

The Use of Theory of Linear Mixed-Effects Models to Detect Fraudulent Erasures at an Aggregate Level

Educational and Psychological
Measurement

2022, Vol. 82(1) 177–200

© The Author(s) 2021

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0013164421994893

journals.sagepub.com/home/epm



Luyao Peng^{1,2} and Sandip Sinharay³ 

Abstract

Wollack et al. (2015) suggested the erasure detection index (EDI) for detecting fraudulent erasures for individual examinees. Wollack and Eckerly (2017) and Sinharay (2018) extended the index of Wollack et al. (2015) to suggest three EDIs for detecting fraudulent erasures at the aggregate or group level. This article follows up on the research of Wollack and Eckerly (2017) and Sinharay (2018) and suggests a new aggregate-level EDI by incorporating the empirical best linear unbiased predictor from the literature of linear mixed-effects models (e.g., McCulloch et al., 2008). A simulation study shows that the new EDI has larger power than the indices of Wollack and Eckerly (2017) and Sinharay (2018). In addition, the new index has satisfactory Type I error rates. A real data example is also included.

Keywords

data forensics, erasure analysis, linear mixed-effects model, empirical best linear unbiased predictor, test fraud

Test administrators are increasingly concerned about fraudulent erasures or “test tampering,” which is a major type of test fraud (e.g., Wollack & Schoenig, 2018).

¹Beijing Language and Culture University, Beijing, China

²ByteDance, Beijing, China

³Educational Testing Service, Princeton, NJ, USA

Corresponding Author:

Luyao Peng, Center for the Cognitive Science of Language, Beijing Language and Culture University, Beijing, China.

Email: luyaopeng@blcu.edu.cn

The educator cheating scandal in Atlanta public schools in 2009 (e.g., Maynes, 2013; Wollack et al., 2015) demonstrated the severity of teacher/administrator tampering and sent a signal that the U.S. states should actively detect test tampering. Consequently, the 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) includes the recommendation that testing programs may use technologies such as computer analyses of erasure patterns in the answer sheets to detect possible irregularities. Erasure analysis is regularly performed for several state tests (e.g., McClintock, 2015)—the analysis mostly involves the flagging of schools with an unexpectedly large number of erasures.

Wollack et al. (2015) suggested the *erasure detection index* (EDI) for detecting fraudulent erasures for individual examinees. Wollack and Eckerly (2017) and Sinharay (2018) extended the index of Wollack et al. (2015) to suggest three EDIs that can be used to perform erasure analysis at an aggregate or group level, where the groups could refer to the classes, schools, or districts that the examinees belong to. The main goal of this article is to continue the research of Wollack and Eckerly (2017) and Sinharay (2018) and to propose a new EDI for performing erasure analysis at an aggregate level using the theory of linear mixed-effects models (LMMs; e.g., McCulloch et al., 2008). It is demonstrated that the new EDI has larger power than the indices of Wollack and Eckerly (2017) and Sinharay (2018). In addition, the new EDI is shown to have satisfactory Type I error rates in all simulation conditions.

The next section includes reviews of the EDIs that were suggested by Wollack et al. (2015), Wollack and Eckerly (2017), and Sinharay (2018) to detect fraudulent erasures. The new EDI, which is based on the theory of LMMs, is introduced in the Method section. In the Simulations section, the Type I error rates and power of the new EDI are compared with those of the existing EDIs. The new EDI is applied to a real data in the Real Data section. Conclusions and extensions are provided in the last section.

Background

Let us consider a test that comprises only dichotomous items whose parameters are assumed to be known and are equal to the estimates computed from a previous calibration using an item response theory (IRT) model. Let us consider the case when the examinees belong to G groups. In most applications of erasure analysis, the groups correspond to the classes or schools or school districts that the examinees belong to. The number of examinees with at least one erasure in group g is denoted as n_g , $g = 1, 2, \dots, G$. The total number of examinees with at least one erasure across all the groups is denoted by $n (= \sum_{g=1}^G n_g)$.¹ The erasures could be fraudulent erasures or benign (where benign means “not fraudulent”) erasures and could have been produced by the examinees and/or educators. Let E_{gk} denote the set of items on which erasures were found for examinee k in group g . For convenience, “examinee

k in group g ” is henceforth used to denote the k th examinee among those with at least one erasure in group g . Thus, the examinees with no erasures contribute only to the item-parameter estimation and are ignored in the erasure analysis. Let E_{gk}^c denote the set of items on which no erasures were found for examinee k in group g . The sets E_{gk} and E_{gk}^c are nonoverlapping and their union is the set of all items administered to examinee k in group g .

Let X_{gk} denote the raw score of examinee k in group g on the items in E_{gk} . The score X_{gk} is also referred as the wrong-to-right (WTR) score (e.g., Wollack et al., 2015). Let μ_{gk} and σ_{gk} , respectively, denote the expectation and standard deviation of X_{gk} given the true ability parameter (θ_{gk}) of the corresponding examinee. The ability θ_{gk} is unknown and is estimated from the responses on the nonerased items, that is, items in E_{gk}^c (Sinharay, 2018; Wollack et al., 2015). The estimates of μ_{gk} and σ_{gk} are denoted as $\hat{\mu}_{gk}$ and $\hat{\sigma}_{gk}$, respectively, and are obtained by replacing θ_{gk} by its estimate $\hat{\theta}_{gk}$ (typically, the maximum likelihood estimate or weighted likelihood estimate of Warm, 1989, is used as an estimate of the true ability) in the expressions of μ_{gk} and σ_{gk} as

$$\hat{\mu}_{gk} = \hat{E}(X_{gk}) = \sum_{i \in E_{gk}} P_i(\hat{\theta}_{gk}),$$

$$\hat{\sigma}_{gk} = \sqrt{\hat{\sigma}_{gk}^2} = \sqrt{\widehat{Var}(X_{gk})} = \sqrt{\sum_{i \in E_{gk}} P_i(\hat{\theta}_{gk})(1 - P_i(\hat{\theta}_{gk}))}, \tag{1}$$

where $P_i(\hat{\theta}_{gk})$ is the estimated probability of a correct answer on item i by examinee k in group g . The terms $P_i(\hat{\theta}_{gk})$ are determined by the underlying IRT model. For example, if the three-parameter logistic model (3PLM) is used in the analysis, $P_i(\hat{\theta}_{gk})$ would be given by

$$P_i(\hat{\theta}_{gk}) = \hat{c}_i + (1 - \hat{c}_i) \frac{e^{\hat{a}_i(\hat{\theta}_{gk} - \hat{b}_i)}}{1 + e^{\hat{a}_i(\hat{\theta}_{gk} - \hat{b}_i)}},$$

where \hat{a}_i , \hat{b}_i , and \hat{c}_i , respectively, are the estimated slope, difficulty, and guessing parameters of item i . Wollack et al. (2015), Wollack and Eckerly (2017), and Sinharay (2018) used the nominal response model (Bock, 1972) as the IRT model, but noted that other IRT models such as the 3PLM can be used to compute the EDIs as well.

The individual-level EDI (Wollack et al., 2015) for examinee k in group g is defined as

$$EDI_{Ind} = \frac{X_{gk} - \hat{\mu}_{gk} - 0.5}{\hat{\sigma}_{gk}}.$$

A large positive value of EDI_{Ind} indicates potentially fraudulent erasures for examinee k in group g .

Wollack and Eckerly (2017) extended the individual-level EDI to define the aggregate-level EDI, or, EDI_g , for detecting fraudulent erasures at an aggregate level, as

$$EDI_g = \frac{\sum_{k=1}^{n_g} (X_{gk} - \hat{\mu}_{gk}) - 0.5}{\sqrt{\sum_{k=1}^{n_g} \hat{\sigma}_{gk}^2}}, \quad g = 1, \dots, G. \quad (2)$$

The groups could refer to the classes, schools, or districts that the examinees belong to. Wollack and Eckerly (2017) assumed that EDI_g approximately follows the standard normal distribution under the null hypothesis of no fraudulent erasures. The null hypothesis is rejected and group g is flagged for potentially fraudulent erasures if EDI_g is a large positive number.

Sinharay (2018) suggested two modified versions of EDI_g . The first modified version, EDI_g^N , is defined as

$$EDI_g^N = \frac{\sum_{k=1}^{n_g} (X_{gk} - \hat{\mu}_{gk})}{\sqrt{\sum_{k=1}^{n_g} \hat{\sigma}_{gk}^2}}, \quad g = 1, \dots, G. \quad (3)$$

The second modified version, EDI_g^A , is defined as

$$EDI_g^A = \frac{\sum_{k=1}^{n_g} (X_{gk} - \hat{\mu}_{gk})}{\sqrt{\sum_{k=1}^{n_g} \widehat{Var}(X_{gk} - \hat{\mu}_{gk})}}, \quad g = 1, \dots, G, \quad (4)$$

where $\widehat{Var}(X_{gk} - \hat{\mu}_{gk}) = \hat{\sigma}_{gk}^2 + \widehat{Var}(\hat{\theta}_{gk}) \left[\sum_{i \in E_{gk}} P'_i(\hat{\theta}_{gk}) \right]^2$, and $P'_i(\hat{\theta}_{gk})$ is the derivative of $P_i(\hat{\theta}_{gk})$. Sinharay (2018) proved, using theoretical derivations and simulations, that both EDI_g^N and EDI_g^A approximately follow the standard normal distribution under the null hypothesis of no fraudulent erasures for group g . Sinharay (2018) found that the power of EDI_g^A was either equal to or 0.01 less than that of EDI_g^N up to two decimal places in all the simulation cases, and that the power of either of EDI_g^A or EDI_g^N was slightly larger, occasionally by up to 0.05, than that of EDI_g .

Equations (2) to (4) show that the numerators of EDI_g , EDI_g^N , and EDI_g^A include the quantity

$$\sum_{k=1}^{n_g} (X_{gk} - \hat{\mu}_{gk}) = \sum_{k=1}^{n_g} X_{gk} - n_g \bar{\mu}_g = \sum_{k=1}^{n_g} (X_{gk} - \bar{\mu}_g),$$

where

$$\bar{\mu}_g = \frac{1}{n_g} \sum_{k=1}^{n_g} \hat{\mu}_{gk}.$$

The quantity $\bar{\mu}_g$ is the estimator of $\frac{1}{n_g} \sum_{k=1}^{n_g} \mu_{gk} = \bar{\mu}_g$, the average expected WTR score for group g . When n_g is small, the estimator $\bar{\mu}_g$ may have a large standard error, and consequently, EDI_g , EDI_g^N , and EDI_g^A , which are estimates themselves,² may be far, on average, from the population quantities they intend to estimate, and may have poor Type I error rates and power. In fact, Sinharay (2018) found the Type I error rates of all the aggregate-level EDIs to be considerably smaller than the nominal level for small schools (e.g., those of EDI_g^A were smaller than 0.007 at nominal level of 0.01 for schools with 15 students). He also found the power of EDI_g^A to be considerably smaller for small schools than for larger schools. Consequently, an extension of EDI_g , EDI_g^N , and EDI_g^A that involves a more accurate estimate (compared with $\bar{\mu}_g$) of $\frac{1}{n_g} \sum_{k=1}^{n_g} \mu_{gk}$ may have better Type I error rates and power compared with EDI_g , EDI_g^N , and EDI_g^A , at least for small schools. We used theory from LMMs (e.g., McCulloch et al., 2008) to suggest such an extension.

Method

In this section, we discuss the derivation of a new index, which is a modified version of EDI_g , for detecting fraudulent erasures at an aggregate level. The new index is based on the theory of LMMs (e.g., McCulloch et al., 2008). Appendix A includes a discussion of the theory of LMMs that is relevant to this article.

The Model

To apply the theory of LMMs to aggregate-level erasure analysis, we express $\hat{\mu}_{gk}$ as

$$\hat{\mu}_{gk} = \mu + b_g + e_{gk} = \mu_g + e_{gk}, \quad k = 1, \dots, n_g, \quad g = 1, \dots, G, \tag{5}$$

where μ is the expected mean WTR score for the whole population, b_g is the random effect for group g ,

$$\mu_g = \mu + b_g,$$

and e_{gk} is the random error of examinee k in group g . Equation (5) looks like a typical representation of an LMM. For example, the equation is equivalent to Equation (2.15) of McCulloch et al. (2008). It is assumed in Equation (5) that b_g and e_{gk} are independent, $b_g \stackrel{iid}{\sim} N(0, \sigma_b^2)$, $e_{gk} \stackrel{iid}{\sim} N(0, \text{Var}(e_{gk}) = \sigma_e^2 w_{gk})$, σ_b^2 and σ_e^2 are unknown variance components, and the w_{gk} 's are known scalar quantities. Since the $\hat{\mu}_{gk}$'s may depend on the individual ability estimates $\hat{\theta}_{gk}$'s, the $\text{Var}(e_{gk})$'s in the above LMM may depend on the $\hat{\theta}_{gk}$'s and hence can be unequal over the examinees. Furthermore, if the estimated variance of $\hat{\theta}_{gk}$, $\widehat{\text{Var}}(\hat{\theta}_{gk})$, is larger, then the extent of uncertainty is larger in estimating $\hat{\mu}_{gk}$. Therefore, we set the w_{gk} 's equal to $\widehat{\text{Var}}(\hat{\theta}_{gk})^3$ so that $\text{Var}(e_{gk})$ is proportional to $\widehat{\text{Var}}(\hat{\theta}_{gk})$. The model provided in Equation (5) is the same as that provided in Equation (A2) in Appendix A, with $\hat{\mu}_{gk}$ in Equation (5) playing the role of y_{gk} in Equation (A2).

Denoting $\hat{\boldsymbol{\mu}}$ to be the $n \times 1$ vector that comprises the $\hat{\mu}_{gk}$'s of all the examinees with erasures over all the groups, Equation (5) can be expressed using matrix notation as

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\mu} \mathbf{1}_n + \mathbf{U} \mathbf{b} + \mathbf{e}, \tag{6}$$

where $\mathbf{1}_n$ is an $n \times 1$ vector of 1s, \mathbf{U} is an $n \times G$ known matrix indicating the group memberships of the examinees and is given by

$$\mathbf{U} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \vdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \vdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_G} & \dots & \dots & \mathbf{1}_{n_G} \end{pmatrix}, \tag{7}$$

where, for example, $\mathbf{0}_{n_1}$ is an $n_1 \times 1$ vector of zeroes, $\mathbf{b} = (b_1, b_2, \dots, b_G)'$ is a $G \times 1$ vector of random effects corresponding to the groups, and \mathbf{e} is an $n \times 1$ vector that comprises all the e_{gk} 's. The above equations imply that $\mathbf{b} \sim MVN(0, \sigma_b^2 \mathbf{I}_G)$ and $\mathbf{e} \sim MVN(0, \sigma_e^2 \mathbf{W})$, where MVN denotes the multivariate normal distribution and \mathbf{W} is an $n \times n$ diagonal matrix comprising the w_{gk} 's. The vectors \mathbf{b} and \mathbf{e} are independent of each other.

The Best Linear Unbiased Predictor and Empirical Best Linear Unbiased Predictor

Equation (6) is just like Equation (A3) of Appendix A, which provides the classic matrix formulation of the LMM, with $\hat{\boldsymbol{\mu}}$ in Equation (6) playing the same role as that of \mathbf{y} in Equation (A3). Therefore, the theory of LMMs that is discussed in Appendix A is directly applicable to the model provided by Equations (5) and (6) and the best linear unbiased predictor (BLUP) of $\boldsymbol{\mu}_g$ can be obtained using Equation (A4) of Appendix A. Simplifications⁴ of the expression provided in Equation (A4) lead to the following expression of the BLUP:

$$\hat{\mu}_g^{BLUP}(\sigma_b^2, \sigma_e^2) = \tilde{\mu} + \gamma_g(\tilde{\mu}_g - \tilde{\mu}) = (1 - \gamma_g)\tilde{\mu} + \gamma_g\tilde{\mu}_g, \quad g = 1, \dots, G, \quad (8)$$

where

$$\begin{aligned} \gamma_g &= \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2 \left(\sum_{k=1}^{n_g} w_{gk}^{-1} \right)^{-1}}, \\ \tilde{\mu} &= \frac{\mathbf{1}' \text{diag}(\sigma_b^2 \mathbf{J}_{n_g} + \sigma_e^2 \mathbf{W}_{n_g})^{-1} \hat{\mu}}{\mathbf{1}' \text{diag}(\sigma_b^2 \mathbf{J}_{n_g} + \sigma_e^2 \mathbf{W}_{n_g})^{-1} \mathbf{1}}, \\ \tilde{\mu}_g &= \frac{\sum_{k=1}^{n_g} w_{gk}^{-1} \hat{\mu}_{gk}}{\sum_{k=1}^{n_g} w_{gk}^{-1}}, \end{aligned}$$

$\text{diag}(\sigma_b^2 \mathbf{J}_{n_g} + \sigma_e^2 \mathbf{W}_{n_g})$ is a block diagonal matrix of order $n \times n$ with blocks $\sigma_b^2 \mathbf{J}_{n_g} + \sigma_e^2 \mathbf{W}_{n_g}$, \mathbf{J}_{n_g} is a matrix of order $n_g \times n_g$ with all elements being 1, and \mathbf{W}_{n_g} is an $n_g \times n_g$ diagonal matrix of the w_{gk} 's for group g , $\tilde{\mu}_g$ is a weighted average of $\hat{\mu}_{gk}$'s over group g .

The right-hand side of Equation (8) is a weighted sum of $\tilde{\mu}_g$, which is the estimated mean WTR score for group g , and $\tilde{\mu}$, which is the estimated overall mean WTR score across all groups. If $\sigma_e^2 (\sum_{k=1}^{n_g} w_{gk}^{-1})^{-1}$ in γ_g is relatively large (that can happen, e.g., if σ_e^2 is large or n_g is small or w_{gk} 's are large), then γ_g will be small and more weight in the BLUP will be given to the estimated overall mean $\tilde{\mu}$. If, on the other hand, σ_b^2 is relatively large, then more weight in the BLUP will be given to the group-specific mean $\tilde{\mu}_g$.

Equation (8) can be used to compute the BLUP when the variance components σ_b^2 and σ_e^2 are known, but they are unknown for a real data set and have to be estimated. The variances σ_b^2 and σ_e^2 can be estimated using the restricted maximum likelihood (REML) method (e.g., McCulloch et al., 2008) under the model in Equation (5). By replacing σ_b^2 and σ_e^2 by their REML estimates in Equation (8), one obtains what is referred to as the empirical best linear unbiased predictor (EBLUP; e.g., Agresti, 2015; Harville, 1991) of μ_g as

$$\hat{\mu}_g^{EBLUP} = \hat{\mu}_g^{EBLUP}(\hat{\sigma}_b^2, \hat{\sigma}_e^2) = (1 - \hat{\gamma}_g^{REML})\hat{\mu}^{REML} + \hat{\gamma}_g^{REML}\tilde{\mu}_g, \quad (9)$$

where $\hat{\mu}^{REML}$ and $\hat{\gamma}_g^{REML}$ are obtained by replacing σ_e^2 and σ_b^2 , respectively, by their corresponding REML estimates ($\hat{\sigma}_e^2$ and $\hat{\sigma}_b^2$), in $\tilde{\mu}$ and γ_g , respectively. Note that the computation of $\hat{\mu}_g^{EBLUP}$ (that corresponds to group g) in Equation (9) involves the borrowing of information from the other groups in the sample. Prasad and Rao (1990) pointed out that the estimator $\hat{\mu}_g^{EBLUP}$ is essentially identical to the empirical Bayes estimator (that involves borrowing of information; Ghosh & Meeden, 1986)

of μ_g , and has a smaller mean square error (*MSE*) compared with $\tilde{\mu}_g$ (i.e., the simpler estimate of μ_g), especially when n_g is small (Ghosh & Meeden, 1986). Note that empirical Bayes estimators have been found to be more accurate compared with traditional (frequentist) estimators in various contexts in educational measurement including validity studies and survival analysis (e.g., Braun, 1989) and differential item functioning (Zwick et al., 1999). Therefore, the EBLUP defined in Equation (9) is anticipated to perform well in aggregate-level erasure analysis. Thus, the EBLUP represents the more accurate estimate of $\frac{1}{n_g} \sum_{k=1}^{n_g} \mu_{gk}$ that we sought toward the end of the previous section of this article.

To estimate $\widehat{Var}(\hat{\mu}_g^{EBLUP})$, which is the variance of $\hat{\mu}_g^{EBLUP}$, we applied the bootstrap method (e.g., Efron, 1981) as discussed in González-Manteiga et al. (2008).

The Index EDI_g^{EBLUP}

After computing $\hat{\mu}_g^{EBLUP}$ and $\widehat{Var}(\hat{\mu}_g^{EBLUP})$ as described above, the new aggregate-level EDI, denoted by EDI_g^{EBLUP} , is defined as

$$\begin{aligned}
 EDI_g^{EBLUP} &= \frac{\sum_{k=1}^{n_g} (X_{gk} - \hat{\mu}_g^{EBLUP})}{\sqrt{\sum_{k=1}^{n_g} \widehat{Var}(X_{gk} - \hat{\mu}_g^{EBLUP})}} \\
 &= \frac{\sum_{k=1}^{n_g} (X_{gk} - \hat{\mu}_g^{EBLUP})}{\sqrt{\sum_{k=1}^{n_g} \sum_{i \in E_{gk}} P_i(\hat{\theta}_{gk})(1 - P_i(\hat{\theta}_{gk})) + n_g \widehat{Var}(\hat{\mu}_g^{EBLUP})}} \tag{10}
 \end{aligned}$$

because $\widehat{Var}(X_{gk} - \hat{\mu}_g^{EBLUP}) = \widehat{Var}(X_{gk}) + \widehat{Var}(\hat{\mu}_g^{EBLUP})$ due to the independence of X_{gk} and $\hat{\mu}_g^{EBLUP}$ given θ_{gk} and $\widehat{Var}(X_{gk}) = \sum_{i \in E_{gk}} P_i(\hat{\theta}_{gk})(1 - P_i(\hat{\theta}_{gk}))$. Note that the

numerator of EDI_g^{EBLUP} is different from that of EDI_g^N and EDI_g^A only in the use of $\hat{\mu}_g^{EBLUP}$ instead of $\hat{\mu}_{gk}$. The denominator of EDI_g^{EBLUP} is larger than that of EDI_g^N , but could be smaller or larger than that of EDI_g^A depending on the magnitude of $\widehat{Var}(\hat{\mu}_g^{EBLUP})$. The R code (e.g., R Core Team, 2020) for computing EDI_g^{EBLUP} using the R package sae (Molina & Marhuenda, 2015) as well as the existing EDIs for all schools for a data set is provided in Appendix B. Note that $\hat{\mu}_g^{EBLUP}$ or EDI_g^{EBLUP} cannot be computed for a group in which erasures were found for all or none of the examinees. For groups without any erasures, erasure analysis cannot be performed and no EDIs can be computed. For groups in which all examinees made erasures, it

is possible to use one of the existing EDIs even if a decision has been taken to employ EDI_g^{EBLUP} for the corresponding test.

Simulation Study

A simulation study was performed to compare the Type I error rates and the power of the new index (EDI_g^{EBLUP}) with those of EDI_g , EDI_g^N , and EDI_g^A .

Design of the Simulation

In the simulation study, we artificially created several data sets that look like those originating from a test with erasures, starting from a real test data set with actual scores on 46 dichotomous items of 2,710 examinees who belonged to a total of 29 schools. The number of examinees in the schools ranged between 6 and 305, with the average number being about 93.

The following three factors were varied in the simulations:

- The number of tampered schools, that is, schools in which tampering (or fraudulent erasures) took place (NTamp: 0 or 6)
- The percentage of erasure victims in a tampered school (PercVictim: 5, 10, or 20)
- The number of fraudulent erasures per erasure victim (Nerasure: 3, 5, or 10)

When NTamp = 0, there are no tampered schools and no erasure victims. When NTamp = 6, the other two factors (PercVictim and Nerasure) are fully crossed. Thus, there are a total of 10 simulation conditions (one for NTamp = 0 and nine for NTamp = 6). For each of the nine simulation conditions with NTamp = 6, where a simulation condition is characterized by a specific combination of PercVictim and Nerasure, the following steps were performed to create the artificial data:

- Randomly sample six schools from the 29 schools and assign them as the “tampered” schools. The 23 schools that are not tampered are referred as the “untampered” schools.
- Draw a random sample of examinees (of size 5%, 10%, or 20%, depending on the value of PercVictim) from each of the six tampered schools. The examinees in these samples play the role of the erasure victims. The remaining examinees are nonvictims. Then,
 1. For each victim, artificially create fraudulent erasures by sampling Nerasure (that is equal to 3, 5, or 10) items from all the items that the victim answered incorrectly and changing the scores of those items from 0 to 1. If the number of items with incorrect answers is smaller than Nerasure for a victim, then change the scores of all of them to 1. Also, for each victim, randomly select two items (that do not overlap

- with the Nerasure items) and mark them as those with benign erasures, but do not change their scores. Such a strategy was used to simulate benign erasures in Sinharay et al. (2017).
2. Randomly sample 95% of the nonvictim examinees in the school. For each sampled examinee, randomly sample two items and mark them as those with benign erasures, but do not change the scores on those items. The remaining 5% nonvictims are assumed to not produce any erasures.
 - All the examinees in the 23 untampered schools are nonvictims. For such schools, sample 95% of the examinees to have two benign erasures, sample the items that have benign erasures for these examinees, and assume that the remaining 5% examinees do not produce any erasures.

For the simulation condition with NTamp = 0, all schools are untampered; so for this condition, for each school, 95% of the examinees are sampled to have two benign erasures and the remaining 5% examinees are assumed to not produce any erasures.

Computations

For each simulation condition, using the resulting artificial data (that not only includes many of the item scores from the original data set but also includes some changed item scores, and the erasure indicators for each examinee–item combination), we computed the $\hat{\theta}_{gk}$'s, which are the weighted likelihood estimates (e.g., Warm, 1989) of the examinee ability parameters, and the $\widehat{Var}(\hat{\theta}_{gk})$'s using the R package *irtoys* (Partchev et al., 2017). The estimated item parameters from the original data set were used in this computation.⁵ The 3PLM was used as the IRT model in all calculations. Then, EDI_g , EDI_g^N , and EDI_g^A were computed for each school using Equations (2) to (4) and EDI_g^{EBLUP} was computed for each school using the following steps:

- Compute $\sum_{k=1}^{n_g} X_{gk}$, the sum of the observed WTR scores, for each school.
- Compute $\sum_{k=1}^{n_g} \sum_{i \in E_{gk}} P_i(\hat{\theta}_{gk})(1 - P_i(\hat{\theta}_{gk}))$, the sum of the estimated variance of the observed WTR scores, for each school.
- Compute the estimated expected WTR score for examinee k in school g as $\hat{\mu}_{gk} = \sum_{i \in E_{gk}} P_i(\hat{\theta}_{gk})$
- Set w_{gk} equal to $\widehat{Var}(\hat{\theta}_{gk})$ for $k = 1, \dots, n_g, g = 1, \dots, 29$.
- Use the R package *sae* (Molina & Marhuenda, 2015) and the w_{gk} 's to fit the LMM given by Equation (5) and compute $\hat{\mu}_g^{EBLUP}$ given by Equation (9) and $\widehat{Var}(\hat{\mu}_g^{EBLUP})$ for each school using the bootstrap method.

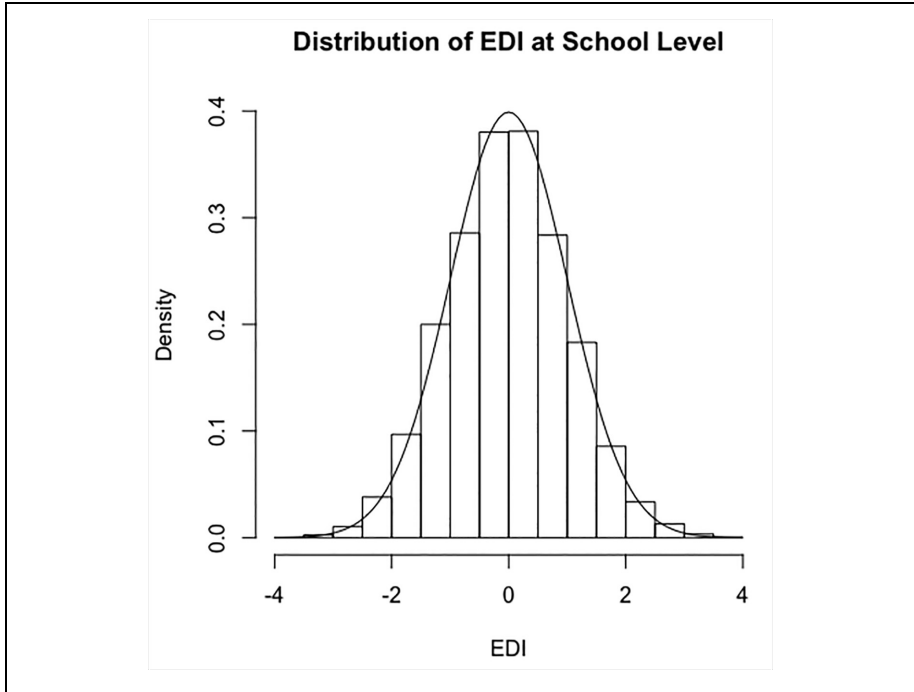


Figure 1. The histogram of the distribution of EDI_g^{EBLUP} for NTamp = 0.

- Compute EDI_g^{EBLUP} using Equation (10) for all schools using $\hat{\mu}_g^{EBLUP}$ and $\widehat{Var}(\hat{\mu}_g^{EBLUP})$ computed in the previous step.

We replicated the above steps 1,000 times for each of the 10 simulation conditions. Then, we used the output of the simulation to compute the Type I error rates and power of $EDI_g, EDI_g^N, EDI_g^A,$ and EDI_g^{EBLUP} .

Results

Distribution of the Indices Under the Null Hypothesis

Figure 1 shows the histogram of the distribution of the values of EDI_g^{EBLUP} for all the schools for the simulation condition with NTamp = 0, a condition that is associated with no fraudulent erasures, and hence the distribution of EDI_g^{EBLUP} should be close to the standard normal distribution. The standard normal distribution is shown in the figure using a solid line for comparison. It seems that the empirical distribution of EDI_g^{EBLUP} under no fraudulent erasures is close to the standard normal distribution, which is a favorable property of EDI_g^{EBLUP} .

Table 1. The Average Type I Error Rates of the Indices.

Index	$\alpha = .05$	$\alpha = .01$	$\alpha = .001$
EDI_g	.0262	.0053	.00028
EDI_g^N	.0352	.0068	.00034
EDI_g^A	.0325	.0061	.00034
EDI_g^{EBLUP}	.0541	.0103	.00110

Table 2. The Average Power of the Indices.

Index	$\alpha = .05$	$\alpha = .01$	$\alpha = .001$
EDI_g	.660	.538	.419
EDI_g^N	.682	.556	.432
EDI_g^A	.668	.536	.408
EDI_g^{EBLUP}	.719	.591	.465

Results on Type I Error Rates

Table 1 shows the Type I error rates at the .05, .01 and .001 levels of significance for the four indices, EDI_g , EDI_g^N , EDI_g^A , and EDI_g^{EBLUP} . The Type I error rates of the existing EDIs are somewhat smaller than the nominal level, while those of EDI_g^{EBLUP} are close to the nominal levels. Though the Type I error rates of EDI_g^{EBLUP} are slightly larger than the nominal level, they are satisfactory in all cases according to Cochran's criterion for robustness (Cochran, 1952) that deems Type I error rates smaller than .06, .015, and .0015 to be satisfactory at levels .05, .01, and .001, respectively.

Results on Power

Table 2 shows the values of power of EDI_g , EDI_g^N , EDI_g^A , and EDI_g^{EBLUP} at significance levels of .05, .01 and .001, averaged over all the levels of the two simulation factors PercVictim and Nerasure for NTamp = 6. The values of the average power for EDI_g^{EBLUP} are the largest for each significance level while those for EDI_g are the smallest for any significance level.

Figure 2 shows the power of EDI_g , EDI_g^N , EDI_g^A , and EDI_g^{EBLUP} for different values of Nerasure and PercVictim for NTamp = 6 at the .05 level (the comparative performance of the indices is similar for levels .01 and .001—so plots for these levels are not shown). Each panel of the figure shows the power (along the Y-axis) for the three values of PercVictim (X-axis) for each index for a significance level. The three

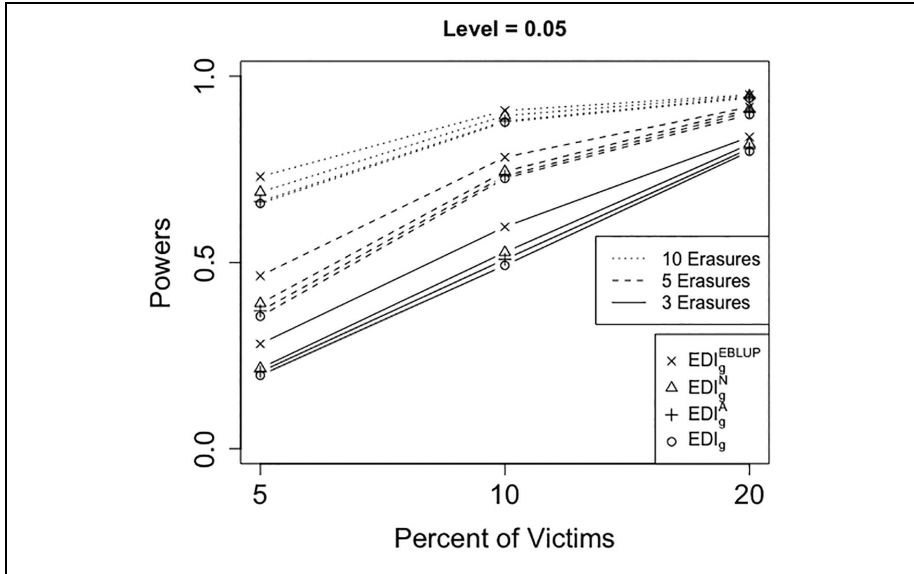


Figure 2. Power of the indices for different simulation conditions at the .05 level.

line types correspond to the three values of *Nerasure* (3, 5, and 10). The values of power for EDI_g , EDI_g^N , EDI_g^A , and EDI_g^{EBLUP} are, respectively, denoted by circle, triangle, plus, and multiplication symbols. The figure shows that the power of EDI_g^{EBLUP} is larger than those of all the other indices across all the simulation conditions. In addition, the difference in power of EDI_g^{EBLUP} and the other indices increases as the number of erasures decreases—this result implies that EDI_g^{EBLUP} would allow investigators to detect fraudulent schools or districts even when the number of fraudulent erasures is small on average.

We studied the comparative performance of the indices using a simulation design similar to the one above starting from two other data sets that include larger numbers of examinees. The results in these additional simulations (not shown here and can be obtained on request from the authors), especially those regarding EDI_g^{EBLUP} being more powerful compared with the other indices especially for small schools, were very similar.

Application to Real Data

We analyzed a data set that includes the responses of 19,107 fifth-grade students to 53 dichotomous items on a state mathematics test—the data set is a subset of the K-12 education data set described in Cizek and Wollack (2017). The students belonged to 820 schools in 446 districts. The data providers did not reveal if there were any fraudulent erasures on the test. Erasures were captured through a scanning process

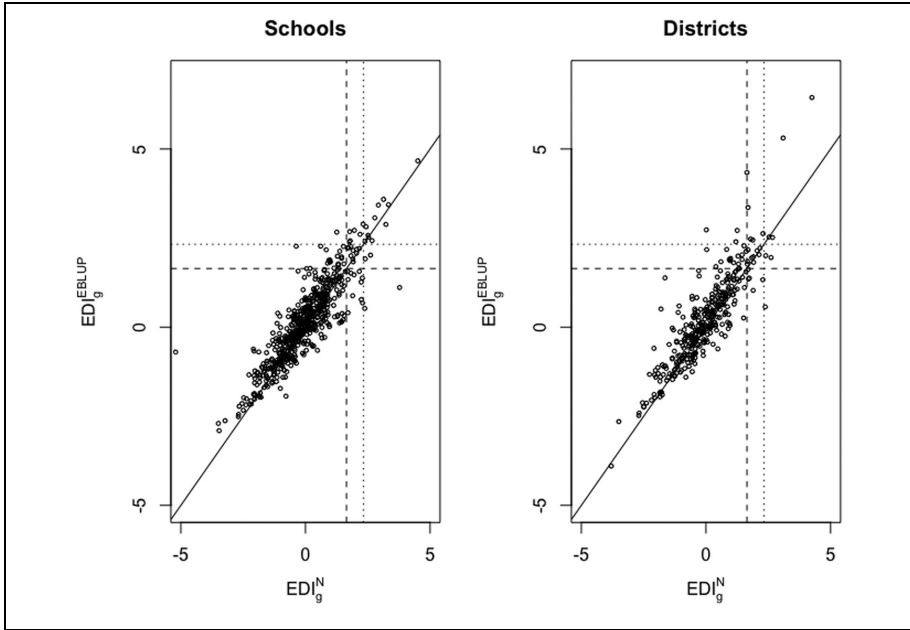


Figure 3. Values of the new and existing EDIs for schools and districts.

by looking for “light marks” (Cizek & Wollack, 2017, p. 15). On average, the number of erasures per examinee is two (i.e., 3.7% of all the items on the test), which is about twice that of what is typically found in similar assessments (see, e.g., Wollack et al., 2015). Since the LMM does not work for schools or districts in which all or none of the examinees have erasures,⁶ those schools or districts were used to estimate the item parameters and then removed from the data, which left 573 schools and 346 districts in our sample. The 3PL model was used as the IRT model. The four EDIs were computed for the 573 schools and 346 districts in the data set.

The left and right panels of Figure 3 show plots of EDI_g^N (along the X -axis) versus EDI_g^{EBLUP} (along the Y -axis) for the districts and schools, respectively.⁷ For convenience, the panels of the figure include vertical and horizontal dashed lines at the 95th percentile point (1.64) of the standard normal distribution and vertical and horizontal dotted lines at the 99th percentile point (2.33) of the standard normal distribution. Figure 3 shows that EDI_g^{EBLUP} is different from EDI_g^N for quite a few districts and schools and these differences will often lead to EDI_g^{EBLUP} being statistically significant even though EDI_g^N is statistically nonsignificant or vice versa. Table 3 shows the number of schools and districts for which EDI_g^{EBLUP} , EDI_g^N , and EDI_g^A are statistically significant at levels of .05 and .01.

Table 3 shows that the use of EDI_g^{EBLUP} instead of EDI_g^N or EDI_g^A will lead to different numbers of schools and districts being flagged for potentially fraudulent

Table 3. The Number of Districts and Schools With Statistically Significant Values of the Indices.

Level	Districts			Schools		
	EDI_g^{EBLUP}	EDI_g^N	EDI_g^A	EDI_g^{EBLUP}	EDI_g^N	EDI_g^A
.05	38	25	20	50	42	40
.01	12	7	7	18	15	14

erasures at both significance levels. The difference between EDI_g^{EBLUP} and the existing indices is larger for districts than for schools. For example, while EDI_g^N and EDI_g^A are significant for seven districts, EDI_g^{EBLUP} is significant for 12 districts at the level of .01. Because it is unknown which schools or districts are truly guilty of fraudulent erasures, it is difficult to comment with certainty on whether the correct decision was made from EDI_g^{EBLUP} or the existing EDIs for the schools/districts for which one of these is significant and the others are not. But Table 3 indicates, like Figure 3, that the conclusion on whether to flag for fraudulent erasures will differ in general between EDI_g^{EBLUP} and any of the existing indices for quite a few schools/districts. Thus, the use of EDI_g^{EBLUP} could lead to practically different conclusions for several schools/districts in an investigation of test fraud. Also, the satisfactory Type I error rate of EDI_g^{EBLUP} (see Table 1) and the larger power of the index compared with the existing indices (see Table 2) make us confident that the correct decision for real data would be made more often by EDI_g^{EBLUP} .

Conclusions

We followed up on the research of Wollack et al. (2015), Wollack and Eckerly (2017), and Sinharay (2018), and suggested a new index, EDI_g^{EBLUP} , for the detection of fraudulent erasures at the aggregate level. The derivation of the new index is based on the theory of linear mixed-effects models. The Type I error rates of the new index are close to the nominal level and the new index is more powerful than the indices of Wollack and Eckerly (2017) and Sinharay (2018) in all the simulation conditions that we considered.

Fraudulent erasures or test tampering is a major problem that is faced by test administrators (e.g., Wollack & Schoenig, 2018). For a large state with many schools, a difference in power of only a couple of percentages may lead to many more detections of fraudulent schools. Thus, the investigator should try to use the most powerful method that also has satisfactory Type I error rates. If computation is not an issue, the theoretical and simulation results in this article indicate that EDI_g^{EBLUP} is preferable compared with the existing indices.

Note that our article is not the first to apply LMMs to aggregate-level erasure analysis. Bishop et al. (2011) suggested the use of three LMMs to detect fraudulent erasures at the aggregate level. However, Bishop et al. (2011) applied LMMs to estimate the mean WTR scores of examinee groups and used those estimates to detect fraudulent erasures. In contrast, we applied LMMs to estimate the expected WTR scores of examinee groups and used those estimates to compute group-level EDIs, which are then used to detect fraudulent erasures. In addition, while it is unclear that Bishop et al. (2011) allowed the error variances to be unequal in their LMM, we performed our derivations under the assumption of unequal error variances.

Note that the model used in this article (those provided in Equations 5 and 6) can also be referred to as a linear random-effects model, which is a special case of the LMMs, because the only fixed effect in our model is the overall mean μ . We refer to the model as an LMM in our article because the results of, for example, C. R. Henderson (1975) and Prasad and Rao (1990), which form the basis of the new index, apply to LMMs.

Erasures usually occur on paper-and-pencil tests and the analysis to detect fraudulent erasures for such tests is referred to as erasure analysis (e.g., McClintock, 2015). However, erasures essentially mean answer changes, and computer-based tests (CBTs) may also suffer from fraudulent answer changes. Foster (2013) noted that hackers can gain unauthorized access to the scoring system of a CBT and change lower scores to higher ones. Thus, the indices discussed in this article also apply to CBTs.

Several extensions on EDI_g^{EBLUP} can be considered in the future. The LMMs are flexible in incorporating auxiliary variables to predict the dependent variable for a subpopulation especially when the sample size in that subpopulation is small. So, it is possible to extend the LMM used in this article by adding auxiliary variables such as the characteristics of the school, to improve the precision of the prediction of the expected mean WTR score at school level, which may improve the power of EDI_g^{EBLUP} further. This extension may be especially beneficial when some schools have small number of examinees with erasures. We used the REMLs of the variance components (σ_e^2 and σ_b^2) of the random effects, but other estimates (e.g., those suggested by Prasad & Rao, 1990) can be used as well. The LMM applied in this article has random intercepts; however, other LMMs including the random-intercept-and-slope LMM (e.g., McCulloch et al., 2008) may be used instead. Finally, the methodology presented in this article can be seen as an extension of the individual-level EDI (Wollack et al., 2015) to the aggregate level using theory from LMMs. It is possible to perform future research on the extension of other individual-level statistics for erasure analysis (e.g., those suggested by Sinharay et al., 2017; Sinharay & Johnson, 2017) and individual-level statistics for detecting other types of test fraud (e.g., those suggested by Sinharay, 2017; Sinharay & Jensen, 2019; Wollack, 1997) to the aggregate level using theory from LMMs.

Appendix A

The Nested-Error Mixed-Effects Model

The General Form of the Linear Mixed-Effects Models

To reduce the estimation error for small samples, LMMs (e.g., McCulloch et al., 2008) are often used because of their flexibility in effectively combining various sources of information and explaining different sources of errors (e.g., Jiang & Lahiri, 2006).

The general LMMs in matrix form are given by

$$y = X\beta + U_1b_1 + \dots + U_sb_s + e, \tag{A1}$$

where y is an $n \times 1$ vector, containing the values of the dependent variable for the sample, X is a known $n \times p$ matrix containing the values of p covariates for the sampled individuals, β is a $p \times 1$ vector of coefficients for the fixed effects, U_i is a known $n \times t_i$ incidence matrix for the i th random effect, b_i is a $t_i \times 1$ vector of the i th random effect, $i = 1, \dots, s$, e is an $n \times 1$ vector of random errors. It is assumed that $b_i \sim MVN(0, \sigma_i^2 I_{t_i})$ and $e \sim MVN(0, \sigma_e^2 W_n)$, where I_{t_i} is an identity matrix of order t_i and W_n is a known diagonal matrix that consists of the weights of the error variance of each individual as the diagonal elements. When the weights are equal to 1, W_n is an identity matrix of order n . The quantities b_i 's and e are assumed to be independent.

Random-Intercept Linear Mixed Models

Random-intercept linear mixed models (RILMM; e.g., Agresti, 2015; McCulloch et al., 2008) are special cases of LMMs described in Equation (A1) and are given by

$$y_{gk} = x_{gk}'\beta + b_g + e_{gk}, j = 1, \dots, n_g, g = 1, \dots, G, \tag{A2}$$

where y_{gk} is the value of the response variable for individual k in group g , x_{gk} is a known $p \times 1$ vector containing the values of p covariates for individual k in group g , b_g is the random effect of group g , and e_{gk} is the random error for individual k in group g . It is assumed that $b_g \stackrel{iid}{\sim} N(0, \sigma_b^2)$. Depending on whether the variances of random errors are assumed to be equal, $e_{gk} \stackrel{iid}{\sim} N(0, \sigma_e^2)$ or $e_{gk} \stackrel{iid}{\sim} N(0, \sigma_e^2 w_{gk})$, where w_{gk} is the weight of the error variance for individual k in group g , the quantities b_g and e_{gk} are assumed to be independent of each other.

The matrix form of the model in Equation (A2) is given by

$$y = X\beta + Ub + e, \tag{A3}$$

where $b \sim MVN(0, D)$, $e \sim MVN(0, R)$, $D = \sigma_b^2 I_{t_1}$, $R = \sigma_e^2 W_n$, and $W_n =$ a diagonal matrix with w_{gk} 's as its elements. If $W_n = I_n$, the variances of e_{gk} 's are equal. Equation (A3) implies that $Var(y) = UDU' + R = V$. The vectors b and e are assumed to be independent.

Best Linear Unbiased Predictors Under Random-Intercept Linear Mixed Models

Users of an RILMM given by Equation (A3) are often interested in the prediction of linear combinations of the form

$$\eta = \mathbf{k}'\boldsymbol{\beta} + \mathbf{m}'\mathbf{b},$$

where \mathbf{k} is an $p \times 1$ vector and \mathbf{m} is a $G \times 1$ vector. Using a result from C. R. Henderson (1975), the BLUP of η under the RILMM given by Equation (A3) for the case of known variance components can be shown to be given by

$$\hat{\eta}^{BLUP}(\sigma_b^2, \sigma_e^2) = \mathbf{k}'\tilde{\boldsymbol{\beta}} + \mathbf{m}'\mathbf{D}\mathbf{U}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}), \quad (\text{A4})$$

where $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})$ is the generalized least square estimator of $\boldsymbol{\beta}$.

Appendix B

R Code to Compute EDI_g^{EBLUP} for a Data Set

```
library(sae)
library(irtoys)
library(ltm)
pr3PL=function(t,a,b,c){return(c + (1-c)/(1 + exp(a*(b-t))))}#Probability under
the 3PL Model
d13PL=function(t,a,b,c){e=exp(a*(t-b)) return(((1-c)*a*e/
((1 + e)*(1 + e)))}#Derivative
#R function to Compute  $EDI_g$ ,  $EDI_g^N$  and  $EDI_g^A$  for a group of examinees
ComputeEDIs=function(scores,erasures,itparm, SchoolInd)
{Nstudents_er=0
EDInumerator=0
EDIdenominator=0
EDIdenominator_A=0
items=1:ncol(scores)
aa=itparm[,1]
bb=itparm[,2]
cc=itparm[,3]
SchoolInds = c()
Mu_gks = c()
Ses = c()
W_sum = 0
Sigma_sum = 0
Nerase_ks = 0
for (k in 1:nrow(scores))
{s=scores[k,]
e=erasures[k,]
```

```

Nerases_k=length(e[e==1])
Nerases_ks=Nerases_ks + Nerases_k
if (Nerases_k<nitem & Nerases_k > 0)
{Nstudents_er=Nstudents_er + 1
W_gk=sum(s[e==1])
ItemE=items[e==1]
ItemNE=items[e==0]
ItemParmNE=as.matrix(itparm[ItemNE,])
NitemNE=nitem-Nerases_k
scoresNE=s[ItemNE]
wl=wle(scoresNE,ItemParmNE) #Compute ability estimate based on nonerased
items
ThetaEst=wl[1]
se=wl[2]
Mu_gk=0
Sigma_gk=0
P_deriv=0
for (i in 1:Nerases_k)
{pr=pr3PL(ThetaEst,aa[ItemE[i]],bb[ItemE[i]],cc[ItemE[i]])
Mu_gk=Mu_gk + pr
Sigma_gk=Sigma_gk + pr*(1-pr)
P_deriv=P_deriv + d13PL(ThetaEst,aa[ItemE[i]],bb[ItemE[i]],cc[ItemE[i]])}
Mu_gks = c(Mu_gks, Mu_gk)
SchoolInds = c(SchoolInds, SchoolInd)
Ses = c(Ses, se)
W_sum = W_sum + W_gk
Sigma_sum = Sigma_sum + Sigma_gk
EDInumerator=EDInumerator + W_gk - Mu_gk
EDIdenominator=EDIdenominator + Sigma_gk
EDIdenominator_A=EDIdenominator_A + Sigma_gk + (se^2)*P_deriv*
P_deriv } }
if(Nerases_ks==0 | Nstudents_er==nrow(scores))
{W_sum=c()
Nstudents_er=c()
Sigma_sum=c()
SchoolInds=c()
Mu_gks=c()
Ses=c()}
EDI_g = (EDInumerator-0.5)/sqrt(EDIdenominator)
EDI_N = (EDInumerator)/sqrt(EDIdenominator)
EDI_A = (EDInumerator)/sqrt(EDIdenominator_A)

```

```

return(list(EDI_g = EDI_g,EDI_N = EDI_N,EDI_A = EDI_A, Mu_gks = Mu_gks,
SchoolInds = SchoolInds, W_sum = W_su Sigma_sum = Sigma_sum, Nstudents_er
= Nstudents_er, Nerase_ks=Nerase_ks, Ses=Ses))}
#####
# R function to Compute  $EDI_g^{EBLUP}$  for all examinee groups
#####
computeEDI_eblup=function(scores,erasures,itparm,SchoolIDs)
{uSchoolInds =unique(SchoolIDs)
SchoolInds = c()
Mu_gks = c()
N_students = c()
Sigma_sums = c()
W_sums = c()
Ses = c()
Nstudents_er = c()
Nerase_ks = c()
Thetas = c()
Schools_EDIs=NULL

for (i in uSchoolInds)
{ sc=scores[schoolIDs==i,]
er=erasures[schoolIDs==i,]
N = nrow(sc)
results = ComputeEDIs(sc,er,itparm, i)
N_students = c(N_students, N)
EDIs=c(results$EDI_g, results$EDI_N, results$EDI_A)
Mu_gks = c(Mu_gks, results$Mu_gks)
Ses = c(Ses, results$Ses)
SchoolInds = c(SchoolInds, results$SchoolInds)
Nstudents_er = c(Nstudents_er, results$Nstudents_er)
Nerase_ks = c(Nerase_ks, results$Nerase_ks)
W_sums = c(W_sums, results$W_sum)
Sigma_sums = c(Sigma_sums, results$Sigma_sum)
Schools_EDIs =rbind(Schools_EDIs, EDIs)}

invalidschools = uSchoolInds [!uSchoolInds %in%unique(SchoolInds)]
validschools = uSchoolInds [!uSchoolInds %in% invalidschools]
Xmean <- data.frame(validschools)
Popn <- data.frame(validschools, N_students[!uSchoolInds
%in% invalidschools])
ediData_school = data.frame(Mu_gks, SchoolInds)
eblupResults = eblupBHF_uneqlvar(Mu_gks ~ 1, dom=SchoolInds,
meanxpop=matrix(Xmean[,1],ncol = 1), weights=Ses^2, popnsize=Popn,
data=ediData_school)

```

```

mseResults = pbmseBHF(Mu_gks ~ 1, dom=SchoolInds,
meanxpop=matrix(Xmean[,1],ncol = 1), popnsize=Popn, B=50,
data=ediData_school)
EDINumerator_eblup = W_sums - Nstudents_er*eblupResults
EDIDenominator_eblup = Sigma_sums + Nstudents_er*(mseResults$mse$mse)
EDI_eblup = EDINumerator_eblup/sqrt(EDIDenominator_eblup)
Output = cbind(uSchoolInds,Schools_EDIs, EDI_eblup)
return(Output)}

# Read the data file that includes the group indicators (1 column), item scores
(I columns, #I=Number of items on the test), and erasure indicators
(I columns)
data=read.csv("~/path/to/the/file.csv")
# Read the item parameters from a file
itparm=matrix(scan("~/path/to/the/item_parameter_file.txt"),ncol = 3,byrow=T)
schoolIDs = data$SchoolIDs #Define the School indicators
scores=data[,grepl("Score",colnames(data))] #Define the item scores
erasures=data[,grepl("Erasure",colnames(data))] #Define the erasure indicators
# Call the R function to compute  $EDI_g^{EBLUP}$  for all examinee groups. The output
of the function
# is a matrix with 5 columns—School ID followed by 3 existing EDI's and the
new EDI
Results = computeEDI_eblup(scores,erasures,itparm,schoolIDs)

```

Authors' Note

The author Luyao Peng is currently an employee at ByteDance. The study was conducted by the authors in 2020 while Luyao Peng was a visiting researcher at Beijing Language and Culture University. The final article was accepted in January 2021.


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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The second author was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant No. R305D170026.

ORCID iD

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

Notes

1. Typically, at least one erasure is found for a large percentage of takers of educational tests. For example, Liu et al. (2015) found at least one erasure for 98.4% and 95.8% of examinees on GRE Verbal and Quantitative, respectively, in two large samples of GRE test takers.
2. For example, EDI_g is an estimate of $\frac{\sum_{k=1}^{n_g} (X_{gk} - \mu_{gk}) - 0.5}{\sqrt{\sum_{k=1}^{n_g} \sigma_{gk}^2}}$.
3. Because $\hat{\theta}_{gk}$ is computed from the nonerased items and the number of nonerased items is large for almost all examinees, this is a reasonable assumption although future research could explore setting the w_{gk} 's equal to other values.
4. The simplifications are facilitated by the standard result (see, e.g., H. Henderson & Searle, 1981) regarding the inversion of the matrix $(A + \mathbf{u}\mathbf{v}')^{-1}$ and the observation that in our case, \mathbf{X} , $\boldsymbol{\beta}$, \mathbf{k} , \mathbf{m} , \mathbf{D} , \mathbf{U} , and \mathbf{V} of Equation (A4) are equal to an $n \times 1$ vector of ones, $\boldsymbol{\mu}$, $\mathbf{1}$, a $G \times 1$ vector with 1 in position g and 0 in all other positions, $\sigma_b^2 \mathbf{I}_G$, the matrix \mathbf{U} given in Equation (7), and $\sigma_e^2 \mathbf{U}\mathbf{U}' + \sigma_e^2 \mathbf{W}$, respectively.
5. The use of the estimated item parameters from the artificial data set does not affect the comparative performance of the indices.
6. The numbers of schools and districts without any erasures are 53 and 21, respectively. The numbers of schools and districts where all examinees made erasures are 194 and 79, respectively.
7. Plots of EDI_g^{EBLUP} versus EDI_g^A show similar patterns as Figure 3 and are not included.

References

- Agresti, A. (2015). *Foundations of linear and generalized linear models*. Wiley.
- American Educational Research Association, American Psychological Association, and National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Bishop, N. S., Liassou, D., Bulut, O., Seo, D. G., & Bishop, K. (2011, April 7-8). *Paper three: Modeling erasure behavior* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1), 29-51. <https://doi.org/10.1007/BF02291411>
- Braun, H. I. (1989). Empirical Bayes methods: A tool for exploratory analysis. In R. Darrell Bock (Ed.), *Multilevel analysis of educational data* (pp. 19-55). Academic Press. <https://doi.org/10.1016/B978-0-12-108840-8.50006-8>
- Cizek, G. J., & Wollack, J. A. (2017). *Handbook of detecting cheating on tests*. Routledge.
- Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, 23(3), 315-345. <https://doi.org/10.1214/aoms/1177729380>
- Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3), 589-599. <https://doi.org/10.1093/biomet/68.3.589>
- Foster, D. (2013). Security issues in technology-based testing. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 299-311). Routledge.

- Ghosh, M., & Meeden, G. (1986). Empirical Bayes estimation in finite population sampling. *Journal of the American Statistical Association*, *81*(396), 1058-1062. <https://doi.org/10.1080/01621459.1986.10478373>
- González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, *78*(5), 443-462. <https://doi.org/10.1080/00949650601141811>
- Harville, D. A. (1991). That BLUP is a good thing: The estimation of random effects [Comment]. *Statistical Science*, *6*(1), 35-39. <https://doi.org/10.1214/ss/1177011928>
- Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, *31*(2), 423-447. <https://doi.org/10.2307/2529430>
- Henderson, H., & Searle, S. (1981). On deriving the inverse of a sum of matrices. *SIAM Review*, *23*(1), 53-60. <https://doi.org/10.1137/1023004>
- Jiang, J., & Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, *15*, Article 1. <https://doi.org/10.1007/BF02595419>
- Liu, O. L., Bridgeman, B., Gu, L., Xu, J., & Kong, N. (2015). Investigation of response changes in the GRE revised general test. *Educational and Psychological Measurement*, *75*(6), 1002-1020. <https://doi.org/10.1177/0013164415573988>
- 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. 187-214). Routledge.
- McClintock, J. C. (2015). Erasure analyses: Reducing the number of false positives. *Applied Measurement in Education*, *28*(1), 14-32. <https://doi.org/10.1080/08957347.2014.973563>
- McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models*. John Wiley.
- Molina, I., & Marhuenda, Y. (2015). Sae: An R package for small area estimation. *R Journal*, *7*(1), 81-98. <https://doi.org/10.32614/RJ-2015-007>
- Partchev, I., Maris, G., & Hattori, T. (2017). *irtoys: A collection of functions related to item response theory (IRT)*. R package Version 0.2.1. R Package Documentation. <https://rdrr.io/cran/irtoys/>
- Prasad, N. N., & Rao, J. N. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association*, *85*(409), 163-171. <https://doi.org/10.1080/01621459.1990.10475320>
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Sinharay, S. (2017). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, *42*(1), 46-68. <https://doi.org/10.3102/1076998616673872>
- Sinharay, S. (2018). Detecting fraudulent erasures at an aggregate level. *Journal of Educational and Behavioral Statistics*, *43*(3), 286-315. <https://doi.org/10.3102/1076998617739626>
- Sinharay, S., Duong, M. Q., & Wood, S. W. (2017). A new statistic for detection of aberrant answer changes. *Journal of Educational Measurement*, *54*(2), 60-217. <https://doi.org/10.1111/jedm.12141>
- Sinharay, S., & Jensen, J. L. (2019). Higher-order asymptotics and its use to test the equality of the examinee ability over two sets of items. *Psychometrika*, *84*(2), 484-510. <https://doi.org/10.1007/s11336-018-9627-8>

- Sinharay, S., & Johnson, M. S. (2017). Three new methods for analysis of answer changes. *Educational and Psychological Measurement, 77*(1), 54-81. <https://doi.org/10.1177/0013164416632287>
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450. <https://doi.org/10.1007/BF02294627>
- Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement, 21*(4), 307-320. <https://doi.org/10.1177/01466216970214002>
- Wollack, J. A., Cohen, A. S., & Eckerly, C. A. (2015). Detecting test tampering using item response theory. *Educational and Psychological Measurement, 75*(6), 931-953. <https://doi.org/10.1177/0013164414568716>
- 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). Routledge. <https://doi.org/10.4324/9781315743097-11>
- Wollack, J. A., & Schoenig, R. W. (2018). Cheating. In B. B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation* (pp. 260-265). Sage.
- Zwick, R., Thayer, D., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1-28. <https://doi.org/10.1111/j.1745-3984.1999.tb00543.x>