The Empirical Power and Type I Error Rates of the GBT and Indices in Detecting Answer Copying on Multiple-Choice Tests

Article *in* Educational and Psychological Measurement · October 2012

DOI: 10.1177/0013164412442941

Cengiz Zopluoglu
University of Oregon

Educational and Psychological Measurement
http://epm.sagepub.com/

The Empirical Power and Type I Error Rates of the GBT and ω Indices in Detecting Answer Copying on Multiple-Choice Tests

Cengiz Zopluoglu and Ernest C. Davenport, Jr.

Educational and Psychological Measurement 2012 72: 975 originally published online 18 July 2012
DOI: 10.1177/0013164412442941

Educational and Psychological
Measurement
72(6) 975–1000

DOI: 10.1177/0013164412442941
http://epm.sagepub.com

The Empirical Power and Type I Error Rates of the GBT and ω Indices in Detecting Answer Copying on Multiple-Choice Tests

Cengiz Zopluoglu[1] and Ernest C. Davenport, Jr.[1]

Abstract

The generalized binomial test (GBT) and ω indices are the most recent methods suggested in the literature to detect answer copying behavior on multiple-choice tests. The ω index is one of the most studied indices, but there has not yet been a systematic simulation study for the GBT index. In addition, the effect of the ability levels of the examinees in answer copying pairs on the statistical properties of the GBT and ω indices have not been systematically addressed as yet. The current study simulated 500 answer copying pairs for each of 1,440 conditions (12 source ability level × 12 cheater ability level × 10 amount of copying) to study the empirical power and 10,000 pairs of independent response vectors for each of 144 conditions (12 source ability level × 12 cheater ability level) to study the empirical Type I error rates of the GBT and ω indices. Results indicate that neither GBT nor ω inflated the Type I error rates, and they are reliable to use in practice. The difference in statistical power of these two methods was very small, and GBT performs slightly better than does ω. The main effect for the amount of copying and the interaction effect between source ability level and the amount of copying are found to be very strong while all other main and interactions effects are negligible.

Keywords

answer copying, item response theory, generalized binomial test, multiple-choice tests, cheating

[1] University of Minnesota, Minneapolis, MN, USA

Corresponding Author:
Cengiz Zopluoglu, 140 Education Sciences Building, 56 East River Road, Minneapolis, MN 55455-0364, USA
Email: zoplu001@umn.edu

Multiple-choice tests are still a common assessment tool used in the schools, and answer copying behavior on multiple-choice tests is a type of academic dishonesty that is all too common. The results of the recent self-report surveys in the United States reveal that about 25% of the high school students cheated once, and about 34% of the high school students cheated two or more times during a test in the previous year (Josephson Institute of Ethics, 2006, 2008, 2010). Other survey studies also indicate that answer copying behavior is an important threat that may invalidate the test scores at different levels of education not only in the United States but also in other countries (Bernardi, Baca, Landers, & Witek, 2008; Bopp, Gleason, & Misicka, 2001; Brimble & Clarke, 2005; Diekhoff, LaBeff, Shinohara, & Yasukawa, 1999; Hughes & McCabe, 2006; Jensen, Arnett, Feldman, & Cauffman, 2002; Lin & Wen, 2007; Lupton & Chapman, 2002; Rakovski & Levy, 2007; Vandehey, Diekhoff, & LaBeff, 2007; Whitley, 1998).

Situational and environmental factors were found to be the most important variables related to answer copying behavior. For instance, McCabe and Trevino (1997) found that the contextual variables explained 27% of the total variance whereas the individual variables explained only 3% of the total variance in answer copying behavior. One of the situational factors that lead students to answer copying behavior is ''the perceived risk of detection'' (Whitley, 1998). Students perceive low risk of being caught, and this perception leads to a cost–benefit process in favor of answer copying (Hughes & McCabe, 2006). Although the fear of punishment was found to be the most effective deterrent of cheating, this fear did not stop the students from answer copying, because the likelihood of being caught was perceived too low by the students (Diekhoff et al., 1999). Another situational factor that triggered the answer copying behavior was found to be ''the negligence of instructors,'' and the primary reason was the insufficient evidence of answer copying (Spiegel, Tabachnick, Whitley, & Washburn, 1998). Similarly, Hughes and McCabe (2006) reported that 46% of the faculty members and 38% of the teaching assistants had ignored answer copying. When they were asked why they ignored the answer copying, the main reason was the lack of evidence, stated by 85% of the faculty members and 79% of the teaching assistants.

The post hoc analytical procedures developed for detecting answer copying behavior may help reduce the frequency of answer copying behavior on multiple-choice tests. The awareness that the responses are analytically screened after the test may increase the perceived cost of answer copying behavior by increasing the likelihood of being caught and reduce the frequency of answer copying behavior. In addition, analytical procedures can provide additional evidence when a proctor suspects that a pair of students exchange answers on the test. However, the evidence provided by the analytical procedures is always a concern from both legal and statistical perspectives (Buss & Novick, 1980). An optimal analytical method should be powerful enough to detect true answer copying pairs and be also reliable enough not to detect the honest pairs of students more than expected by chance. Especially, it may be very destructive to accuse an honest student of answer copying. Therefore, the empirical power and the Type I error rates of the analytical procedures used in detecting answer copying are very essential for practitioners, and investigating the statistical properties of these methods under different conditions is necessary before implementing them in practice.

Research Purpose

Many post hoc analytical methods have been developed and suggested in the literature to provide additional evidence of answer copying behavior between two students. While most of these methods are based on unreasonable assumptions and have only historical importance, some recent ones have been found to be useful in detecting answer copying behavior. Two of these recently developed methods, the ω index (Wollack, 1996) and the generalized binomial test (GBT; Van der Linden & Sotaridona, 2006), are particularly the interest of this study, because they are the only ones that use the item response theory (IRT) framework in implementation.

The GBT and ω indices originally use the nominal response model (Bock, 1972) to compute the likelihood of matching between two response vectors, but they can be adjusted for the use of other dichotomous and polytomous IRT models. While the GBT index is based on the exact null distribution of the number of matching response alternatives between two response vectors, the ω index is a normal approximation to this exact null distribution (see Appendix B). The empirical Type I error rates and power of the ω index are well addressed in the previous simulation studies (Sotaridona & Meijer, 2002, 2003; Wollack, 1996, 2003; Wollack & Cohen, 1998), but there has not yet been a systematic investigation for the empirical performance of the GBT index. Also, there has not yet been a systematic comparison between the empirical performances of the ω and GBT indices under different conditions. In addition, the effects of ability levels of the examinees in answer copying pairs on the empirical power of these methods in detecting answer copying do not seem to be well addressed in the previous literature.

Therefore, the purpose of this research is to contribute to the current literature of detecting answer copying on multiple choice tests by systematically investigating the empirical power and Type I error rates of the GBT index in a simulation environment under different conditions, comparing two IRT-based detection methods, GBT and ω, in terms of their empirical performances in detecting answer copying and exploring the effects of ability levels of the examinees in answer copying pairs on the empirical power of the GBT and ω indices.

Theoretical Background

Sotaridona and Meijer (2002) stated that ‘‘the variety of methods to cheat on educational tests seems to be only restricted to one’s imagination.’’ Similarly, the variety of methods to detect answer copying on multiple-choice tests are only restricted to the statisticians’ imagination. There are about 20 different statistical procedures proposed in the current literature to detect answer copying on multiple-choice tests (Angoff, 1972; Anikeeff, 1954; Bay, 1995; Bellezza & Bellezza, 1989; Bird, 1927; 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, 1996). But most of them have just historical importance and are not convenient to use in practice due to the unreasonable assumptions. All these methods attempt to find an unusual agreement between two response vectors by computing the likelihood of matching response alternatives. In addition to these methods, Levine and Rubin (1979) also recommended a method for detecting answer copying based on detecting irregular response patterns from the IRT perspective, but it did not take much attention in the literature, because irregular response patterns may occur due to several reasons other than answer copying (Frary, 1993). Although most copiers have irregular response patterns, it does not indicate that the examinees with irregular response patterns are always copiers. It is very difficult to accuse an examinee of cheating without demonstrating an unusual answer similarity with another examinee in an acceptable proximity (Wollack, 1996).

The statistical methods to detect answer copying can be classified into three groups based on the evidence used for the unusual agreement between two examinees' response vectors. Some researchers suggested using only the identical incorrect responses as evidence of answer copying (Angoff, 1972; Anikeef, 1954; Bellezza & Bellezza, 1989; Bird, 1927; Cody, 1985; Dickenson, 1945; Hanson et al., 1987; Holland, 1996; Sotaridona & Meijer, 2002, 2003), because they thought that two examinees could not be accused for answer copying due to the identical correct responses. These researchers developed their statistical methods based on the agreement between two examinees' incorrect responses. Other researchers recommended obtaining information not only from identical incorrect responses but also from identical correct responses between two response vectors (Saupe, 1960; Sotaridona & Meijer, 2003). Therefore, they based their methods on the agreement between two examinees' identical correct and incorrect responses. However, Buss and Novick (1980) criticized both types of methods from a legal perspective. If identical responses are considered as evidence of answer copying, then nonidentical responses should also be considered as evidence of no answer copying. They argued that the methods that did not take all items into consideration were unfair, and the decisions made based on those methods were legally questionable. Their argument was as follows:

> Some testing programs now examine only the number of identical incorrect responses, and reject the hypothesis of independent response if the number of such items for any pair of examinees is substantially more than might be expected by chance. When the statistical index of cheating is computed in this way, evidence that may be favorable to the examinee from the items that were answered correctly by one examinee and incorrectly by the other (or incorrectly in different ways by two examinees) is ignored, and this is unfair to any examinee for whom such evidence would bear in a strong way in favor of the hypothesis of independent responses. (pp. 11-12)

In the last group, other methods are also available that takes all items into account when computing the degree of similarity between two response vectors (Bay, 1995; Frary et al., 1977; Van der Linden & Sotaridona, 2006; Wollack, 1996).

In another type of classification, the analytical methods can be placed into four different categories based on the statistical distribution they use to determine the degree of unusual agreement. These analytical methods use empirical null distribution of identical correct and incorrect responses (Angoff, 1972; Bird, 1927; Dickenson, 1945; Hanson et al., 1987; Saupe, 1960), binomial distribution (Anikeef, 1954; Bellezza & Bellezza, 1989; Holland, 1996; Sotaridona & Meijer, 2002), Poisson distribution (Sotaridona & Meijer, 2003), generalized binomial distribution (Bay, 1995; Van der Linden & Sotaridona, 2006), or normal approximation to the generalized binomial distribution (Cody, 1985; Frary et al., 1977; Wollack, 1996).

The basic idea in all these methods is to compute the likelihood of agreement between two response vectors. A critical step in computing the likelihood of unusual agreement between two response vectors is to estimate the probability of selecting a response alternative k of item i for an examinee j with a certain ability level (p_{jik}). The ω and GBT indices are the only methods that propose using IRT models when estimating p_{jik}. In implementation of these two methods, an IRT model, originally a nominal response model, is first fitted to response data, and person and item parameters are estimated. Then, p_{jik} is estimated using the IRT model parameter estimates. All other indices proposed using the classical test theory–based parameters when estimating p_{jik}. Number-correct or number-incorrect scores for examinees and proportion corrects for items are first calculated from response data, and then p_{jik} is empirically estimated by either conditioning or not conditioning on the number-correct or -incorrect scores of the examinees.

Despite the large number of different analytical methods, a few of them are really useful in practice in detecting answer copying in multiple-choice tests, whereas others have just historical importance due to the unreasonable assumptions they make in implementation. The single and comparative performances of these methods are studied mostly in a simulation environment. Power and Type I error rates are the two most important concerns when a statistical procedure is used to make a decision. Statistical methods that provide the most power and also control the Type I error rate at the nominal level are the best candidates to use in practice. In the context of detecting answer copying research, the Type I error rate is falsely detecting an honest pair of examinees as copiers, while the power is truly detecting an answer copying pair. Since there are so many statistical indices in the literature proposed to provide additional evidence for answer copying, it is important for practitioners to know which methods are most powerful and control the Type I error rate at the nominal level. However, this is not an easy task to study in practice, because researchers need true answer copying pairs to study the power, and true honest pairs to study the Type I error rates of an index. It is not easy to obtain these answer copying and honest pairs in practice, because it is not always possible to be sure of the answer copying behavior in reality.

Simulation Studies to Evaluate the Performance of Detection Indices

A practical way to obtain these answer copying and honest pairs of examinees is by creating them in a simulation environment. In fact, it is the only way for studying the power efficiently, because the true copiers are mostly not known in reality. Simulation studies also provide a more flexible environment to manipulate several variables that may affect the statistical power and Type I error rates of these indices.

Table 1 summarizes previous simulation studies that investigated the empirical Type I error rates and power of some statistical indices in the literature. The goal of the previous research was to provide information for practitioners about the indices relative to their power to detect true copiers and not to detect honest pairs. In these studies, answer copying pairs are simulated by first generating two independent response vectors and then overwriting one response vector (source) on another (cheater) for a number of selected items. The items to overwrite are selected from all items, so some items that had already been identical between two response vectors may be selected, and the answer copying may or may not change the cheater examinee's responses. It reflects a real-life situation where a cheater examinee looks at a source examinee's response for an item, and keeps his/her response the same if the response is already the same with the source examinee's response; or the cheater examinee may change his/her response if it is not the same with the source examinee's response (Wollack, 2009, personal communication).

Five types of answer copying scenarios were used in previous studies: random copying, difficulty weighted copying, random string copying, string beginning copying, and string end copying. Random copying reflects the situation that a cheater randomly chooses the items to copy among all items in the test. Difficulty weighted imitates the situation that harder items are more likely to be copied. String beginning and string end reflect the situation that a cheater copies the first or last *n* items of the test. In random string copying, the items are divided into blocks and each block includes five consecutive items. Then a number of blocks based on the desired amount of copying are selected for answer copying. Many studies used only the random copying situation. A couple studies reported that the type of copying has a small influence on the power of ω after comparing random, random string, and difficulty weighted copying conditions (Wollack, 1996; Wollack & Cohen, 1998).

Other variables manipulated in previous studies were the test length, amount of copying, and sample size. In general, the statistical power of these indices increased as the sample size, test length, and amount of copying increased. Bigger sample size provides more accurate parameter estimation for both IRT and non-IRT based methods to be used in the probabilistic models, so the power of the test is more reliably estimated. Longer tests and higher amounts of copying provide more statistical information regarding answer copying behavior (if it exists), so the empirical power also increases as the test length and amount of copying increase.

The previous research suggested that the ω index is the best choice for detecting answer copying, if the IRT model parameters can be estimated reliably. The ω index provided the highest detection rate, while controlling the Type I error rates at the

Table 1. Simulation Studies Assessed the Indices Developed to Detect Answer Copying

	Compared indices	Sample size	Test length	Amount of copying	Type of copying	Type of data
Hanson et al. (1987)	g_2, B, H, P, CP, Pair I, Pair II	—	100	10%, 20%, 30%, 40%, 50%	Random, difficulty weighted, string end, string beginning, random string	Real
Bay (1995)	B_m, ESA, g_2	100, 200	20, 50	10%, 25%, 50%, 75%, 90%	Random	Real
Wollack (1996)	g_2, ω	100, 500	40, 80	10%, 20%, 30%, 40%	Random, difficulty weighted, random string	Simulated
Wollack and Cohen (1998)	ω	100, 500	40, 80	10%, 20%, 30%, 40%	Difficulty weighted, random strings	Simulated
Sotaridona and Meijer (2002)	ω, K, K_1, K_2	100, 500, 2,000	40, 80	10%, 20%, 30%, 40%	Random	Simulated
Sotaridona and Meijer (2003)	ω, K_2, S_1, S_2	100, 500	40, 80	10%, 20%, 30%, 40%	Random	Simulated
Wollack (2003)	K, Scrutiny, g_2, ω	50, 100, 250, 500, 1,000, 2,000, 5,000, 10,000	20, 40, 80	10%, 20%, 30%, 40%	Random	Real

nominal level. However, K and its variants (Holland, 1996; Sotaridona & Meijer, 2002) as well as the S_1 and S_2 indices (Sotaridona & Meijer, 2003) also held the Type I error rates at the nominal level and provided reasonable detection rates. Therefore, K and its variants and S_1 and S_2 are also recommended in detecting answer copying. All other statistical indices either inflated the Type I error rates or did not provide enough power to detect true answer copying pairs. The GBT index was the most recently developed index, and the developers suggested a formula to obtain its theoretical power, but there has not yet been a systematic simulation study to compare its empirical power and Type I error rates to other methods.

The accuracy of parameter calibration is important for the IRT-based methods, because the computations depend on the model parameters. Previous simulation studies that use known item parameters when computing the ω index are criticized, because the item parameters are unknown in practice and have to be estimated from the sample. Wollack and Cohen (1998) studied whether using estimated item parameters rather than known item parameters had influence on the ω index. They reported that there was a small difference between using known or using estimated item parameters regarding the empirical power and Type I error rates of the ω index.

Although there are promising methods available for detecting answer copying, an important factor that may affect the power of these methods seems to be ignored by previous studies. Most of these studies put constrains on the ability level of the hypothetical cheater and source examinees when simulating answer copying. For instance, the ability levels of the hypothetical cheater examinees were assumed to be much lower than the ability levels of the hypothetical source examinees. As pointed out by Lewis and Thayer (1998), the power in detecting true answer copying pairs is very likely to decrease as the ability of the source examinee increase. For instance, let us think about two hypothetical cheater examinees, Examinee A and Examinee B, whose ability levels are equal to -1, and two hypothetical source examinees, Examinee C and Examinee D, whose ability levels are equal to 0 and 2, respectively, in a standard normal distribution. Let us say that Examinee A copies 40% of the items from Examinee C, while Examinee B copies the same amount of items from Examinee D in the same multiple-choice test. It is expected that the likelihood of being detected by any analytical method is higher for Examinee A than for Examinee B, although they have copied the same amount of items. The degree of reduction in the statistical power in detecting answer copying at different ability levels of the source examinee is an empirical question, and this fact brings up some questions regarding the utility of these analytical methods. So far, none of the previous studies focused and systematically manipulated the ability levels of the examinees in answer copying pairs, and there is little empirical information regarding the degree of reduction in statistical power at different ability levels of the examinees in answer copying pairs.

In light of previously mentioned facts, there are two main objectives of this study. The first objective is to study the empirical power and the Type I error rates of the GBT index in detecting answer copying and to compare its performance to the ω index, which is another IRT-based method in detecting answer copying. The second objective of this study is to investigate the empirical power of the GBT and ω indices for various combinations of cheater and source examinee ability levels and to examine the reduction in empirical power of these two indices in detecting answer copying for increasing levels of source examinee ability.

Method

Independent and Dependent Variables

There are many variables to consider for a simulation study of empirical power and Type I error rates in detecting answer copying. These variables are test length, sample size, amount of copying, type of copying, and the ability levels of the examinees in answer copying pairs. Manipulating all these variables at once is a challenging task, because the fully crossed matrix would have thousands of different experimental cell conditions. Therefore, the current study chose to manipulate the amount of copying and ability levels of the examinees in answer copying pairs as the independent variables. The type of copying was not manipulated, because Wollack and Cohen (1998) reported that the type of copying had small influence on the power of ω after comparing random, random string, and difficulty weighted copying conditions. Similar to most of the previous simulation studies, the random copying situation was used in the current study. Test length was fixed, and a hypothetical 40-item test is used when simulating response vectors. In this study, the single pairs of response vectors are simulated, so the sample size was not an issue. The GBT and ω indices can be calculated for a single pair of response vectors once the IRT item parameters are known as assumed in this study.

The study manipulated the amount of copying in ten different levels from 10% to 100% in increments of 10%. The IRT ability scale between −3 and +3 was divided into 12 equal intervals in increments of .5 to manipulate the ability levels of the examinees in answer copying pairs. The number of possible combinations between the source and the cheater ability level intervals was 144 (12 × 12). Each level of amount of copying is fully crossed with 144 combinations of the source and cheater ability levels, yielding a total of 1,440 conditions (12 × 12 × 10) for studying the empirical power of the GBT and ω indices. Five hundred answer copying pairs were generated within each experimental condition (see Appendix A). For studying Type I error rates, there was no need to simulate answer copying. Therefore, each level of the source ability interval is fully crossed with the levels of the cheater ability interval, yielding a total of 144 conditions (12 × 12) for studying the empirical Type I error rates of the GBT and ω indices. Ten thousand honest pairs (pairs of independent response vectors) were generated within each experimental condition (see Appendix A). The dependent variable was the proportion of answer copying pairs being truly detected for the power analysis, and the proportion of honest pairs being falsely detected for the Type I error analysis.

Data Analysis

First the GBT and ω indices were computed for each pair of response vectors simulated. The proportion of 500 answer copying pairs being successfully detected within each of the 1,440 experimental conditions, and the proportion of 10,000 honest pairs being falsely detected within each of the 144 experimental conditions was computed based on a specified alpha level (α = .05, .01, .001) for both indices. The descriptive proportions were cross tabulated in several tables for each condition.

Although presenting the descriptive statistics is both effective and informative, these representations lack the detection of important effects and the magnitude of significant effects. In contrast, inferential analysis provides information and insights beyond that of descriptive statistics (Harwell, 1997; Harwell, Stone, Hsu, & Kirisci, 1996). Given this, a multiple logistic regression model was run to examine the effects of source ability, cheater ability, and number of items copied on the statistical power of the GBT and ω indices. The outcome variable for this analysis was the proportion of answer copying pairs being truly detected, and the independent variables were the source ability level, cheater ability level, and the number of items copied. McFadden's pseudo R^2 (McFadden, 1973) was also computed as an effect size measure to evaluate the relative contribution of each significant term in reducing the prediction error. It was calculated for a model term based on the following formula:

$$R_{\mathrm{M}}^2 = \frac{D_{\mathrm{R}} - D_{\mathrm{F}}}{D_0},$$

where R_{M}^2 is the contribution of model term in reducing the total error, D_0 is the deviance for the null model (no predictor), D_{F} is the deviance for the full model with all terms included, and D_{R} is the deviance for the reduced model after excluding the model term from the full model.

Results

It is important to use a reasonable alpha level in detecting answer copying. Committing an error is more serious in answer copying, because detecting an honest pair of students as an answer copying pair may cause serious problems. The α level of .05 is generally thought of very liberal and not very likely to be used in practice. On the other hand, the α level of .001 may be very conservative and reduce the power in detecting true answer copying pairs. The α level of .01 seems more reasonable and likely to be used in practice. Therefore, the results are reported for the empirical power and Type I errors rates of the GBT and ω indices at the α level of .01. Similar patterns were observed for other levels of alpha, so they are not reported. All tabulated results are available from the author.

Empirical Power of the GBT and ω Indices

The empirical power of the indices was tabulated in Tables 2, 3, and 4 that allowed to examine the two-way interaction effects among the source ability level, cheater ability level, and amount of copying. The final column and rows in these tables presented the main effect of the independent variables on the empirical power of the GBT and ω indices.

A quick look at the Table 2 revealed that the cheater ability level had little effect, whereas the source ability level had substantial effect on the empirical power of both indices, especially when the source ability level was above zero. The overall power for both indices was around .60 regardless of the ability level of the cheater examinee in the answer copying pair. The overall power for both indices was around .8 when the ability of the source examinee is below zero, but the power decreased to .05 as the ability of the source examinee increased beyond zero ability level. Although Table 2 suggests an interaction effect between source and cheater ability levels, the interaction did not seem to be strong. The power of the GBT index slightly increased, remained same, and slightly decreased for the source ability levels below −1.5, between −1.5 and −0.5, and above 0.5, respectively, as the cheater ability increased. The proportion of answer copying pairs did not reach to 0.5 in most cases when the source ability is above 1.

The main effect of the amount of copying and the two-way interaction effects between the amount of copying and the ability levels of the examinees in answer copying pairs can be examined through Tables 3 and 4. The results indicated that the amount of copying had considerable effect on the power of these two indices. The proportion of being detected increased rapidly from .02 to .61 until the half of the items were copied, and then increased at a slower rate after 50% copying. Table 3 suggested a significant interaction effect between the ability level of source examinee and the amount of copying. When the source ability is below 0, the empirical power quickly hit the ceiling for both indices after 30% copying. As the source ability increased, the power in detecting answer copying power showed different patterns. For instance, the power reached .9 only when the source ability is between 1 and 1.5 and the amount of copying was 90%. The power was not above 0.5 in most cases even for 90% copying when the source ability was above 1.5. This is not surprising since we can only estimate the observed ability of the cheater examinees after answer copying. When the cheater examinees copy answers from high ability students, they are likely to copy correct answers and increase their observed ability estimates. Therefore, the likelihood of matching between two response vectors becomes higher and makes the detection difficult. Table 4 suggested that there is not an interaction effect between the ability levels of cheater examinees in answer copying pairs and the amount of copying. The effect of amount of copying on the empirical power was very similar within different ability levels of cheater examinees.

The probability of being detected was averaged across all 144 combinations of the source and cheater ability levels within a level of amount of copying and presented in Figure 1 at different alpha levels. The overall empirical power of the GBT index

Table 2. The Empirical Power of the GBT and ω Indices: Interaction Between Source and Cheater Ability Levels (α = .01)

Source ability	Cheater ability level												
level	−2.75	−2.25	−1.75	−1.25	−0.75	−0.25	0.25	0.75	1.25	1.75	2.25	2.75	
GBT Index													
−2.75	0.76	0.77	0.77	0.77	0.78	0.80	0.82	0.83	0.84	0.86	0.86	0.86	0.81
−2.25	0.78	0.77	0.78	0.78	0.79	0.80	0.82	0.83	0.84	0.84	0.84	0.85	0.81
−1.75	0.79	0.79	0.79	0.80	0.80	0.80	0.81	0.82	0.82	0.83	0.83	0.83	0.81
−1.25	0.79	0.81	0.80	0.81	0.81	0.81	0.81	0.82	0.81	0.81	0.81	0.81	0.81
−0.75	0.80	0.80	0.81	0.81	0.81	0.80	0.79	0.79	0.80	0.80	0.79	0.79	0.80
−0.25	0.80	0.80	0.79	0.78	0.78	0.77	0.77	0.76	0.76	0.76	0.76	0.75	0.77
0.25	0.77	0.75	0.74	0.72	0.71	0.71	0.71	0.71	0.70	0.70	0.71	0.71	0.72
0.75	0.71	0.67	0.64	0.62	0.59	0.59	0.59	0.60	0.61	0.61	0.61	0.62	0.62
1.25	0.56	0.53	0.49	0.44	0.42	0.43	0.43	0.44	0.45	0.47	0.49	0.49	0.47
1.75	0.38	0.34	0.30	0.26	0.25	0.25	0.26	0.27	0.29	0.31	0.34	0.34	0.30
2.25	0.21	0.18	0.14	0.12	0.12	0.12	0.12	0.14	0.15	0.17	0.18	0.19	0.15
2.75	0.10	0.09	0.06	0.05	0.05	0.05	0.05	0.06	0.06	0.08	0.08	0.09	0.07
	0.62	0.61	0.59	0.58	0.57	0.58	0.58	0.59	0.59	0.60	0.61	0.61	
ω Index													
−2.75	0.72	0.72	0.72	0.73	0.74	0.77	0.79	0.81	0.83	0.85	0.86	0.86	0.78
−2.25	0.74	0.74	0.73	0.74	0.75	0.76	0.79	0.81	0.83	0.84	0.85	0.86	0.79
−1.75	0.76	0.76	0.76	0.76	0.76	0.77	0.79	0.81	0.83	0.84	0.85	0.85	0.79
−1.25	0.78	0.78	0.77	0.77	0.77	0.78	0.79	0.81	0.82	0.83	0.84	0.85	0.80
−0.75	0.78	0.78	0.77	0.77	0.76	0.77	0.77	0.79	0.82	0.83	0.83	0.85	0.79
−0.25	0.77	0.76	0.75	0.74	0.74	0.73	0.74	0.76	0.79	0.80	0.82	0.83	0.77
0.25	0.73	0.71	0.69	0.67	0.66	0.66	0.67	0.70	0.73	0.75	0.77	0.79	0.71
0.75	0.63	0.58	0.55	0.53	0.51	0.52	0.55	0.59	0.62	0.65	0.68	0.70	0.59
1.25	0.42	0.40	0.36	0.34	0.33	0.35	0.37	0.41	0.45	0.49	0.53	0.56	0.42
1.75	0.23	0.21	0.19	0.18	0.18	0.19	0.21	0.23	0.27	0.30	0.35	0.36	0.24
2.25	0.10	0.09	0.08	0.08	0.08	0.08	0.09	0.11	0.13	0.15	0.17	0.19	0.11
2.75	0.04	0.04	0.03	0.03	0.03	0.04	0.04	0.05	0.05	0.07	0.07	0.09	0.05
	0.56	0.55	0.53	0.53	0.53	0.54	0.55	0.57	0.60	0.62	0.63	0.65	

Note: GBT = generalized binomial test. The ability level intervals are represented by their midpoints. The cell proportions are based on 5000 answer copying pair.

Table 3. The Empirical Power of the GBT and ω Indices: Interaction Between Source Ability Level and Amount of Copying (α = .01)

Source ability	Amount of copying										
level	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	
GBT Index											
−2.75	0.04	0.32	0.77	0.96	1.00	1.00	1.00	1.00	1.00	1.00	0.81
−2.25	0.03	0.30	0.78	0.97	1.00	1.00	1.00	1.00	1.00	1.00	0.81
−1.75	0.04	0.27	0.79	0.98	1.00	1.00	1.00	1.00	1.00	1.00	0.81
−1.25	0.04	0.26	0.79	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.81
−0.75	0.03	0.23	0.75	0.98	1.00	1.00	1.00	1.00	1.00	1.00	0.80
−0.25	0.02	0.18	0.60	0.93	1.00	1.00	1.00	1.00	1.00	1.00	0.77
0.25	0.02	0.11	0.41	0.75	0.93	0.99	1.00	1.00	1.00	1.00	0.72
0.75	0.01	0.06	0.20	0.45	0.70	0.87	0.95	0.99	1.00	1.00	0.62
1.25	0.00	0.02	0.08	0.20	0.37	0.55	0.72	0.85	0.93	0.98	0.47
1.75	0.00	0.01	0.03	0.08	0.15	0.24	0.38	0.55	0.68	0.85	0.30
2.25	0.00	0.01	0.01	0.03	0.06	0.09	0.16	0.23	0.37	0.56	0.15
2.75	0.00	0.01	0.01	0.01	0.03	0.03	0.06	0.10	0.15	0.29	0.07
	0.02	0.15	0.43	0.61	0.68	0.73	0.77	0.81	0.84	0.89	
ω Index											
−2.75	0.03	0.26	0.64	0.90	0.99	1.00	1.00	1.00	1.00	1.00	0.78
−2.25	0.02	0.25	0.66	0.93	1.00	1.00	1.00	1.00	1.00	1.00	0.79
−1.75	0.02	0.25	0.71	0.96	1.00	1.00	1.00	1.00	1.00	1.00	0.79
−1.25	0.03	0.26	0.72	0.97	1.00	1.00	1.00	1.00	1.00	1.00	0.80
−0.75	0.03	0.25	0.70	0.97	1.00	1.00	1.00	1.00	1.00	1.00	0.79
−0.25	0.02	0.20	0.59	0.90	0.99	1.00	1.00	1.00	1.00	1.00	0.77
0.25	0.02	0.12	0.39	0.70	0.91	0.98	1.00	1.00	1.00	1.00	0.71
0.75	0.01	0.06	0.18	0.38	0.61	0.81	0.91	0.97	0.99	1.00	0.59
1.25	0.00	0.03	0.07	0.15	0.27	0.43	0.60	0.77	0.89	0.97	0.42
1.75	0.00	0.01	0.02	0.05	0.09	0.16	0.26	0.42	0.58	0.81	0.24
2.25	0.00	0.00	0.01	0.02	0.03	0.05	0.09	0.15	0.26	0.51	0.11
2.75	0.00	0.00	0.00	0.01	0.01	0.02	0.03	0.05	0.10	0.26	0.05
	0.01	0.14	0.39	0.58	0.66	0.70	0.74	0.78	0.82	0.88	

Note: GBT = generalized binomial test. The ability level intervals are represented by their midpoints. The cell proportions are based on 6000 answer copying pair.

Table 4. The Empirical Power of the GBT and ω Indices: Interaction Between Cheater Ability Level and Amount of Copying (α = .01)

Cheater ability level	Amount of copying										
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	
GBT Index											
−2.75	0.03	0.15	0.45	0.68	0.75	0.77	0.80	0.83	0.85	0.89	0.62
−2.25	0.03	0.15	0.42	0.65	0.72	0.76	0.79	0.82	0.85	0.89	0.61
−1.75	0.03	0.14	0.40	0.62	0.70	0.74	0.77	0.81	0.83	0.89	0.59
−1.25	0.03	0.13	0.39	0.60	0.67	0.72	0.76	0.80	0.83	0.89	0.58
−0.75	0.03	0.13	0.39	0.58	0.66	0.70	0.75	0.79	0.83	0.89	0.57
−0.25	0.03	0.14	0.40	0.58	0.65	0.70	0.75	0.79	0.84	0.89	0.58
0.25	0.02	0.14	0.42	0.59	0.65	0.71	0.75	0.79	0.84	0.89	0.58
0.75	0.01	0.15	0.45	0.59	0.67	0.72	0.76	0.80	0.84	0.89	0.59
1.25	0.01	0.15	0.46	0.60	0.67	0.73	0.77	0.81	0.85	0.89	0.59
1.75	0.01	0.17	0.47	0.61	0.68	0.73	0.78	0.82	0.85	0.90	0.60
2.25	0.01	0.16	0.48	0.62	0.70	0.74	0.79	0.83	0.86	0.89	0.61
2.75	0.01	0.16	0.49	0.63	0.69	0.75	0.80	0.83	0.86	0.89	0.61
	0.02	0.15	0.43	0.61	0.68	0.73	0.77	0.81	0.84	0.89	
ω Index											
−2.75	0.01	0.10	0.34	0.58	0.67	0.70	0.73	0.76	0.80	0.88	0.56
−2.25	0.01	0.09	0.31	0.55	0.65	0.70	0.72	0.76	0.80	0.88	0.55
−1.75	0.01	0.08	0.28	0.53	0.63	0.68	0.71	0.75	0.79	0.88	0.53
−1.25	0.01	0.07	0.27	0.51	0.62	0.67	0.71	0.75	0.79	0.88	0.53
−0.75	0.01	0.06	0.28	0.50	0.61	0.66	0.70	0.75	0.80	0.88	0.53
−0.25	0.01	0.08	0.31	0.52	0.62	0.67	0.71	0.76	0.81	0.88	0.54
0.25	0.01	0.10	0.34	0.55	0.63	0.68	0.72	0.77	0.81	0.88	0.55
0.75	0.02	0.14	0.41	0.58	0.65	0.70	0.74	0.79	0.83	0.88	0.57
1.25	0.02	0.19	0.48	0.61	0.67	0.72	0.76	0.80	0.84	0.88	0.60
1.75	0.02	0.23	0.52	0.64	0.69	0.74	0.78	0.81	0.85	0.88	0.62
2.25	0.02	0.27	0.56	0.67	0.72	0.76	0.79	0.83	0.85	0.88	0.63
2.75	0.03	0.30	0.59	0.68	0.73	0.77	0.81	0.84	0.86	0.88	0.65
	0.01	0.14	0.39	0.58	0.66	0.70	0.74	0.78	0.82	0.88	

Note: GBT = generalized binomial test. The ability level intervals are represented by their midpoints. The cell proportions are based on 6000 answer copying pair.

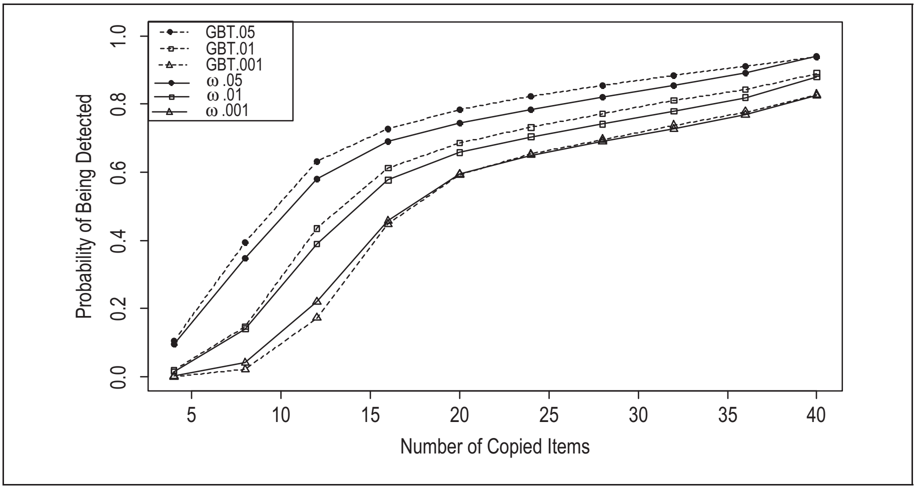

Figure 1. The overall empirical power of the generalized binomial test and ω indices

was slightly higher than the overall empirical power of the ω index. The biggest difference was .039 when the amount of copying was 50% at the α level of .05.

Empirical Type I Error Rates of the GBT and ω Indices

Empirical Type I error rates were calculated based on 10,000 honest pairs at three different theoretical alpha levels. After the GBT and ω indices were calculated for the honest pairs (independent response vectors), the proportion of falsely detected pairs was computed. The empirical Type I error rates of the GBT and ω indices at the α level of .05, .01, and .001 are available from the author. Table 5 presents the results for only α level of .01.

The empirical type I error rates of the GBT index never exceed its nominal level in any of the conditions for different theoretical alpha levels. In most conditions, it was much smaller than the theoretical alpha level. The empirical Type I error rates of the ω index exceeds its nominal level in some conditions. The maximum empirical Type I error rate for the ω index was .071 at $\alpha = .05$, .013 at $\alpha = .01$, and .002 at $\alpha = .001$. In most conditions, the ω index was more successful to hold the Type I error rate around its nominal level. When the empirical Type I error rates were averaged across all 144 combinations of the source and cheater ability levels, the empirical Type I error rates at the α levels of .05, .01, and .001 were .0126, .0015, and .0001 for the GBT index, and .0369, .0064, and .0006 for the ω index, respectively.

The empirical Type I error rates were highest when both examinees' ability levels were below zero, and it was smallest when both examinee's ability levels were above zero. This may be because of the fact that matching incorrect responses contributes to the likelihood of similarity between two response vectors more than matching

Table 5. Empirical Type I Error Rates of the GBT and ω Indices at the Nominal Alpha Level of .01

Source ability level	Cheater ability level											
	−2.75	−2.25	−1.75	−1.25	−0.75	−0.25	0.25	0.75	1.25	1.75	2.25	2.75
GBT Index												
−2.75	0.003	0.005	0.004	0.002	0.003	0.002	0.002	0.001	0.001	0.000	0.000	0.000
−2.25	0.002	0.007	0.006	0.004	0.005	0.003	0.001	0.001	0.000	0.000	0.000	0.001
−1.75	0.005	0.005	0.006	0.006	0.005	0.004	0.002	0.001	0.001	0.000	0.000	0.000
−1.25	0.006	0.006	0.005	0.006	0.005	0.003	0.002	0.001	0.000	0.000	0.000	0.000
−0.75	0.003	0.005	0.004	0.006	0.006	0.006	0.003	0.002	0.000	0.000	0.000	0.000
−0.25	0.003	0.002	0.004	0.004	0.006	0.004	0.002	0.001	0.000	0.000	0.000	0.000
0.25	0.002	0.002	0.002	0.003	0.003	0.002	0.002	0.001	0.000	0.000	0.000	0.000
0.75	0.001	0.000	0.000	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000
1.25	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.75	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2.25	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2.75	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ω Index												
−2.75	0.009	0.009	0.010	0.008	0.008	0.007	0.008	0.009	0.010	0.011	0.010	0.012
−2.25	0.010	0.011	0.010	0.009	0.009	0.008	0.007	0.008	0.010	0.008	0.008	0.007
−1.75	0.013	0.012	0.010	0.012	0.010	0.011	0.008	0.008	0.008	0.009	0.007	0.007
−1.25	0.013	0.013	0.012	0.012	0.012	0.012	0.009	0.010	0.008	0.008	0.006	0.003
−0.75	0.013	0.013	0.012	0.011	0.011	0.013	0.011	0.012	0.010	0.010	0.005	0.004
−0.25	0.011	0.010	0.010	0.011	0.012	0.009	0.009	0.010	0.010	0.009	0.006	0.005
0.25	0.011	0.009	0.007	0.006	0.006	0.006	0.006	0.008	0.009	0.009	0.009	0.006
0.75	0.010	0.007	0.004	0.003	0.003	0.001	0.003	0.003	0.007	0.007	0.006	0.006
1.25	0.008	0.007	0.004	0.001	0.001	0.001	0.001	0.001	0.002	0.003	0.004	0.004
1.75	0.006	0.006	0.002	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.002	0.002
2.25	0.007	0.004	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.001
2.75	0.007	0.005	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note: GBT = generalized binomial test. The ability level intervals are represented by their midpoints. The cell proportions are based on 10,000 pairs of independent response vectors.

correct responses. When the GBT and ω indices are computed between two low ability examinees, they are more likely to have matching incorrect responses, and the high number of matching incorrect responses in low-ability examinees may increase the Type I error rates when these two indices are computed.

Inferential Analysis on the Empirical Power

A multiple logistic regression full model including the three main effects, three two-way interaction effects, and one three-way interaction effect was fitted to the empirical power of the two indices at the α level .01. Two separate runs were conducted for the GBT and ω indices. All terms in the model were statistically significant at the α level of .05 in the analysis for the GBT index. The relative contribution of each term in reducing the total error (null deviance) was computed by excluding the term from the full model and computing the R^2_{M} while all other terms were kept in the model. The results indicated that the main effect for the number of items copied and the interaction effect between the source ability level and the number of items copied had substantial effects on the empirical power of the GBT index. The proportional reduction in total error was 48% and 12% for the main effect of number of items copied and the interaction effect between the source ability level and the number of items copied, respectively. The single proportional reductions in error for all other terms were below .002. Similar results were obtained from the analysis for the ω index. All terms in the model were statistically significant at the α level of .05 in the analysis, but only two of them were practically important. The proportional reduction in error was 41% and 10% for the main effect of number of items copied and the interaction effect between the source ability level and the number of items copied respectively. The single proportional reductions in error for other terms were all below .003.

Discussion

This study compared the empirical power and Type I error rates of the two IRT-based indices, GBT and ω, in detecting answer copying pairs in multiple-choice tests. In addition, the performances of these two indices were examined under different conditions by manipulating the amount of copying and the ability levels of the examinees in answer copying pairs.

In general, the empirical power of both indices was very close to each other, but the GBT index performed slightly better than did the ω index. The power was very low for both indices at low amount of copying conditions. The results indicated that the empirical power of the GBT and ω indices at the α level of .01 were below .5 until 30% of the items are copied. The power reaches about .8 for both indices at the α level of .01 when the 80% of the items are copied.

The study found that the main effect for the amount of copying and the interaction effect between the source ability level and the amount of copying were the dominant factors that affect the power of the GBT and ω indices. In short, the power of these two indices in detecting answer copying highly depends on how many items are copied and from whom the items are copied. For instance, the power does not reach to .5 unless the cheater examinee copies 50% of the items from a source examinee with the ability level above 1, or copies 80% of the items from a source examinee with the ability level above 2.

The results of this study bring up the question of whether these post hoc procedures are useful in practice in detecting answer copying. Although it is always possible for a high-ability examinee to copy from a low-ability examinee, it is not common in reality. Therefore, some of the results in this research might be informative in terms of looking at the whole picture for the observed power of the GBT and ω indices but not meaningful in practice. The examinees who copy answers from a lower ability examinee are more likely to harm themselves rather than to benefit, and these examinees may not be the main concern in practice. It may also not very attractive to detect answer copying pairs when both the source and cheater examinees are low ability examinees because the cheater examinees do not receive much benefit from answer copying. It is practically more important to consider how effectively the GBT and ω indices perform for answer copying pairs in which the cheater ability level is low and the source ability level is high. The results are not promising for those answer copying pairs, and the GBT and ω indices perform worse in detecting answer copying pairs when it is more needed to do so.

The results of the study provide evidence that both indices are reliable to use in practice. The empirical Type I error rates were under their theoretical levels in most conditions for both indices. The empirical Type I error rates for the GBT index was much lower than was the theoretical level compared to the ω index. The empirical Type I error rate was higher when the GBT and ω indices were calculated between two low ability examinees' response vectors, but it was smaller when two examinees have high ability levels. This might be because of the fact that matching incorrect responses contribute more than matching correct responses to the likelihood of correspondence between two response vectors.

The research has many limitations. The first limitation is using known item parameters when calculating the ω and GBT indices. In most cases, it is not expected to have a precalibrated item bank, so the item parameters of the nominal response model have to be estimated. The actual power of the GBT and ω indices is expected to be less than the power in this research, when the estimated item parameters are used. However, Wollack (1998) reported a very slight decrease in the power of the ω index when the item parameters were estimated. The second limitation was using model-based response vectors. In practice, the responses of the real examinees may not follow the nominal response model. The third limitation is the type of copying used in the simulation. It was assumed that cheater examinees copy answers randomly from source examinees. This assumption may not be correct in real life and the results may differ for different types of copying such as difficulty-weighted copying.

Appendix A

There were four steps in simulating answer copying pairs for power analysis. At the first step, the IRT ability level scale was divided into 12 equal intervals from -3 to $+3$ by increment of .5 as denoted by C_t ($t = 1, 2, \ldots, 12$) for cheater ability level intervals and S_p ($p = 1, 2, \ldots, 12$) for source ability level intervals. For instance, C_1 represents the hypothetical cheater examinees whose ability levels are between -3 and -2.5, and S_5 represents the hypothetical source examinees whose ability levels are between -1 and -0.5. Two ability levels were drawn from a uniform distribution within the C_t and S_p intervals, respectively, for a hypothetical cheater examinee and for a hypothetical source examinee. Then, two independent response vectors for a 40-item hypothetical test were simulated based on the nominal response model given the generated ability levels and the item parameters. The nominal response model item parameters were taken from Wollack's (1996) study. At the second step, a random copying situation was simulated between two independent response vectors by randomly selecting m% of the 40 items and overwriting the response vector of hypothetical source examinee on the response vector of hypothetical cheater examinee for the randomly selected items. The same process was replicated until the 500 answer copying pairs were obtained for power analysis within the each level of 1,440 experimental conditions.

At the third step, the ability levels of both hypothetical cheater and source examinees in answer copying pairs were reestimated based on the response vectors after simulating answer copying using maximum-likelihood estimation in MULTILOG (Thissen, 2003). Item parameters were assumed to be known when reestimating the ability levels. At the final step, the reestimated ability levels were used in computing the GBT and ω indices for 500 answer copying pairs in each experimental condition. S-plus code for the GBT index was kindly provided by its developers and modified to be able to use in R (R Development Core Team, 2009) to compute the GBT index. The code to calculate the ω index was written in R by the authors and used to calculate the ω index. Both R scripts take the item parameters, reestimated ability levels, and response vectors after answer copying as inputs and return the chance probability of matching response alternatives between two examinees' response vectors (see Appendix B for a numerical illustration regarding how to compute these indices). If the probability value is below an alpha level, the pair is assumed to be detected as an answer copying pair by the index. Alpha levels of .001, .01, and .05 were used in the research.

The data generation process was very similar for the Type I error rate analysis. It included all four steps as described above except the second step, simulating answer copying. Two ability levels for two independent hypothetical examinees were drawn from a uniform distribution within the C_t and S_p intervals, respectively. Then, two independent response vectors for a 40-item hypothetical test were simulated based on the nominal response model given the generated ability levels and the item parameters. The same process was replicated until 10,000 honest independent pairs of response vectors were obtained for Type I error analysis within the each level of 144 experimental conditions.

Appendix B

Computing the GBT and ω Indices

Both indices use the nominal response model to compute the probability of selecting the response alternative k of item i for person j:

$$p_{jik} = \frac{\exp(\zeta_{ik} + \lambda_{ik} * \theta_j)}{\sum_{k=1}^{m} \exp(\zeta_{ik} + \lambda_{ik} * \theta_j)},$$

where p_{jik} is the probability of choosing response alternative k of item i for person j, ζ_{ik} is the threshold parameter estimate for response alternative k of item i, λ_{ik} is the slope parameter estimate for response alternative k of item i, and θ_j is the ability estimate for person j.

When computing the ω index, the following quantities are first computed:

$$E_{cs} = \sum_{i=1}^{N} p_{cik},$$

$$s = \sqrt{\sum_{i=1}^{N} p_{cik} * (1 - p_{cik})},$$

where k is the response alternative chosen by the source examinee for item i, c is the cheater examinee with an ability level of θ_c, E_{cs} is the expected number of identical responses between two response vectors, and s is the standard error of the expected number. Finally, the z statistic is equal to

$$z = \frac{O_{cs} - E_{cs}}{s},$$

where O_{cs} is the observed number of identical responses between two response vectors. The z statistic is compared to a normal distribution using a criterion alpha level to test the hypothesis of independence between two response vectors.

The computations are more complex for the GBT index. First, the probability of choosing the same response alternative simultaneously by the source and the cheater examinees is defined as the following:

$$P_i = \sum_{a=1}^{k} p_{cik} * p_{sik},$$

where P_i is the probability of matching on item i for two examinees, p_{cik} and p_{sik} are the probabilities of selecting response alternative k of item i for the cheater examinee with the ability level of θ_c and source examinee with the ability level of θ_s, respectively. Then, the probability of observing m matches in N *items* is computed as following:

$$f_N(m) = \sum \left(\prod_{i=1}^{N} P_i^{u_i} Q_i^{1-u_i} \right),$$

where u_i is equal to 1 if the responses match and 0 if the responses do not match on item i, and Q_i is equal to $1 - P_i$. The summation is over all possible combinations of m matches in N items. For instance, the probability of observing two matches in a three-item test would be

$$P_1P_2Q_3 + P_1Q_3P_3 + Q_1P_2P_3.$$

Finally, the GBT index is computed as the probability of observing m or more matches in N items:

$$GBT = \sum_{t=m}^{N} f_N(t).$$

Numerical Illustration

In the following tables, Table A1 illustrates the nominal response model item parameters for five hypothetical items, and Table A2 illustrates the observed responses of two hypothetical examinees as well as the probability of selecting each response option for these examinees.

Table A1. Nominal Response Model Item Parameters for Five Hypothetical Items

	Slope estimates					Threshold estimates				
	A	B	C	D	E	A	B	C	D	E
Item 1	−.831	−.754	.080	**1.549**	−.043	.016	−1.394	.430	**1.716**	−.766
Item 2	−.336	−.239	**1.264**	−.631	−.059	−.055	−.188	**1.635**	−.443	−.948
Item 3	−.762	−.085	**1.068**	−.898	.678	−.367	.873	**1.693**	−1.547	−.650
Item 4	−.774	−.201	−.292	**1.896**	−.631	−.995	−.484	1.335	**1.165**	−1.023
Item 5	−.319	−.114	−.276	−.401	**1.111**	−.311	−.144	−.196	−1.176	**1.829**

Note: Bold items are the key responses for the corresponding items.

Table A2. Probability of Selecting Response Options for Each Examinee Based on the Ability Level Estimated from Observed Response Pattern

	True ability level	Estimated ability level[a]		Probability of choosing response option					Observed response
				A	B	C	D	E	
Examinee 1 (cheater)	−1.484	−.731	Item 1	.310	.072	.241	.298	.080	A
			Item 2	.214	.174	.360	.180	.072	B
			Item 3	.173	.365	.357	.059	.046	C
			Item 4	.087	.096	.632	.108	.077	C
			Item 5	.153	.156	.166	.068	.457	E
Examinee 2 (source)	−.478	−.445	Item 1	.224	.053	.226	.425	.072	D
			Item 2	.178	.149	.472	.137	.064	B
			Item 3	.129	.330	.448	.042	.051	C
			Item 4	.071	.091	.587	.187	.065	C
			Item 5	.123	.133	.136	.054	.554	C

[a] Maximum Likelihood ability estimate is obtained by using MULTILOG.

The ω Index

The probability of selecting the second examinee's (source) response options by the first examinee (cheater) is

$$E_{cs} = .298 + .174 + .357 + .632 + .166 = 1.627.$$

The expected number of match is 1.627. The variance of this estimate is

$$.298 * (1 - .298) + .174 * (1 - .174) + .357 * (1 - .357) \\ + .632 * (1 - .632) + .166 * (1 - .166) = .953.$$

The observed number of matches is three; therefore, the ω index is

$$\omega = \frac{3 - 1.627}{\sqrt{.953}} = 1.40.$$

The ω index is compared with the critical values of 1.64, 2.32, and 3.09 for the alpha level of .05, .01, and .001, respectively.

The GBT Index

The probabilities of matching for two examinees on the first, second, third, fourth, and fifth items are the following:

$$P_1 = (.310 * .224) + (.072 * .053) + (.241 * .226) + (.298 * .425) + (.080 * .072) = .26$$

$$P_2 = (.214 * .178) + (.174 * .149) + (.360 * .472) + (.180 * .137) + (.072 * .064) = .26$$

$$P_3 = (.173 * .129) + (.365 * .330) + (.357 * .448) + (.059 * .042) + (.046 * .051) = .31$$

$$P_4 = (.087 * .071) + (.096 * .091) + (.632 * .587) + (.108 * .187) + (.077 * .065) = .41$$

$$P_5 = (.153 * .123) + (.156 * .133) + (.166 * .136) + (.068 * .054) + (.457 * .554) = .32$$

The probabilities of 0, 1, 2, 3, 4, or 5 matches between two response vectors in the hypothetical five-item test are the following:

$$f_5(0) = (1 - .26)(1 - .26)(1 - .31)(1 - .41)(1 - .32) = .152$$

$$\begin{aligned} f_5(1) = {} & [(.26)(1 - .26)(1 - .31)(1 - .41)(1 - .32)] + [(1 - .26)(.26)(1 - .31)(1 - .41)(1 - .32)] \\ & + [(1 - .26)(1 - .26)(.31)(1 - .41)(1 - .32)] + [(1 - .26)(1 - .26)(1 - .31)(.41)(1 - .32)] \\ & + [(1 - .26)(1 - .26)(1 - .31)(1 - .41)(.32)] = .351 \end{aligned}$$

$$\begin{aligned} f_5(2) = {} & [(.26)(.26)(1 - .31)(1 - .41)(1 - .32)] + [(.26)(1 - .26)(.31)(1 - .41)(1 - .32)] \\ & + [(.26)(1 - .26)(1 - .31)(.41)(1 - .32)] + [(.26)(1 - .26)(1 - .31)(1 - .41)(.32)] \\ & + [(1 - .26)(.26)(.31)(1 - .41)(1 - .32)] + [(1 - .26)(.26)(1 - .31)(.41)(1 - .32)] \\ & + [(1 - .26)(.26)(1 - .31)(1 - .41)(.32)] + [(1 - .26)(1 - .26)(.31)(.41)(1 - .32)] \\ & + [(1 - .26)(1 - .26)(.31)(1 - .41)(.32)] + [(1 - .26)(1 - .26)(1 - .31)(.41)(.32)] = .319 \end{aligned}$$

$$\begin{aligned} f_5(3) = {} & [(.26)(.26)(.31)(1 - .41)(1 - .32)] + [(.26)(.26)(1 - .31)(.41)(1 - .32)] \\ & + [(.26)(.26)(1 - .31)(1 - .41)(.32)] + [(.26)(1 - .26)(.31)(.41)(1 - .32)] \\ & + [(.26)(1 - .26)(.31)(1 - .41)(.32)] + [(.26)(1 - .26)(1 - .31)(.41)(.32)] \\ & + [(1 - .26)(.26)(.31)(.41)(1 - .32)] + [(1 - .26)(.26)(.31)(1 - .41)(.32)] \\ & + [(1 - .26)(.26)(1 - .31)(.41)(.32)] + [(1 - .26)(1 - .26)(.31)(.41)(.32)] = .143 \end{aligned}$$

$$\begin{aligned} f_5(4) = {} & [(.26)(.26)(.31)(.41)(1 - .32)] + [(.26)(.26)(.31)(1 - .41)(.32)] \\ & + [(.26)(.26)(1 - .31)(.41)(.32)] + [(.26)(1 - .26)(.31)(.41)(.32)] \\ & + [(1 - .26)(.26)(.31)(.41)(.32)] = .032 \end{aligned}$$

$$f_5(5) = (.26)(.26)(.31)(.41)(.32) = .003$$

The probability of observing three or more matches between two response vectors is

$$\mathrm{GBT} = .143 + .032 + .003 = .178$$

The GBT index can be compared with .05, .01, or .001 to test the independence of two response vectors.

Authors' Note

Additional specific tables for each simulation condition and figures are available on the website www.tc.umn.edu/~zoplu001

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

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

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). *Detection of cheating on multiple-choice examinations*. Paper presented at the annual 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.

Bernardi, R. A., Baca, A. V., Landers, K. S., & Witek, M. B. (2008). Methods of cheating and deterrents to classroom cheating: An international study. *Ethics & Behavior, 18*, 373-391.

Bird, C. (1927). The detection of cheating in objective examinations. *School and Society, 25*, 261-262.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 46*, 443-459.

Bopp, M., Gleason, P., & Misicka, S. (2001). *Reducing incidents of cheating in adolescence* (Master's project). 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.

Buss, W. G., & Novick, M. R. (1980). The detection of cheating on standardized tests: Statistical and legal analysis. *Journal of Law and Education, 9*, 1-64.

Cody, R. P. (1985). Statistical analysis of examinations to detect cheating. *Journal of Medical Education, 60*, 136-137.

Dickenson, H. (1945). Identical errors and deception. *Journal of Educational Research, 38*, 534-542.

Diekhoff, G. M., LaBeff, E. E., Shinohara, K., & Yasukawa, H. (1999). College cheating in Japan and the United States. *Research in Higher Education, 40*, 343-353.

Frary, R. B. (1993). Statistical detection of multiple-choice answer copying: Review and commentary. *Applied Measurement in Education, 6*, 153-165.

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics, 2*, 235-256.

Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying*. Iowa City, IA: American College Testing.

Harwell, M. R. (1997). Analyzing the results of Monte Carlo studies in item response theory. *Educational and Psychological Measurement, 57,* 266-279.

Harwell, M. R., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-125.

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). *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

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269-290.

Lewis, C., & Thayer, D. T. (1998). *The power of the K-Index to detect copying* (ETS Research Report No.98-49). Princeton, NJ: Educational Testing Service.

Lin, C. S., & Wen, L. M. (2007). Academic dishonesty in higher education: A nationwide study in Taiwan. *Higher Education, 54*, 85-97.

Lupton, R. A., & Chapman, K. J. (2002). Russian and American college students' attitudes, perceptions, and tendencies towards cheating. *Educational Research, 44*(1), 17-27.

McCabe, D. L., & Trevino, L. K. (1997). Individual and contextual influences on academic dishonesty: A multicampus investigation. *Research in Higher Education, 38*, 379-396.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105-142). New York, NY: Academic Press.

R Development Core Team. (2009). *R: A language and environment for statistical computing* (ISBN 3-900051-07-0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rakovski, C. C., & Levy, E. S. (2007). Academic dishonesty: Perceptions of business students. *College Student Journal, 41*, 466-481.

Saupe, J. (1960). An empirical model for the corroboration of suspected cheating on multiple-choice tests. *Educational and Psychological Measurement, 20*, 475-489.

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.

Spiegel, P. K., Tabachnick, B. G., Whitley, B. E., & Washburn, J. (1998). Why professors ignore cheating: Opinions of a national sample of psychology instructors. *Ethics & Behavior, 8*, 215-227.

Thissen, D. (2003). Multiple categorical item analysis and test scoring using item response theory [Computer Software]. Chicago, IL:Scientific Software.

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.

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.

Whitley, B. E. (1998). Factors associated with cheating among college students: A review. *Research in Higher Education, 39*, 235-274.

Wollack, J. A. (1996). Detection of answer copying using item response theory. *Dissertation Abstracts International, 57/05*, 2015.

Wollack, J. A. (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement, 40*, 189-205.

Wollack, J. A., & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22*, 144-152.