33, the average power of WTR_{std} over the same simulation cases was only 0.29. Wollack and Eckerly (2017) found the correlation between EDI_g and the WTR count and several other similar and popular statistics to be rather small (0.51 or smaller).

Application to Real Data

Data Set and Analyses

Wollack and Eckerly (2017) analyzed a data set that includes the responses of 72,686 fifth-grade students to 53 dichotomous items on a state mathematics test. The students

[13] whereas the WTR count would be powerful only when the level of fraud is very low in most schools and very high in a few schools.

belonged to 3,213 classes in 1,187 schools in 630 districts. The data providers did not reveal if there were any fraudulent erasures on the test. Erasures were captured through a scanning process by looking for "light marks" (Cizek & Wollack, 2017, p. 15). On average, the number of erasures per examinee is two (that is 3.7% of all the items on the test), which is about twice of what is typically found in similar assessments (see, e.g., Primoli et al., 2011; Wollack et al., 2015). About 50.9% of the total number of erasures were WTR erasures. As in Wollack and Eckerly (2017), the NRM was used to analyze these data in this paper and the erased responses were treated as missing data in estimating the item parameters. The missing responses were also treated as missing data in estimating the item and ability parameters. The items had four response categories each; an additional response category ("missing") was assumed in the analysis with the NRM so that five intercept parameters and five slope parameters [14] were estimated for each item using the R package *mirt* (Chalmers, 2012). The statistics EDI_g, EDI_g^A and EDI_g^N were computed for each district, school, and class in the data set.

A significance level of 0.001 was used to determine the statistical significance of the statistics, as in Wollack and Eckerly (2017). A standard normal null distribution assumption for the statistics implies that a value larger than 3.09 of any of this statistics is statistically significant.

Results

The top left and middle panels of Figure 8 show plots of EDI_g vs EDI_g^A and of EDI_g vs EDI_g^N for the districts. The bottom left and middle panels of the figure show similar plots for the schools. In each panel, a diagonal solid line and a horizontal and vertical dotted line at 3.09 (the critical value at level 0.001) are shown. The figure shows that EDI_g^N is always larger than EDI_g, as expected from their definitions, which means that EDI_g^N would flag a larger number of schools/districts compared to EDI_g. The figure also shows that EDI_g^A is mostly larger than EDI_g except for very large values (larger than about 3.3) of these

[14] For each item, the sum of the five intercept parameters is zero and the sum of the five slope parameters is zero.

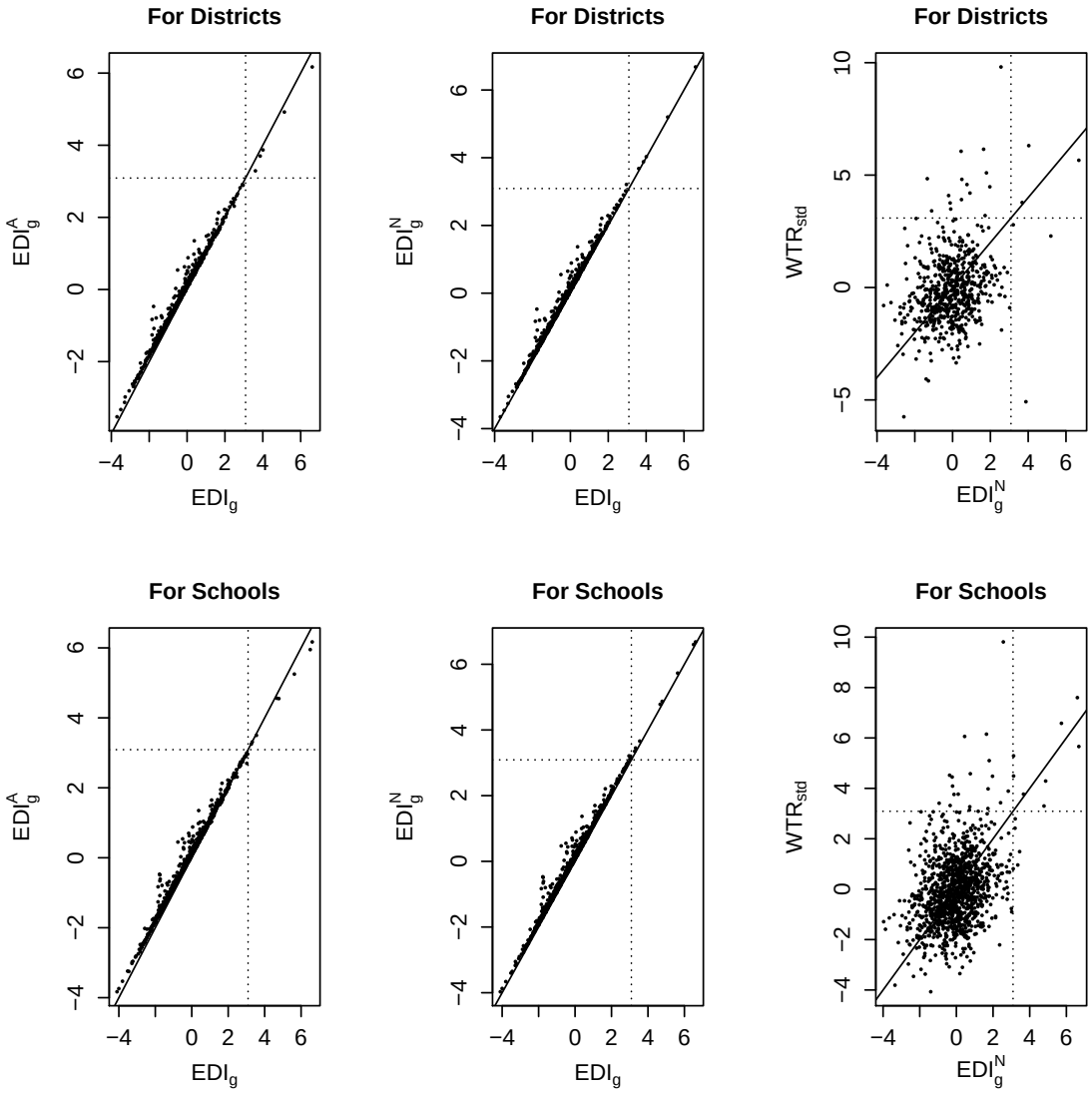

Figure 8. Plots of EDI_g vs EDI_g^A, EDI_g vs EDI_g^N, and EDI_g^N vs WTR_{std} for Districts (Top Row) and Schools (Bottom Row) for the Real Data Set

statistics for which EDI_g^A is mostly smaller than EDI_g.

Among the 630 districts in the data set, EDI_g, EDI_g^A and EDI_g^N were statistically significant for 5, 5, and 6 districts, respectively. Among the 1,187 schools in the data set, EDI_g, EDI_g^A and EDI_g^N were statistically significant for 8, 8, and 13 schools, respectively. Among the 3,213 classes in the data set, EDI_g, EDI_g^A and EDI_g^N were statistically significant

Table 4. Districts, Schools, and Classes for which EDI_g, EDI_g^N, or EDI_g^A was Statistically Significant.

District				School				Class			
ID	EDI_g	EDI_g^N	EDI_g^A	ID	EDI_g	EDI_g^N	EDI_g^A	ID	EDI_g	EDI_g^N	EDI_g^A
401600	6.61	6.68	6.17	344969	6.61	6.68	6.17	9	7.72	7.82	6.96
274475	5.14	5.20	4.92	273425	6.50	6.60	5.95				
71771	4.01	4.03	3.87	244544	3.32	3.45	3.31	5010	3.30	3.47	3.32
				274152	3.06	**3.12**	2.96				
55558	***3.86***	3.88	3.70	354770	4.78	4.87	4.55	331	4.89	4.99	4.59
13758	3.61	3.68	3.29	65825	5.63	5.73	5.25				
424557	2.96	**3.21**	2.89	165894	2.96	**3.21**	2.89				
102235	NS	NS	NS	391665	4.67	4.78	4.56				
88033	NS	NS	NS	187462	3.55	3.66	3.50				
123845	NS	NS	NS	335982	***3.27***	3.35	3.25				
24093	NS	NS	NS	125561	3.05	**3.20**	2.97				
182640	NS	NS	NS	241507	3.02	**3.12**	2.70	3667	4.09	4.24	3.47
								3666	3.24	3.49	2.75
350388	NS	NS	NS	15517	3.01	**3.10**	2.99	5108	3.23	3.40	3.26
374941	NS	NS	NS	388551	NS	NS	NS	102	3.29	3.44	3.28
38891	NS	NS	NS	216471	NS	NS	NS	1000	3.64	3.75	3.62
243971	NS	NS	NS	12900	NS	NS	NS	8	***3.26***	3.40	3.03
190527	NS	NS	NS	351598	NS	NS	NS	441	***3.15***	3.31	3.19
362963	NS	NS	NS	367962	NS	NS	NS	512	3.04	**3.23**	**3.17**
305498	NS	NS	NS	204528	NS	NS	NS	444	2.80	**3.21**	**3.17**

Note: The values of EDI_g written in bold and italicized font were not significant in Wollack and Eckerly (2017), but are significant here. The values of EDI_g^N or EDI_g^A written in bold and regular font correspond to cases for which EDI_g is not significant, but EDI_g^N or EDI_g^A is significant. "NS" means not statistically significant.

for 10, 11, and 12 classes, respectively. Table 4 shows the districts, schools, or classes for which at least one of EDI_g, EDI_g^A and EDI_g^N was statistically significant. For any school, the district that the school belongs to is shown in the same row of the table. For any class, the district and the school that the class belongs to is shown in the same row of the table. For example, the first row includes Class 9, which is within School 344969, which is within District 401600. All districts with significant values of the statistics included at least a school with significant values of the statistics.

The number of significant values for EDI_g is one more here for both schools and districts and two more here for classes compared to that in Wollack and Eckerly (2017) who found four districts, seven schools, and eight classes to have statistically significant values of EDI_g.[15] This difference is most likely an outcome of the way missing data were handled in these two studies and because of the use of different software packages (MULTILOG by Wollack & Eckerly, 2017, vs R in this study) in these two studies. However, the values of EDI_g from our calculations were very close to those of Wollack and Eckerly (2017) for the classes, schools and districts that are listed in Table 11.7 of Wollack and Eckerly (2017). For example, while Wollack and Eckerly (2017) reported the value of EDI_g of District 401600 to be 6.54 in their Table 11.7, the corresponding value here is 6.61. Further, all the EDI_g values (for either classes, schools or districts) that were statistically significant in Wollack and Eckerly (2017) were significant here as well. [16] Some of the values of EDI_g that are significant here and not in Wollack and Eckerly (2017) are justified; for example, Wollack and Eckerly (2017) found EDI_g to be significant (and very large) for School 354770, but did not find EDI_g to be significant for District 55558 that the school belonged to; in contrast, EDI_g was significant for the district in this paper.

There are one district, five schools, and two classes for which EDI_g was not statistically significant, but EDI_g^N was significant; there were two classes for which EDI_g was not statistically significant, but EDI_g^A was significant. [17] Especially, note that School 165894 belongs to District 424557, and EDI_g^N was found significant and EDI_g was found not significant for both of them. Some of the instances with significant EDI_g^N and non-significant EDI_g provide strong evidence in favor of the use of EDI_g^N. For example, EDI_g is significant for both Classes 3667 and 3666; however, EDI_g is not significant for School 241507 that these two classes belong two; in contrast, EDI_g^N is significant for School 241507. Thus, the larger power of EDI_g^N and EDI_g^A (observed in the simulation studies earlier) may manifest

[15] EDI_g was significant here but not in Wollack and Eckerly (2017) for District 55558, School 335982, and Classes 8 and 441.

[16] It was confirmed with James Wollack that in the third row and fourth column of Table 11.7 of Wollack and Eckerly (2017), 13758 should be replaced by 65825.

[17] EDI_g was statistically significant, but EDI_g^A was not significant for Class 8 in School 12900.

itself as a practically different result for certain groups of examinees for real data.

The statistic WTR_{std} provided by Equation 1 was also computed for each district and school. The correlation between EDI_g and WTR_{std} was found to be 0.30 for districts and 0.39 for schools. The two rightmost panels of Figure 8 show plots of EDI_g^N versus WTR_{std} for districts and schools, respectively. The value of WTR_{std} was significant at the level of 0.001 for 18 districts and 29 schools, that is, for a much larger number of districts and schools compared to the other statistics. Among the six districts for which EDI_g^N was significant, WTR_{std} was significant for four, but was 2.79 and -5.08 (and hence not significant) for the remaining two. For one district (with more than 300 students), EDI_g^N was -1.3 (that is, far from being statistically significant), but WTR_{std} was 4.8, which is significant and indicates that statistics such as EDI_g^N can be small even for groups that produce a large number of WTR changes on average. For another district, EDI_g^N was 3.88 (that is, statistically significant), but WTR_{std} was -5.08, which is not significant and indicates that statistics such as EDI_g^N can be significant even for groups that produce fewer number of WTR changes on average.

Erasure analysis was also performed at the individual level using the L-index (Sinharay et al., 2017). The values of the L-index agree with the values of EDI_g, EDI_g^N and EDI_g^A for the data set. For example, the L-index was significant at the level of 0.01 for 13, 12, 15, 14, and 10 percent of examinees, respectively, in the schools 344969, 273425, 65825, 354770, and 391665 that had the largest values of EDI_g^N in Table 4.

Conclusions

This paper follows up on the research of Wollack and Eckerly (2017) by suggesting two modifications of their index for detection of fraudulent erasures at the group level. The suggested modifications have slightly larger power compared to the index of Wollack and Eckerly (2017). The Type I error rate of one of the modified indices is smaller than or equal to the nominal level while that of the other modified index is close to the nominal level, but can occasionally be slightly larger than the nominal level. The choice of an index in a particular application would depend on the preference of the testing program. If

one is willing to allow a couple more false alarms in exchange for a slightly larger power, EDI_g^N would be a better choice. If controlling of the false alarms is the top priority, then EDI_g^A would be a better choice. Note that the computational complexity of the indices is about the same; EDI_g^A involves the derivatives of the response probabilities and estimated variances of the ability estimates, but, remembering that all these indices require the fitting of an IRT model, it is natural to assume that an investigator who has the capability of fitting IRT models should be able to compute derivatives of the response probabilities and estimated variances of the ability estimates.[18] Each of EDI_g, EDI_g^N, or EDI_g^A was modestly correlated with and much more powerful than standardized average WTR count, which is operationally used in erasure analysis by several states.

The choice of the significance level to be used with EDI_g, EDI_g^N, or EDI_g^A is an important issue. Wollack and Eckerly (2017, p. 227) used the significance level of 0.001 in their real data example to limit the number of false positives and commented that states or test sponsors would apply a more conservative criterion in practice. Another option to limit the number of false positives is to choose a critical value that adjusts for multiple comparisons by controlling the family-wise error rate (using, for example, a Bonferroni correction) or controlling the false discovery rate (using the procedure of Benjamini & Hochberg, 1995). If one applies the Bonferroni correction to the real data example discussed above, then critical values of 4.16, 4.30, and 4.52, respectively, would allow one to control the familywise Type I error rate at 0.01 for districts, schools, and classes, respectively (EDI_g^N is significant for two districts, five schools, and two classes if one applies this Bonferroni correction).

Though EDI_g and the suggested modifications were applied in the context of erasure analysis, it is possible to apply them to problems in which (i) examinees belong to groups (like districts, schools, or classes), (ii) the investigator is interested in the difference, at the group level, between the performance of the examinees on two sets of items that are supposed to measure a common construct, and (iii) the estimates of the parameters of

[18]The Newton-Raphson algorithm for computing ability estimates involves both derivatives of response probabilities and estimated variances.

the two sets of items are available. [19] At an individual level, the difference between the performance of the examinees on two sets of items was of interest in Finkelman, Weiss, and Kim-Kang (2010) because the difference would quantify the change that occurred in the examinee abilities, in Guo and Drasgow (2010) because a difference would indicate possible cheating, and in Sinharay (2017) because a difference would indicate possible item preknowledge; in all these applications, the null hypothesis is that the ability of the examinees is the same on average over the two sets of items and the alternative hypothesis is that the ability is not the same (due to change/growth or cheating or item preknowledge). One may be interested in quantifying such differences at an aggregate level and EDI_g, EDI_g^A, and EDI_g^N may be applied to quantify the differences. For example, if it is known that a certain subset of items may have been compromised (a problem considered by, for example, Sinharay, 2017), one can apply EDI_g, EDI_g^A, and EDI_g^N to detect possible item preknowledge at an aggregate level; the set of compromised items and the remaining items on the test would constitute the two sets of items in such an application.

Statistical indices for determination of fraudulent erasures 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 (1987) commented that no statistical method on its own can provide conclusive proof that copying occurred (p. 25); the comment is true about erasures 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. However, test-security experts such as Wollack and Cizek (2017, p. 200) have recently presented the viewpoint that statistical evidence based on even a single statistic may constitute conclusive proof of cheating provided the statistic has been properly vetted and accepted by the research community and the degress of aberrance is clearly extreme.

There are several limitations of this paper and, consequently, several related topics can be further investigated. First, it is possible to extend other indices for detection of

[19] Note that in erasure analysis in this paper, the two sets of items are E_j and $\overline{E}_j$.

fraudulent erasures for individual examinees including those suggested by Sinharay and Johnson (2017), Sinharay et al. (2017), and van der Linden and Lewis (2015) to the group level and a future study may compare the extensions suggested in this paper to extensions of other individual-level statistics for detecting fraudulent erasures. Second, while our simulation study was detailed, it is possible to perform more simulations, possibly with other IRT models. Similarly, it is possible to consider applications of the indices to more real data examples. Finally, since classes and schools involve a hierarchical structure where students are nested within classes, which are nested within schools, it is possible to apply a hierarchical model to perform erasure analysis; Skorupski and Egan (2014) suggested a hierarchical linear model for detection of group-level cheating; their approach uses the score on a vertical scale as the response variable and treats an unusually large increase in score for a group from a previous grade as possible evidence of cheating. It may be possible to use a similar approach for an aggregate-level erasure analysis.

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.

Baker, F. B., & Kim, H., S. (2004). *Item response theory: Parameter estimation techniques (2nd ed.).* New York, NY: Marcel Dekker.

Belov, D. I. (2015). Robust detection of examinees with aberrant answer changes. *Journal of Educational Measurement*, *52*, 437–456.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, *57*, 289–300.

Bishop, S., & Egan, K. (2017). Detecting erasures and unusual gain scores: Understanding the status quo. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 193–213). Washington, DC: Routledge.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Casella, G., & Berger, R. L. (2002). *Statistical inference (2nd edition).* Pacific Grove, CA: Duxbury.

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

Chang, H. H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 37–52.

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

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, *23*, 315–345.

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.

Finkelman, M., Weiss, D. J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the adaptive measurement of change. *Applied Psychological Measurement*, *34*, 238–254.

Fremer, J., & Olson, J. F. (2015). *TILSA test security: Lessons learned by state assessment programs in preventing, detecting, and investigating test security irregularities.* Washington DC: Council of Chief State School Officers.

Furr, R. (2010). Yates's correction. In N. Salkind (Ed.), *Encyclopedia of research design* (pp. 1646–1649). Thousand Oaks, CA: SAGE Publications, Inc.

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored internet tests: The z-test and the likelihood ratio test. *International Journal of Selection and Assessment*, *18*, 351–364.

Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying (ACT research report series no. 87-15).* Iowa City, IA: American College Testing.

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

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.

McClintock, J. C. (2015). Erasure analyses: Reducing the number of false positives. *Applied Measurement in Education*, *28*, 14-32.

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.

R Core Team. (2017). *R: A language and environment for statistical computing.* Vienna, Austria.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: John Wiley.

Sinharay, S. (2016). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika*, *81*, 992–1013.

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

Sinharay, S., Duong, M. Q., & Wood, S. W. (2017). A new statistic for detection of aberrant answer changes. *Journal of Educational Measurement*, *54*, 200–217.

Sinharay, S., & Johnson, M. S. (2017). Three new methods for analysis of answer changes. *Educational and Psychological Measurement*, *77*, 54–81.

Skorupski, W. P., & Egan, K. (2014). A Bayesian hierarchical linear modeling approach for detecting cheating and aberrance. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 121–133). New York, NY: Routledge.

Snijders, T. (2001). Asymptotic distribution of person-fit statistics with estimated person parameter. *Psychometrika*, *66*, 331–342.

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. 158–174). New York, NY: Routledge.

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., & Lewis, C. (2015). Bayesian checks on cheating on tests. *Psychometrika*, *80*, 689–706.

Wollack, J. A., & Cizek, G. J. (2017). Test security for licensure and certification examination programs. In S. Davis-Becker & C. Buckendahl (Eds.), *Testing in the professions* (pp. 178–209). New York, NY: Routledge/Taylor & Francis.

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.

Wollack, J. A., & Eckerly, C. (2017). Detecting test tampering at the group level. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 214–231). Washington, DC: Routledge.

Yates, F. (1934). Contingency table involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society*, *1*, 217-235.