Classification Performance of Answer-Copying Indices Under Different Types of IRT Models

Applied Psychological Measurement
2016, Vol. 40(8) 592–607
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621616664724
apm.sagepub.com

Cengiz Zopluoglu¹

Abstract

Test fraud has recently received increased attention in the field of educational testing, and the use of comprehensive integrity analysis after test administration is recommended for investigating different types of potential test frauds. One type of test fraud involves answer copying between two examinees, and numerous statistical methods have been proposed in the literature to screen and identify unusual response similarity or irregular response patterns on multiple-choice tests. The current study examined the classification performance of answer-copying indices measured by the area under the receiver operating characteristic (ROC) curve under different item response theory (IRT) models (one- [1PL], two- [2PL], three-parameter [3PL] models, nominal response model [NRM]) using both simulated and real response vectors. The results indicated that although there is a slight increase in the performance for low amount of copying conditions (20%), when nominal response outcomes were used, these indices performed in a similar manner for 40% and 60% copying conditions when dichotomous response outcomes were utilized. The results also indicated that the performance with simulated response vectors was almost identically reproducible with real response vectors.

Keywords

response similarity, answer copying, test security, test score integrity, person fit, item response theory

Test fraud has recently received increased attention in the field of educational testing. This is due in part to the increased stakes being associated with tests used to measure student and teacher performance, and screening for a possible contamination of item response data by test fraud is essential for maintaining test score integrity and validity. As a result of increasing concerns regarding the integrity of test scores, some states are pushing bills that require state departments to regularly look for potential fraud in standardized tests (Crouch, 2012), while some others are creating test security units within their departments of education (Hildebrand, 2012). A handbook on test and data integrity (National Council on Measurement in Education [NCME], 2012) and a symposium on test integrity (U.S. Department of Education, Institute of

¹University of Miami, FL, USA

Corresponding Author:

Cengiz Zopluoglu, Department of Educational and Psychological Studies, University of Miami, Max Orovitz Building, 304-A, 1570 Levante Avenue, Coral Gables, FL 33146, USA.
Email: c.zopluoglu@miami.edu

Education Sciences [IES], & National Center for Education Statistics [NCES], 2013) both recommended the use of comprehensive integrity analysis after test administration for investigating different types of potential test frauds in academic testing.

Different types of test frauds on examinations have been discussed (Cizek, 1999), and different irregularities may occur in testing data such as suspicious changes in test scores in adjoining years, suspicious erasure behavior, or unusual answer similarity due to different types of frauds. One type of test fraud involves answer copying between two examinees and may occur in many ways such as by glancing at a nearby student's work or by sending text messages and photos of answers (e.g., Green, 2014). Many self-report surveys indicate that answer copying is a type of test fraud that is all too common (Bopp, Gleason, & Misicka, 2001; Brimble & Clarke, 2005; Hughes & McCabe, 2006; Jensen, Arnett, Feldman, & Cauffman, 2002; Josephson Institute of Ethics, 2006, 2008, 2010, 2012; Lin & Wen, 2007; McCabe, 2001; Rakovski & Levy, 2007; Vandehey, Diekhoff, & LaBeff, 2007; Whitley, 1998). Answer copying may result in unusual response patterns for a copier examinee and/or unusual answer similarity between two response vectors, which can be investigated through probabilistic models.

Numerous statistical methods have been proposed in the literature to screen and identify unusual response similarity or irregular response patterns on multiple-choice tests, and interest in such methods can be traced back to the late 1920s (Bird, 1927). These methods can be classified into two main categories, as they either look at the likelihood of similarity between two response vectors (Angoff, 1972; Anikeeff, 1954; Bay, 1995; Bellezza & Bellezza, 1989; Belov, 2011; Cody, 1985; Dickenson, 1945; Frary, Tideman, & Watts, 1977; Hanson, Harris, & Brennan, 1987; Holland, 1996; Saupe, 1960; Sotaridona & Meijer, 2002, 2003; van der Linden & Sotaridona, 2006; Wollack, 1997) or look at the consistency of a single response vector with a response model (Drasgow, Levine, & Williams, 1985; Harnisch & Linn, 1981; Karabatsos, 2003; Sato, 1975; Sijtsma & Meijer, 1992; Tatsuoka, 1984; Trabin & Weiss, 1983; van der Flier, 1982). Recently, Belov and Armstrong (2010) proposed a two-stage procedure and combined the features of both approaches. This procedure first uses a person fit index from the second category to identify potential copiers, and then uses a similarity index from the first category to identify unusual similarity between the potential copiers and their neighbor examinees.

Some of these indices are similar to each other as many of them were proposed as an improvement or alternative to a previous one. For instance, ω utilizes the same idea as g_2 but uses an item response theory (IRT) model rather than classical test theory (CTT)-based sample statistics to compute the probability of selecting the k th response alternative of the j th item for the i th examinee i (p_{ijk}). Similarly, generalized binomial test (GBT) is identical to B_M (Bay, 1995) in terms of computational procedure but estimates p_{ijk} by using a known response model rather than using sample proportion-correct for items. The K-index and its variants (K_1 , K_2 , S_1 , S_2) are similar to each other with slight differences based on the type of evidence and statistical distribution used. For instance, K , K_1 , K_2 , and S_1 only use identical incorrect responses as evidence of copying, while S_2 takes both identical correct and identical incorrect responses into account. Also, K , K_1 , and K_2 use a binomial distribution, whereas S_1 and S_2 use a Poisson distribution.

A practical concern for these indices is their statistical power and Type I error rates, and simulation studies provide an efficient and flexible way for studying the statistical properties of these indices under different conditions as answer copying cannot typically be observed. A few simulation studies examined and compared the performance of some of these indices under different conditions in terms of their empirical power and Type I error rates (Bay, 1995; Belov, 2011; Hanson et al., 1987; Karabatsos, 2003; Sotaridona & Meijer, 2002, 2003; Wollack, 1997, 2003, 2006; Wollack & Cohen, 1998; Zopluoglu, Chen, Yu, & Mroch, 2013; Zopluoglu &

Davenport, 2012). In general, the IRT-based ω and GBT indices and their non-IRT counterparts, K , K_1 , K_2 , S_1 , S_2 , and the Variable Match Index are found to be providing reasonable detection rates while holding the Type I error rates at or below the nominal level. Most other indices either inflated the Type I error rate, which has serious practical consequences, or had very low statistical power. The exception was the indices based on empirical null distributions that did not receive as much attention as other methods in the literature, which is likely due to the practical limitations of accessing large databases for creating empirical null distributions for the number of matching correct and incorrect responses.

The standard application of these indices in the literature involves using response vectors with nominal response outcomes. In previous simulation studies regarding the empirical power and Type I error rates of the answer-copying indices, examinees' responses were mostly assumed to follow the nominal response model (NRM; Bock, 1972) for generating data, and nominal response outcomes were also used in implementation, but these indices can also be accommodated for use with dichotomous response vectors and dichotomous IRT models. Although these indices are expected to perform optimally due to the increased amount of information when nominal responses are utilized, researchers and practitioners may have different motivations to use dichotomous responses and dichotomous IRT models (Birnbaum, 1968) in practical applications. For instance, the test questions may be open ended, not include distractors, and be naturally dichotomously scored (e.g., correct/incorrect, true/false), or researchers and practitioners may believe that guessing is an important component of the multiple-choice test under investigation, and be faced with a dilemma whether to use nominal responses with more information but a model with no guessing component or to use dichotomous responses with less information but a model with guessing component. (e.g., three-parameter [3PL] model). Apparently, there is not enough information in the literature regarding the utility of these indices when used with dichotomous scores and dichotomous IRT models.

In addition, simulated data with a well-defined unidimensional IRT model are not contaminated with "nuisance" factors that are likely to exist in real data. In real data, there may be factors that cause systematic variation in item response data other than the target latent trait intended to be measured. For instance, other traits that items do not intend to measure (e.g., reading component in mathematics items) or passage effect in item bundles for a reading test are likely to lead some systematic variance not accounted for by a unidimensional model. Also, personal factors such as motivation and test fatigue in longer tests or environmental factors such as testing time and condition may lead unaccounted systematic variance in item response data. In such cases, the item and person parameter estimates of a unidimensional model may be inaccurate and may indirectly influence the results of indices that make use of those model parameters. The current literature is lacking on utilizing real response vectors in investigating the performance of these methods, and more research with real data is certainly needed assuming that these indices will be potentially used as a part of comprehensive integrity analysis in an era of high-stakes testing. Therefore, there are two main research questions to be addressed in the current article:

Research Question 1: To what extent are answer-copying indices effective in detecting answer copying when used with dichotomous response vectors and dichotomous IRT models as opposed to be used with nominal response vectors and NRM?

Research Question 2: To what extent are the results robust when response vectors from a real dataset with likely nuisance factors are used?

Method

Two separate Monte Carlo simulations were designed to address the two research questions outlined above. First, NRM and dichotomous IRT models (1PL, 2PL, and 3PL) were fitted to a real dataset, and item parameters were estimated for each model. To address the first question, performance of answer-copying indices was examined under different conditions using simulated data based on four different IRT models with no nuisance factors. Then, a very similar design was replicated using the response vectors from the real data in which item parameters were obtained, and results were compared with those obtained from simulated data with similar item and test characteristics to address the second research question.

Sample and Real Dataset

The real dataset in the current study came from a census administration of a graduation-required assessment taken by 67,896 eighth-grade students in mathematics in a large Midwestern state. The datasets comprised of common items administered to all students in mathematics. The portion of the mathematics test used in the current study included 40 items from four content strands (Whole Number, Percentage and Ratio, Measurement, and Table and Graphs). Each item had four options. The average number-correct score was 32.1 (80% correct) with a standard deviation of 6.7. The number-correct scores were negatively skewed with a skewness value of -1.13 and a kurtosis value of 0.65 , so it was a relatively easy test with most students scoring high on the test. The item difficulty ranged from 0.56 to 0.97 with an average of 0.79 and a standard deviation of 0.11 . The point-biserial correlation ranged from $.23$ to $.61$ with an average of 0.46 and a standard deviation of 0.09 .

NRM was fitted to the dataset with nominal responses, and dichotomous IRT models were fitted to the corresponding dichotomous responses. Dichotomous IRT models were fitted using IRTPRO (Cai, du Toit, & Thissen, 2011) with default options. The only exception was to increase the maximum number of E-steps to 10,000 to ensure convergence and to specify a prior distribution, $\beta \sim (5, 17)$, for the guessing parameter when fitting the 3PL model. NRM was fitted using Multilog (Thissen, Chen, & Bock, 2003) with default options, except that the maximum number of E-steps was increased to 10,000 to ensure convergence. Estimated item parameters are reported in Online Appendix Tables A1 and A2, and the averages and standard deviations of item parameters for the dichotomous IRT models along with model fit information are shown in Table 1. Among the dichotomous IRT models, 3PL model provided the best fit based on the Akaike information criterion (AIC) and Bayesian information criterion (BIC) values, indicating a likely guessing component for the items in the test.

Study Design

Simulation 1. The first simulation study was designed to examine the performance of answer-copying indices under nominal and dichotomous IRT models with purely simulated unidimensional data that contain no nuisance factors. Five factors were manipulated: sample size, amount of copying, type of copying, test difficulty level, and generating IRT model. Generating IRT model was manipulated at four levels, and data were simulated under NRM and 1PL, 2PL, 3PL models. Two levels of test difficulty were considered in the current study as easy and medium, and overall test difficulty was manipulated through the b parameters for the dichotomous IRT models and location parameters for NRM. The item parameters derived from the real dataset and reported in Online Appendix Tables A1 and A2 served to generate data for easy test conditions under different IRT models. For easy test difficulty conditions, test difficulty was around

Table 1. Summary Statistics From Fitting Dichotomous IRT Models to Real Dataset.

	A		B		C		AIC	BIC
	M	SD	M	SD	M	SD		
1PL	—	—	-.52	0.70	—	—	2,124,201	2,124,575
2PL	1.41	0.43	-1.56	0.81	—	—	2,100,531	2,101,259
3PL	1.69	0.80	-1.28	0.95	0.16	0.11	2,091,895	2,092,988

Note. IRT = item response theory; AIC = Akaike information criterion; BIC = Bayesian information criterion; 1PL = one-parameter model; 2PL = two-parameter model; 3PL = three-parameter model.

.80 for typical simulated response data which was similar to the real dataset. For medium test difficulty conditions, the *b* parameters for the dichotomous IRT models and location parameters for the NRM were manipulated, such that test difficulty was around .50 for typical simulated response data. This was accomplished by adding a constant of 1.52 for the 1PL, 1.56 for the 2PL, and 1.53 for the 3PL model to the *b* parameters used for the easy test conditions. Similarly, a constant of 1.5 was subtracted from the location parameter for the correct option, and a constant of 0.5 (to preserve the sum-to-zero property of the NRM location parameters) was added to the location parameters for the distractors for each item to maintain a test difficulty level of .5 for the simulated data based on NRM. While manipulating the *b* and location parameters, all other parameters, discrimination and/or guessing, were kept same. The item parameters for the medium test difficulty conditions are reported in Online Appendix Tables A3 and A4. Sample size was manipulated at three levels as 100, 500, and 2,500, and amount of copying was manipulated at three levels as 20% (low), 40% (medium), and 60% (high). Two types of copying were considered: random copying and difficulty weighted copying. In random copying condition, copier examinee was assumed to copy responses from source examinee in a random fashion, so each item was equally likely to be chosen for copying. In difficulty weighted copying, items were ranked from the easiest to the most difficult, and the probability of each item being copied was proportional to its rank (Wollack, 1997). For instance, the probability of being copied was about .001 and .049 for the easiest and most difficult items, respectively, in a 40-item test. All factors were fully crossed, yielding a total of 144 conditions ($4 \times 2 \times 3 \times 3 \times 2$).

The following steps were followed for one replication in the simulation study:

1. Simulating data: Item responses for a sample of *n* (100, 500, and 2,500) students were generated using an IRT model (1PL, 2PL, 3PL, and NRM) and corresponding item parameters are reported in Online Appendix.
2. Obtaining answer-copying pairs: Four percent of the students in the sample were randomly chosen as copiers. For each copier, another student within the sample was randomly selected and matched with the copier student. A certain number of items (20%, 40%, and 60%) was selected for each pair (by random or difficulty weighted sampling), and the source student's response vector was overwritten on the copier student's response vector for the selected items to mimic answer copying.
3. Obtaining honest pairs: After excluding the students in the answer-copying pairs obtained in the second step, 8% of the students were again randomly sampled. For each selected student, another student within the sample was randomly selected and matched. No answer copying was simulated.
4. Computing answer-copying indices: Item parameters and ability parameters were estimated under the same model used to generate data at Step 1. Then, answer-copying

Figure 1. The distribution of p values for a sample of answer-copying and honest pairs.

indices (ω and GBT, K , K_1 , K_2 , S_1 , and S_2) were computed for each answer copying and honest pair obtained in Steps 2 and 3 using a R package CopyDetect (R Core Team, 2014; Zopluoglu, 2013).

Simulation 2. A second simulation was designed to respond to the second research question, and the results of the first simulation served as a benchmark to evaluate the results of the second simulation. In the second simulation, all four steps outlined above were similarly implemented. The only difference was that the original real dataset served as population to draw response vectors in Step 1. Sample size, amount of copying, and type of copying were kept same and manipulated at the same levels. Test difficulty was not considered as the average item difficulty was .79 for the real dataset and matched the easy test condition from the first simulation. Also, there were only two IRT models considered for the second simulation. NRM and best-fitting dichotomous IRT model (3PL) were used to analyze the real response vectors and to compute similarity indices with nominal and corresponding dichotomous response outcomes.

Outcome of interest. Current practice in performance of answer-copying indices is to report empirical Type I and Type II error rates for a given alpha level (e.g., .001, .01, and .05). The challenge of this practice is that not all indices operate at the same empirical Type I error rate. While some indices are more conservative and their empirical Type I error rates are much below the nominal alpha level, some indices are relatively more liberal and operate at a higher empirical Type I error rate. Therefore, it is difficult to make comparisons across indices as well as across conditions for the same index. As a possible solution to this challenge, current study used the area under the receiver operating characteristic (ROC) curve (AUC) as a performance measure which was previously used by Karabatsos (2003) for comparing person fit indices. To illustrate the idea, Figure 1 shows the distribution of p values computed for a sample of true answer-copying pairs and honest pairs using an answer-copying index. If a nominal alpha level of .01 was used, the area under the distribution on the left side of $\alpha = .01$ line for the honest pairs of examinees would be the empirical Type I error rate and equal to .0042 (proportion of falsely detected honest pairs), while the area under the distribution on the left side of $\alpha = .01$ line for

Figure 2. Receiver operating characteristic curve for the illustration in Figure 1.

answer-copying pairs would be the empirical power and equal to .2648 (proportion of truly detected copying pairs).

If a nominal alpha level of .05 was used, the corresponding empirical Type I error rate and power would be equal to 0.0306 and 0.4954, respectively. Suppose that the empirical Type I error rates and power for all possible nominal alpha levels are computed in a similar fashion and plotted against each other. The resulting graph is a ROC curve as shown in Figure 2, and the AUC is used as a measure of classification accuracy for how well an index separates the answer-copying and honest pairs of students. While an area of 1 indicates a perfect classification, an area of .5 indicates a classification performance not beyond the chance success. Roughly, an area of .5 to .7, .7 to .8, .8 to .9, and .9 to 1 would indicate poor, fair, good, and excellent classification performance. For instance, the AUC in Figure 2 was equal to .89 indicating a good classification performance.

Number of replications. The number of replications was chosen such that the maximum possible standard error of AUC estimate is equal to .005. The standard error of AUC is known to be equal to (Hanley & McNeil, 1982)

$$\sigma_{\text{AUC}} = \sqrt{\frac{A(1-A) + (m-1)\left(\frac{A}{2-A} - A^2\right) + (n-1)\left(\frac{2A^2}{1+A} - A^2\right)}{mn}},$$

where A represents the estimated AUC, m can be considered as the number of answer-copying pairs, and n can be considered as the number of honest pairs used to estimate AUC in this context. The standard error of an AUC estimate reaches its maximum when the AUC estimate is equal to .5 for the fixed levels of m and n . With a number of 5,000 answer-copying pairs and 10,000 honest pairs, the standard error for an AUC estimate of .5 is equal to 0.005. Therefore, the number of replications was chosen, such that the number of answer-copying pairs and the number of honest pairs analyzed were equal to 5,000 and 10,000, respectively, for each

condition (50, 250, and 1,250 replications for conditions with a sample size of 2,500, 500, and 100, respectively).

Results

The AUC estimates for each index under all conditions for the first simulation are presented in Table 2 through Table 5. To better capture the important effects on these tables, an ANOVA with only main effects for five between-subjects factors (generating IRT model, amount of copying, sample size, test difficulty, type of copying) and one within-subjects factor (answer-copying index) was run. All main effects were statistically significant at the level of .001, and the model with only main effects accounted for 94.7% of the variance in the AUC estimates. The effect sizes as measured by η^2 were 0.850, 0.036, 0.037, 0.005, 0.013, and 0.006 for amount of copying, generating IRT model, test difficulty, sample size, type of copying, and type of index, respectively. While the amount of copying was the most important factor that affects the classification performance, the type of IRT model and the test difficulty level had small effects, and the sample size, type of copying, and type of answer-copying index had negligible effects. The main effects for amount of copying, test difficulty, and generating IRT model were further analyzed. Post hoc comparisons revealed that the average AUC estimate was equal to .960 when the amount of copying was 60% and significantly higher than the average AUC estimate when the amount of copying was 40%, $\Delta_{AUC} = .048$, $p < .001$, and when the amount of copying was 20%, $\Delta_{AUC} = .183$, $p < .001$. The average AUC estimate was equal to .899 when the test difficulty level was medium and significantly higher than the average AUC estimate when the test was easy, $\Delta_{AUC} = .033$, $p < .001$. The average AUC estimate was equal to .910 when the generating IRT model was NRM and significantly higher than the average AUC estimate when the generating IRT model was 3PL, $\Delta_{AUC} = .037$, $p < .001$; when the generating model is 2PL, $\Delta_{AUC} = .039$, $p < .001$; and when the generating model is 1PL, $\Delta_{AUC} = .034$, $p < .001$. There were no significant differences in average AUC estimates among dichotomous IRT models.

The main interest for the current article was to investigate to what extent answer-copying indices were effective when used with dichotomous response vectors and dichotomous IRT models as opposed to be used with nominal response vectors and NRM. As expected, there was a reduction in classification performance as measured by AUC due to the information loss when dichotomous response outcomes were used in computing answer-copying indices. However, the loss in performance was small, and the average AUC estimate was around .76, .90, and .96 when the amount of copying is 20%, 40%, and 60%, respectively, when the dichotomous IRT models were used as opposed to .82, .94, and .97 when NRM was used. The largest difference occurred for low amount of copying conditions where using more information provided some advantages to NRM, but the difference became negligible as the amount of copying increased.

To respond to the second research question, a very similar design was replicated using a real dataset, and the results from the second simulation are presented in Table 6. The results at the upper portion of Tables 4 and 5 can serve as a benchmark for the results in Table 6 to compare the classification performance of answer-copying indices under different conditions for simulated and real datasets. Initial exploration of the results indicated that the classification performance of answer-copying indices was very similar across different conditions under simulated and real datasets. To test and confirm the initial exploration, another ANOVA was run with five between-subjects factors (generating IRT model, amount of copying, sample size, type of copying, and type of dataset) and one within-subjects factor (answer-copying index) using the data only from Table 6 and the upper portions of Tables 4 and 5. The ANOVA indicated that there was not a significant effect of the type of dataset (simulated vs. real) on the classification

Table 2. Area Under the ROC Curve When Generating IRT Model is IPL.

Sample size	Amount of copying	Type of copying	ω	GBT	K	K ₁	K ₂	S ₁	S ₂
Easy test difficulty (0.80)									
100	20	Random	0.745	0.741	0.693	0.727	0.718	0.718	0.744
100	20	Difficulty W.	0.793	0.793	0.734	0.779	0.768	0.774	0.800
100	40	Random	0.881	0.881	0.820	0.869	0.864	0.863	0.879
100	40	Difficulty W.	0.917	0.914	0.863	0.907	0.904	0.903	0.914
100	60	Random	0.943	0.940	0.899	0.937	0.936	0.931	0.936
100	60	Difficulty W.	0.966	0.964	0.931	0.962	0.963	0.959	0.958
500	20	Random	0.763	0.758	0.737	0.748	0.742	0.730	0.758
500	20	Difficulty W.	0.804	0.798	0.777	0.789	0.782	0.770	0.801
500	40	Random	0.887	0.885	0.869	0.876	0.875	0.865	0.881
500	40	Difficulty W.	0.923	0.918	0.911	0.917	0.915	0.907	0.919
500	60	Random	0.945	0.940	0.937	0.938	0.941	0.930	0.937
500	60	Difficulty W.	0.964	0.959	0.959	0.960	0.962	0.955	0.958
2,500	20	Random	0.754	0.752	0.731	0.738	0.734	0.715	0.742
2,500	20	Difficulty W.	0.794	0.787	0.772	0.778	0.773	0.754	0.785
2,500	40	Random	0.888	0.886	0.878	0.880	0.879	0.863	0.880
2,500	40	Difficulty W.	0.924	0.920	0.915	0.917	0.917	0.905	0.918
2,500	60	Random	0.949	0.946	0.944	0.942	0.945	0.933	0.940
2,500	60	Difficulty W.	0.969	0.965	0.969	0.967	0.969	0.961	0.961
Medium test difficulty (0.50)									
100	20	Random	0.776	0.772	0.723	0.781	0.773	0.726	0.759
100	20	Difficulty W.	0.786	0.785	0.727	0.789	0.783	0.740	0.769
100	40	Random	0.932	0.932	0.868	0.929	0.928	0.895	0.913
100	40	Difficulty W.	0.933	0.934	0.868	0.928	0.928	0.909	0.920
100	60	Random	0.979	0.980	0.921	0.973	0.975	0.957	0.961
100	60	Difficulty W.	0.978	0.978	0.920	0.971	0.974	0.958	0.960
500	20	Random	0.777	0.771	0.766	0.782	0.775	0.734	0.759
500	20	Difficulty W.	0.781	0.776	0.777	0.789	0.784	0.757	0.781
500	40	Random	0.927	0.927	0.917	0.924	0.924	0.902	0.911
500	40	Difficulty W.	0.932	0.930	0.924	0.931	0.932	0.916	0.924
500	60	Random	0.978	0.978	0.972	0.973	0.976	0.961	0.962
500	60	Difficulty W.	0.976	0.975	0.971	0.970	0.974	0.959	0.960
2,500	20	Random	0.778	0.772	0.776	0.784	0.778	0.740	0.766
2,500	20	Difficulty W.	0.781	0.771	0.782	0.784	0.783	0.751	0.776
2,500	40	Random	0.929	0.927	0.927	0.928	0.928	0.905	0.915
2,500	40	Difficulty W.	0.933	0.933	0.932	0.933	0.933	0.918	0.928
2,500	60	Random	0.975	0.974	0.973	0.971	0.974	0.958	0.960
2,500	60	Difficulty W.	0.974	0.974	0.971	0.969	0.972	0.959	0.960

Note. ROC = receiver operating characteristic; IRT = item response theory; IPL = one-parameter model; GBT = generalized binomial test; Difficulty W. = difficulty weighted.

performance of answer-copying indices, $F(1, 64) = 0.913$, $p = .343$, after controlling for other factors in the model. The results were reproducible using a real dataset.

Discussion

Given the increasing use of high-stakes testing, test score integrity has become a major concern in educational testing, and identifying unusual answer copying among examinees is a type of integrity analysis recommended in several recent platforms. Fortunately, the technical literature

Table 3. Area Under the ROC Curve When Generating IRT Model is 2PL.

Sample size	Amount of copying	Type of copying	ω	GBT	K	K_1	K_2	S_1	S_2
Easy test difficulty (0.80)									
100	20	Random	0.737	0.728	0.697	0.733	0.724	0.728	0.752
100	20	Difficulty W.	0.774	0.762	0.731	0.768	0.760	0.762	0.789
100	40	Random	0.873	0.868	0.820	0.868	0.864	0.862	0.876
100	40	Difficulty W.	0.909	0.902	0.861	0.904	0.902	0.902	0.913
100	60	Random	0.939	0.931	0.900	0.937	0.936	0.930	0.935
100	60	Difficulty W.	0.956	0.947	0.916	0.953	0.953	0.950	0.953
500	20	Random	0.746	0.739	0.721	0.734	0.727	0.720	0.747
500	20	Difficulty W.	0.802	0.793	0.775	0.789	0.781	0.772	0.802
500	40	Random	0.880	0.874	0.859	0.867	0.865	0.857	0.876
500	40	Difficulty W.	0.919	0.914	0.904	0.913	0.910	0.905	0.918
500	60	Random	0.944	0.939	0.936	0.938	0.940	0.930	0.938
500	60	Difficulty W.	0.965	0.960	0.961	0.962	0.964	0.958	0.960
2,500	20	Random	0.754	0.748	0.732	0.740	0.734	0.716	0.744
2,500	20	Difficulty W.	0.797	0.789	0.772	0.779	0.773	0.756	0.788
2,500	40	Random	0.888	0.883	0.871	0.873	0.872	0.859	0.877
2,500	40	Difficulty W.	0.922	0.915	0.911	0.913	0.912	0.905	0.918
2,500	60	Random	0.949	0.945	0.944	0.942	0.945	0.933	0.939
2,500	60	Difficulty W.	0.966	0.961	0.963	0.961	0.964	0.956	0.959
Medium test difficulty (0.50)									
100	20	Random	0.768	0.763	0.719	0.780	0.774	0.726	0.753
100	20	Difficulty W.	0.765	0.757	0.719	0.775	0.769	0.723	0.753
100	40	Random	0.918	0.920	0.864	0.922	0.920	0.890	0.905
100	40	Difficulty W.	0.918	0.919	0.856	0.921	0.920	0.900	0.909
100	60	Random	0.974	0.975	0.918	0.971	0.972	0.952	0.956
100	60	Difficulty W.	0.971	0.971	0.915	0.965	0.969	0.951	0.950
500	20	Random	0.775	0.767	0.770	0.783	0.778	0.733	0.761
500	20	Difficulty W.	0.772	0.765	0.766	0.779	0.774	0.743	0.769
500	40	Random	0.925	0.924	0.918	0.926	0.926	0.904	0.910
500	40	Difficulty W.	0.918	0.917	0.913	0.919	0.921	0.905	0.914
500	60	Random	0.977	0.978	0.971	0.972	0.976	0.958	0.957
500	60	Difficulty W.	0.973	0.973	0.966	0.967	0.971	0.953	0.953
2,500	20	Random	0.777	0.770	0.780	0.786	0.781	0.737	0.765
2,500	20	Difficulty W.	0.772	0.766	0.778	0.783	0.779	0.741	0.775
2,500	40	Random	0.928	0.925	0.924	0.925	0.926	0.904	0.913
2,500	40	Difficulty W.	0.916	0.914	0.916	0.917	0.919	0.906	0.914
2,500	60	Random	0.974	0.974	0.969	0.968	0.971	0.954	0.955
2,500	60	Difficulty W.	0.970	0.969	0.965	0.963	0.967	0.949	0.949

Note. ROC = receiver operating characteristic; IRT = item response theory; 2PL = two-parameter model; GBT = generalized binomial test; Difficulty W. = difficulty weighted.

has already provided a number of useful tools for identifying pairs of examinees with unusual response similarity, which may potentially arise from answer copying. The performance of answer-copying indices has only been explored with a certain type of response outcome (nominal) and a certain type of IRT model (NRM). In practice, however, the type of items used in testing may have dichotomous outcomes by nature, or the practitioners may choose a dichotomous IRT model for different motivations. In addition, the research on these indices heavily relied on simulated data with no nuisance factors and aligned with a well-specified response model. The real response data used in the current study included four content areas, and these

Table 4. Area Under the ROC Curve When Generating IRT Model is 3PL.

Sample size	Amount of copying	Type of copying	ω	GBT	K	K ₁	K ₂	S ₁	S ₂
Easy test difficulty (0.80)									
100	20	Random	0.723	0.698	0.704	0.738	0.730	0.731	0.753
100	20	Difficulty W.	0.760	0.742	0.737	0.777	0.767	0.775	0.795
100	40	Random	0.846	0.828	0.827	0.869	0.862	0.864	0.878
100	40	Difficulty W.	0.872	0.854	0.862	0.906	0.901	0.902	0.910
100	60	Random	0.909	0.893	0.898	0.936	0.935	0.930	0.934
100	60	Difficulty W.	0.926	0.908	0.929	0.958	0.958	0.955	0.955
500	20	Random	0.730	0.715	0.726	0.736	0.730	0.721	0.744
500	20	Difficulty W.	0.767	0.751	0.765	0.780	0.773	0.767	0.791
500	40	Random	0.856	0.843	0.866	0.874	0.871	0.863	0.877
500	40	Difficulty W.	0.883	0.871	0.906	0.914	0.912	0.905	0.915
500	60	Random	0.917	0.906	0.938	0.941	0.942	0.934	0.937
500	60	Difficulty W.	0.927	0.914	0.956	0.957	0.959	0.952	0.952
2,500	20	Random	0.729	0.714	0.732	0.739	0.733	0.723	0.746
2,500	20	Difficulty W.	0.768	0.754	0.772	0.778	0.774	0.760	0.786
2,500	40	Random	0.856	0.846	0.869	0.872	0.871	0.859	0.875
2,500	40	Difficulty W.	0.877	0.866	0.905	0.907	0.907	0.898	0.909
2,500	60	Random	0.919	0.908	0.945	0.943	0.945	0.935	0.938
2,500	60	Difficulty W.	0.936	0.926	0.964	0.962	0.964	0.957	0.956
Medium test difficulty (0.50)									
100	20	Random	0.754	0.754	0.733	0.777	0.773	0.753	0.754
100	20	Difficulty W.	0.769	0.765	0.748	0.796	0.791	0.779	0.778
100	40	Random	0.922	0.928	0.883	0.937	0.935	0.919	0.912
100	40	Difficulty W.	0.911	0.913	0.886	0.937	0.937	0.924	0.915
100	60	Random	0.976	0.979	0.943	0.984	0.984	0.973	0.964
100	60	Difficulty W.	0.971	0.972	0.949	0.984	0.986	0.975	0.966
500	20	Random	0.767	0.767	0.776	0.787	0.784	0.765	0.766
500	20	Difficulty W.	0.765	0.765	0.780	0.790	0.787	0.773	0.774
500	40	Random	0.925	0.931	0.933	0.942	0.941	0.927	0.920
500	40	Difficulty W.	0.920	0.926	0.940	0.945	0.947	0.934	0.927
500	60	Random	0.978	0.981	0.984	0.986	0.987	0.976	0.969
500	60	Difficulty W.	0.973	0.977	0.984	0.986	0.987	0.977	0.970
2,500	20	Random	0.768	0.769	0.782	0.787	0.784	0.763	0.766
2,500	20	Difficulty W.	0.765	0.769	0.783	0.788	0.785	0.773	0.776
2,500	40	Random	0.922	0.929	0.938	0.941	0.941	0.926	0.920
2,500	40	Difficulty W.	0.923	0.931	0.948	0.948	0.950	0.939	0.933
2,500	60	Random	0.975	0.981	0.985	0.985	0.986	0.977	0.973
2,500	60	Difficulty W.	0.974	0.978	0.985	0.985	0.986	0.977	0.972

Note. ROC = receiver operating characteristic; IRT = item response theory; 3PL = three-parameter model; GBT = generalized binomial test; Difficulty W. = difficulty weighted.

four content areas are likely to yield weaker group dimensions in addition to the general dimension intended to be measured. Due to testing conditions, there may be personal or environmental factors causing some variability in the real dataset. Therefore, the real dataset is expected to have a latent structure slightly deviating from a unidimensional structure and to have some systematic noise that cannot be captured by a well-specified unidimensional model. For such conditions, the literature is lacking on how these indices would perform. The current study aimed to fill this gap by investigating the classification performance of answer-copying indices with different types of IRT models using simulated response vectors as well as real response vectors.

Table 5. Area Under the ROC Curve When Generating IRT Model is NRM.

Sample size	Amount of copying	Type of copying	ω	GBT	K	K_1	K_2	S_1	S_2
Easy test difficulty (0.80)									
100	20	Random	0.772	0.752	0.724	0.756	0.754	0.767	0.770
100	20	Difficulty W.	0.810	0.794	0.758	0.797	0.795	0.806	0.809
100	40	Random	0.894	0.888	0.828	0.864	0.867	0.872	0.874
100	40	Difficulty W.	0.927	0.926	0.867	0.899	0.903	0.905	0.907
100	60	Random	0.948	0.950	0.886	0.916	0.922	0.921	0.922
100	60	Difficulty W.	0.967	0.966	0.915	0.939	0.945	0.941	0.941
500	20	Random	0.788	0.786	0.763	0.765	0.767	0.764	0.782
500	20	Difficulty W.	0.832	0.834	0.809	0.813	0.814	0.814	0.833
500	40	Random	0.917	0.925	0.885	0.886	0.889	0.885	0.895
500	40	Difficulty W.	0.943	0.952	0.923	0.924	0.926	0.923	0.930
500	60	Random	0.963	0.971	0.941	0.939	0.942	0.937	0.940
500	60	Difficulty W.	0.976	0.981	0.962	0.960	0.964	0.958	0.959
2,500	20	Random	0.801	0.805	0.772	0.771	0.772	0.768	0.793
2,500	20	Difficulty W.	0.838	0.846	0.809	0.807	0.809	0.804	0.831
2,500	40	Random	0.917	0.928	0.893	0.890	0.893	0.889	0.900
2,500	40	Difficulty W.	0.942	0.954	0.922	0.918	0.921	0.917	0.927
2,500	60	Random	0.963	0.974	0.945	0.940	0.945	0.940	0.944
2,500	60	Difficulty W.	0.978	0.984	0.965	0.960	0.965	0.959	0.960
Medium test difficulty (0.50)									
100	20	Random	0.860	0.853	0.770	0.827	0.817	0.824	0.826
100	20	Difficulty W.	0.886	0.877	0.804	0.861	0.856	0.855	0.858
100	40	Random	0.977	0.978	0.908	0.960	0.958	0.956	0.956
100	40	Difficulty W.	0.987	0.988	0.928	0.975	0.974	0.971	0.971
100	60	Random	0.995	0.996	0.947	0.987	0.988	0.985	0.984
100	60	Difficulty W.	0.997	0.998	0.962	0.993	0.994	0.992	0.991
500	20	Random	0.872	0.865	0.823	0.838	0.831	0.833	0.841
500	20	Difficulty W.	0.894	0.887	0.857	0.868	0.864	0.862	0.874
500	40	Random	0.980	0.982	0.956	0.962	0.959	0.957	0.957
500	40	Difficulty W.	0.990	0.991	0.977	0.980	0.980	0.976	0.976
500	60	Random	0.997	0.998	0.989	0.990	0.991	0.988	0.986
500	60	Difficulty W.	0.998	0.999	0.995	0.994	0.996	0.994	0.992
2,500	20	Random	0.874	0.867	0.837	0.846	0.838	0.839	0.848
2,500	20	Difficulty W.	0.903	0.896	0.873	0.876	0.873	0.868	0.882
2,500	40	Random	0.979	0.981	0.958	0.961	0.958	0.956	0.956
2,500	40	Difficulty W.	0.990	0.990	0.979	0.980	0.979	0.976	0.977
2,500	60	Random	0.996	0.997	0.991	0.990	0.991	0.988	0.985
2,500	60	Difficulty W.	0.996	0.996	0.993	0.992	0.993	0.991	0.990

Note. ROC = receiver operating characteristic; IRT = item response theory; NRM = nominal response model; GBT = generalized binomial test; Difficulty W. = difficulty weighted.

All answer-copying indices included in the current study showed a very similar classification performance, and there was not a particular index outperformed the others in terms of classification accuracy as measured by the AUC. Using nominal response outcomes and NRM in implementation provided some advantage, particularly when the amount of copying is low. The difference in estimated AUC was about .06 for 20% copying. The difference was very small when the amount of copying is 40% or 60%, and using dichotomous IRT models with dichotomous response outcomes performed almost as well as the nominal response outcomes and NRM in those conditions.

Table 6. Area Under the ROC Curve Under 3PL and NRM for the Real Dataset.

Sample size	Amount of copying	Type of copying	ω	GBT	K	K_1	K_2	S_1	S_2
3PL									
100	20	Random	0.719	0.699	0.690	0.726	0.718	0.720	0.744
100	20	Difficulty W.	0.750	0.729	0.732	0.769	0.759	0.765	0.791
100	40	Random	0.848	0.829	0.822	0.869	0.863	0.865	0.881
100	40	Difficulty W.	0.864	0.844	0.858	0.900	0.896	0.896	0.906
100	60	Random	0.906	0.885	0.897	0.935	0.934	0.929	0.934
100	60	Difficulty W.	0.922	0.904	0.918	0.953	0.953	0.950	0.950
500	20	Random	0.737	0.722	0.736	0.742	0.739	0.728	0.754
500	20	Difficulty W.	0.768	0.750	0.762	0.774	0.767	0.762	0.792
500	40	Random	0.851	0.839	0.861	0.870	0.866	0.862	0.878
500	40	Difficulty W.	0.879	0.865	0.899	0.905	0.904	0.899	0.913
500	60	Random	0.916	0.906	0.939	0.941	0.943	0.935	0.939
500	60	Difficulty W.	0.932	0.920	0.956	0.956	0.958	0.951	0.953
2,500	20	Random	0.726	0.711	0.729	0.733	0.728	0.713	0.739
2,500	20	Difficulty W.	0.767	0.752	0.771	0.778	0.772	0.763	0.790
2,500	40	Random	0.850	0.837	0.865	0.868	0.867	0.855	0.874
2,500	40	Difficulty W.	0.878	0.867	0.906	0.908	0.908	0.900	0.911
2,500	60	Random	0.916	0.909	0.940	0.938	0.941	0.931	0.937
2,500	60	Difficulty W.	0.932	0.922	0.960	0.958	0.960	0.952	0.954
NRM									
100	20	Random	0.763	0.743	0.724	0.755	0.751	0.765	0.767
100	20	Difficulty W.	0.820	0.799	0.767	0.802	0.797	0.813	0.815
100	40	Random	0.887	0.880	0.829	0.860	0.860	0.867	0.868
100	40	Difficulty W.	0.923	0.918	0.863	0.895	0.897	0.901	0.902
100	60	Random	0.945	0.943	0.884	0.916	0.922	0.922	0.922
100	60	Difficulty W.	0.959	0.958	0.909	0.935	0.941	0.937	0.937
500	20	Random	0.790	0.786	0.768	0.772	0.771	0.773	0.791
500	20	Difficulty W.	0.836	0.832	0.809	0.815	0.812	0.816	0.832
500	40	Random	0.912	0.919	0.878	0.881	0.883	0.880	0.892
500	40	Difficulty W.	0.942	0.949	0.914	0.914	0.917	0.912	0.920
500	60	Random	0.958	0.966	0.935	0.934	0.936	0.932	0.935
500	60	Difficulty W.	0.976	0.980	0.958	0.956	0.959	0.954	0.956
2,500	20	Random	0.794	0.793	0.770	0.769	0.770	0.766	0.793
2,500	20	Difficulty W.	0.837	0.838	0.815	0.815	0.815	0.812	0.837
2,500	40	Random	0.912	0.922	0.886	0.884	0.885	0.882	0.897
2,500	40	Difficulty W.	0.946	0.954	0.924	0.921	0.923	0.919	0.930
2,500	60	Random	0.964	0.974	0.945	0.941	0.945	0.940	0.946
2,500	60	Difficulty W.	0.978	0.984	0.967	0.963	0.967	0.962	0.962

Note. ROC = receiver operating characteristic; 3PL = three-parameter model; NRM = nominal response model; GBT = generalized binomial test; Difficulty W. = difficulty weighted.

A potential criticism for the simulation studies regarding the performance of answer-copying indices was the lack of utilizing real response vectors containing likely nuisance factors that may not be captured with a well-defined generating response model. The current study compared the performance of these indices under similar conditions using simulated and real response vectors, and the results showed that the performance was pretty robust when the real response vectors were used. This was surprising and not expected, particularly when nominal response outcomes were used. The fit of 3PL model to the real dataset in the current study indicated that there was a guessing component for these items as shown by the guessing parameter

above .10 for 25 items. Neither the K-index and its variants using the CTT item statistics nor the ω and GBT indices using the NRM item parameters account for guessing success in implementation. It was expected that the performance of these indices would be reduced to some degree when used with real response vectors due to the lack of modeling the guessing component in items. However, the results were almost identically reproducible regardless of the type of response vectors (simulated vs. real) used. Future studies can be designed using different types of real datasets with different characteristics to see whether the findings of the current study can be replicated.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

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

Supplemental Material

The online appendices are available at <http://apm.sagepub.com/supplemental>

References

- Angoff, W. (1972). *The development of statistical indices for detecting cheaters*. Berkeley, CA: Educational Testing Service.
- Anikeeff, A. (1954). Index of collaboration for test administrators. *Journal of Applied Psychology*, 38, 174-177.
- Bay, L. (1995, April). *Detection of cheating on multiple-choice examinations*. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.
- Bellezza, F. S., & Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology*, 16, 151-155.
- Belov, D. I. (2011). Detection of answer copying based on the structure of a high-stakes test. *Applied Psychological Measurement*, 35, 495-517.
- Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback-Leibler divergence and K-index. *Applied Psychological Measurement*, 34, 379-392.
- Bird, C. (1927). The detection of cheating in objective examinations. *School and Society*, 25(635), 261-262.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
- Bopp, M., Gleason, P., & Misicka, S. (2001). *Reducing incidents of cheating in Adolescence* (Master's thesis). Saint Xavier University, Chicago, IL.
- Brimble, M., & Clarke, P. S. (2005). Perceptions of the prevalence and seriousness of academic dishonesty in Australian universities. *Australian Educational Researcher*, 32(3), 19-44.
- Cai, L., du Toit, S. H. C., & Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago, IL: Scientific Software International.
- Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum.

- Cody, R. P. (1985). Statistical analysis of examinations to detect cheating. *Journal of Medical Education*, 60, 136-137.
- Crouch, E. (2012, April 4). School test fraud is target of Missouri bill. *St. Louis Today*. Retrieved from http://www.stltoday.com/news/local/education/school-test-fraud-is-target-of-missouri-bill/article_ab95386b-48c2-50ef-b5c6-e110a445019b.html
- Dickenson, H. (1945). Identical errors and deception. *Journal of Educational Research*, 38, 534-542.
- 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, 67-85.
- Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, 2, 235-256.
- Green, J. (2014, June 4). Cheating at Paly High: Algebra 2 final exams invalidated. *San Jose Mercury News*. Retrieved from http://www.mercurynews.com/ci_25900169/algebra-2-final-exam-scores-at-palo-alto
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29-36.
- Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying* (ACT Research Report No. 87-15). Iowa City, IA: American College Testing.
- Harnisch, D. I., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18, 133-146.
- Hildebrand, J. (2012, March 15). State creates unit to combat test fraud. *Newsday*. Retrieved from <http://www.newsday.com/long-island/education/state-creates-unit-to-combat-test-fraud-1.3605217>
- Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-Index: Statistical theory and empirical support* (ETS Research Report No. 96-97). Princeton, NJ: Educational Testing Service.
- Hughes, J. M. C., & McCabe, D. L. (2006). Academic misconduct within higher education in Canada. *Canadian Journal of Higher Education*, 36(2), 1-21.
- Jensen, A. L., Arnett, J. J., Feldman, S. S., & Cauffman, E. (2002). It's wrong, but everybody does it: Academic dishonesty among high school and college students. *Contemporary Educational Psychology*, 27, 209-228.
- Josephson Institute of Ethics. (2006). *The ethics of American youth*. Retrieved from <http://charactercounts.org/pdf/reportcard/2006/reportcard-all.pdf>
- Josephson Institute of Ethics. (2008). *The ethics of American youth*. Retrieved from http://charactercounts.org/pdf/reportcard/2008/Q_all.pdf
- Josephson Institute of Ethics. (2010). *The ethics of American youth*. Retrieved from http://charactercounts.org/pdf/reportcard/2010/ReportCard2010_data-tables.pdf
- Josephson Institute of Ethics. (2012). *The ethics of American youth*. Retrieved from <http://charactercounts.org/pdf/reportcard/2012/ReportCard-2012-DataTables.pdf>
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16, 277-298.
- Lin, C. H. S., & Wen, L. Y. M. (2007). Academic dishonesty in higher education: A nationwide study in Taiwan. *Higher Education*, 54(1), 85-97.
- McCabe, D. (2001). Cheating: Why students do it and how we can help them stop. *American Educator*, 25(4), 38-43.
- National Council on Measurement in Education. (2012). *Testing and data integrity in the administration of statewide student assessment programs*. Retrieved from <http://ncme.org/default/assets/File/Committee%20Docs/Test%20Score%20Integrity/Test%20Integrity-NCME%20Endorsed%20%282012%20FINAL%29.pdf>
- Rakovski, C. C., & Levy, E. S. (2007). Academic dishonesty: Perceptions of business students. *College Student Journal*, 41, 466-481.
- R Core Team. (2014). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

- Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo, Japan: Tosho.
- Sauepe, J. (1960). An empirical model for the corroboration of suspected cheating on multiple-choice tests. *Educational and Psychological Measurement*, 20, 475-489.
- Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, 16, 149-157.
- Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the K-index for detecting answer copying. *Journal of Educational Measurement*, 39, 115-132.
- Sotaridona, L. S., & Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement*, 40, 53-69.
- Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, 49, 95-110.
- Thissen, D., Chen, W. H., & Bock, R. D. (2003). Multilog (Version 7) [Computer software]. Lincolnwood, IL: Scientific Software International.
- Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 83-108). New York, NY: Academic Press.
- U.S. Department of Education, Institute of Education Sciences, & National Center for Education Statistics. (2013). *Testing Integrity Symposium: Issues and Recommendations for Best Practice*. Retrieved from <http://nces.ed.gov/pubs2013/2013454.pdf>
- Vandehey, M. A., Diekhoff, G. M., & LaBeff, E. E. (2007). College cheating: A twenty-year follow-up and the addition of an honor code. *Journal of College Student Development*, 48, 468-480.
- van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, 13, 267-298.
- van der Linden, W. J., & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, 31, 283-304.
- Whitley, B. E. (1998). Factors associated with cheating among college students: A review. *Research in Higher Education*, 39, 235-274.
- Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, 21, 307-320.
- Wollack, J. A. (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement*, 40, 189-205.
- Wollack, J. A. (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education*, 19, 265-288.
- Wollack, J. A., & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement*, 22, 144-152.
- Zopluoglu, C. (2013). CopyDetect: An R package for computing statistical indices to detect answer copying on multiple-choice examinations. *Applied Psychological Measurement*, 37, 93-95.
- Zopluoglu, C., Chen, T. T., Huang, C. Y., & Mroch, A. A. (2013, October). *The performance of statistical indices in detecting answer copying on multiple-choice examinations using dichotomous item scores*. Paper presented at the 2nd Annual Conference on Statistical Detection of Potential Test Fraud, Madison, WI.
- Zopluoglu, C., & Davenport, E. C., Jr. (2012). The empirical power and type I error rates of the GBT and ω indices in detecting answer copying on multiple-choice tests. *Educational and Psychological Measurement*, 72, 975-1000.