Computer Software Review

Cheating Detection at Your Fingertips: A Review of INTEGRITY

James A. Wollack, University of Wisconsin-Madison

Description

Research on cheating detection, most notably the detection of answer copying, has gained in popularity during the past decade, resulting in a number of statistical indices that have been shown to work quite well at identifying the most serious offenders. However, in spite of the existence of these methods and in spite of the widespread attention that cheating on tests has drawn in recent years, all too often these indices are not applied in practice, even when answer copying is suspected. The most likely reasons for this are that (a) the methods are not known to exist and (b) the methods are not sufficiently easy to compute. This is particularly true in classroom situations, where teachers lack the time, resources, and expertise to keep abreast of the literature on copying detection methods.

Although making known the existence of statistical approaches to detect answer copying remains a marketing challenge, Castle Rock Research Corporation has gone a long way toward making statistical detection of answer copying a viable option for classroom teachers. Castle Rock Research Corporation (2005b) now offers a service called INTEGRITY that performs a number of analyses to evaluate the statistical integrity of the test and the student's test scores. INTEGRITY provides item analysis and subgroup and test center comparisons that are as detailed and prescriptive as those provided by any of the other test analysis packages on the market. However, what clearly offsets INTEGRITY from other packages is its ability to perform collusion detection.

Collusion Detection in INTEGRITY

At the click of a radial button, INTEGRITY performs five different answer copying detection indices: B-index (Angoff, 1974); PAIR1 and PAIR2 (Hanson, Harris, & Brennan, 1987); modified error similarity analysis (MESA), based on the error similarity analysis (ESA) method (Bellezza & Bellezza, 1989); and g_2 (Frary, Tideman, & Watts, 1977). INTEGRITY does not allow users to input a seating chart, but users do specify whether they would like to investigate all possible pairs or target their investigation to consider only specific examinees (although the same critical values, or CVs, are used in both scenarios, even though the probability of falsely detecting an examinee is considerably higher when all pairs are computed). The collusion detection report identifies all pairs for whom at least one of the copying indices was deemed significant and provides the value of the test statistics for any significant indices, along with one of three qualitative interpretations—low, moderate, or high—indicating the "statistical certainty that the pair of examinees engaged in collusion" (Castle Rock Research Corporation, 2005c, p. 4).

Applied Psychological Measurement, Vol. 31 No. 3, May 2007, 233–239
DOI: 10.1177/0146621607300853

Much of what is offered in the collusion detection segment is very helpful. Reports are organized well, and the presentation of data makes the information very accessible. The reports are organized as a set of documents interconnected by hyperlinks. The initial report provides summary information, but it is possible to receive more detailed information (e.g., a visual comparison of the answers provided by a pair of examinees, a histogram showing the distribution of any one of the five detection indices, a listing of all examinee pairs identified by any one detection index, descriptions of the indices, CVs, criteria for the qualitative interpretations, etc.) by clicking on the appropriate buttons and links.

INTEGRITY is very fast. Submitting a data set for analysis takes about a minute. The data set must be comma delimited, and the answer key must be provided in a separate file. Sample files are provided to assist in getting the data into the proper format. Once submitted to the INTEGRITY service, the average time required to complete item analysis, subtest analysis, and collusion detection for all possible pairs on an 80-item test was 5 s for data sets with 100 examinees and 52 s for 1,000 examinees. Even when a data set with 10,000 examinees was used, results were available in just under 45 min. Times were appreciably shorter if either subtest analyses or collusion analyses were not required.

INTEGRITY provides teachers the means to easily test examinee pairs easily for answer copying on tests. Thus, INTEGRITY certainly helps eliminate access and difficulty computing as major barriers to performing collusion detection. Still, two aspects of INTEGRITY remain troublesome. The primary concern is that the procedures used in INTEGRITY do not work particularly well at identifying known copiers. Consequently, many examinees who copy large numbers of questions may go undetected, resulting in some examinees having test scores of questionable validity and some teachers having a false sense of security. Routinely failing to detect copiers may also cause frustration in teachers who have strong evidence of copying from other sources and look to copying indices as the final piece of corroborating evidence. The second concern is that INTEGRITY places an enormous amount of data at the user's fingertips yet gives little direction on how to use this information to make smart, legally defensible assertions that a particular examinee engaged in answer copying. These two concerns are addressed more fully below.

INTEGRITY Indices and Detection Rates

There exist many different indices for copying detection, and not all are created equally. Research by Hanson et al. (1987), Sotaridona and Meijer (2002, 2003), and Wollack (2003, 2006), among others, has shown that indices vary widely in terms of their statistical properties. Several indices tend to have inflated Type I error rates. Others have the opposite problem of being overly conservative, which, though not a problem to the same degree, often results in low detection rates. Even after controlling for differences in Type I error rate, as is the case when empirical CVs are used, indices still vary with respect to their statistical power. What tends not to vary across these studies, however, is the collection of detection indices that perform best at identifying known or simulated copiers. Curiously, none of those "best" indices are used by INTEGRITY. In fact, the primary criticism of the computer program Scrutiny! (Assessment Systems Corporation, 1993)—the only commercially available competition to INTEGRITY—is that it adopts an ESA algorithm that has very low power to detect true copiers. Indices such as ω (Wollack, 1997), S_2 and S_1 (Sotaridona & Meijer, 2003), and if strings-based copying is suspected, H (Angoff, 1974), would have been preferred to the indices used.

A related concern centers on the way INTEGRITY determines CVs for each of the five detection indices. Theory suggests that MESA values can be interpreted directly as probabilities, whereas B-index and g_2 values should be approximately distributed according to the standard

Table 1
Critical Values (and *p* Values) for INTEGRITY Collusion Indices

	B-Index	MESA[a]	g_2
Low	7.0 (1.28×10^{-12})	1.6×10^{-5}	5.75 (4.46×10^{-9})
Moderate	8.0 (6.22×10^{-16})	1.6×10^{-8}	6.25 (2.05×10^{-10})
High	9.0 (1.13×10^{-19})	1.6×10^{-10}	7.25 (2.08×10^{-13})

Note. MESA = modified error similarity analysis. Critical values (CVs) for MESA become gradually more extreme as test length increases to $n = 24$ items, where they stabilize. The CVs provided are for $n \geq 24$.
a. MESA CVs are directly interpretable as *p* values.

normal. However, empirical work has shown that these methods tend not to hold the nominal Type I error rate well. Two suggestions have been made to address this concern: Either use empirical CVs (Hanson et al., 1987) or disregard those indices with liberal error rates and suffer the natural consequences (i.e., reduced power) of those with conservative error rates (Wollack, 1997). PAIR1 and PAIR2 do not have well-known theoretical sampling distributions; consequently, empirical CVs must be set prior to their use.

The procedure for setting empirical CVs is relatively straightforward. This process typically involves analyzing real or simulated data known to come from the null hypothesis (H_0) and finding the value of the test statistic that falls at precisely the $100(1-\alpha)$th percentile, where α is the desired rate at which H_0 will be erroneously rejected. The location of a CV is inextricably connected to a tolerance for false-positive results. With INTEGRITY, this necessary link between CVs and Type I error rates appears to be absent. Furthermore, the empirical CVs for INTEGRITY are independent of the test data. For each index, the same CVs are used for all data sets, regardless of the characteristics of the data set being analyzed. The online documentation on interpreting results from collusion detection analyses says that CVs "were set based on theoretical attributes of each method, peer-reviewed research literature, and estimated false-positive versus true-positive rates" (Castle Rock Research Corporation, 2005c, p. 4). Yet, theoretical interpretations of the CVs (at least for those indices having well-known theoretical sampling distributions) are very different, and based on empirical results, it is clear that that the number of false positives differs dramatically across methods. To illustrate this point, consider the CVs and critical *p* values (given in Table 1) corresponding to low, moderate, and high probabilities of cheating for the B-index, g_2, and MESA indices. Based solely on the indices' theoretical distributions, the CVs do not appear comparable. In fact, if the theoretical distributions hold, it is considerably easier to be detected as a high-likelihood cheater under MESA than as a low-likelihood cheater using the B-index.

Empirical Type I error rates further suggest that the CVs are not equivalent across indices. Empirical Type I error rates were investigated by reanalyzing actual test data from Wollack (2006). The test data were collected in a manner such that copying could not have occurred, but copying was simulated for 8% of the examinees (such that equal numbers of examinees were made to copy 10%, 20%, 30%, or 40% of the items from a single source). Estimates of Type I error rates were based on all pairs of examinees for whom copying was not simulated.

Three different sample sizes were studied: 100, 1,000, or 10,000 examinees. For the two smaller sample size conditions, multiple unique data sets of the same size were submitted to INTEGRITY for analysis, such that a total of 10,000 examinees for each sample size were analyzed for answer copying. The number of replications for the 100- and 1,000-examinee conditions were 100 and 10, respectively. Only one 10,000-examinee data set was analyzed.

Table 2
Empirical Type I Error Rates[a] for INTEGRITY Detection Indices

Sample Size	Confidence Level	B-Index	PAIR1	PAIR2	MESA	g_2
100	Low	.00	.00	.00	.00	.00
	Moderate	.00	.00	.00	.00	.00
	High	.00	.00	.00	.00	.00
1,000	Low	.0000004	.0000004	.0000004	.0000024	.0000016
	Moderate	.0000002	.0000004	.0000004	.0000012	.0000012
	High	.0000002	.0000002	.0000004	.0000002	.0000011
10,000	Low	.00	.0000001	.00	.0000014	.0000009
	Moderate	.00	.0000001	.00	.0000005	.0000007
	High	.00	.0000001	.00	.0000002	.0000004
Average	Low	.0000001	.0000002	.0000001	.0000013	.0000008
	Moderate	.0000001	.0000002	.0000001	.0000006	.0000006
	High	.0000001	.0000001	.0000001	.0000001	.0000005

Note. MESA = modified error similarity analysis.
a. Because the B-index, PAIR1, PAIR2, and MESA produce the same index, regardless of which examinee in a pair is treated as the copier, the empirical Type I error rates for $N = 100$ are based on 493,400 pairs. Type I error rates for $N = 1{,}000$ are based on 4,993,400 pairs. Type I error rates for $N = 10{,}000$ are based on 49,993,400 pairs. For g_2, which produces different values depending on which examinee is treated as the copier, empirical Type I error rates for $N = 100$ are based on 988,400 values. Type I error rates for $N = 1{,}000$ are based on 9,988,400 values. Type I error rates for $N = 10{,}000$ are based on 99,988,400 values.

Empirical Type I error results are given in Table 2. From these data, three things appear clear. First, the CVs that were selected do not result in comparable interpretations of low, moderate, or high confidence of copying across the five indices. As an example, with 1,000 examinees, the low CV for MESA resulted in six times more false positives than the low CV for the B-index, PAIR1, or PAIR2. Also, for larger sample sizes, MESA and, to a lesser extent, g_2 showed, as expected, decreasing numbers of Type I errors as the confidence standard increased from low to high. However, with the other three indices, such differentiation was not always present. In fact, within each of the three sample size conditions, the PAIR2 Type I error rates were identical for the high, moderate, and small categories. The second observation is that interpretations of CVs appear different for different sample sizes. For the 100-examinee condition, none of the five indices falsely identified even a single noncopying examinee at any of the three confidence levels. Type I error rates were more similar for the two larger sample sizes but appeared somewhat larger for the 1,000-examinee condition. Finally, the CVs selected result in excessively stringent false-positive rates. Even at the low confidence level, the highest observed false-positive rate of any index in any condition was .0000016 (for g_2 with $N = 1{,}000$). If one were to adopt a personwise Type I error rate perspective (i.e., controlling the overall probability of any particular noncopying examinee being falsely detected), as has been suggested by Wollack, Cohen, and Serlin (2001) and Wollack (2006), the maximum Type I error rate was .0000032 (for low confidence with $N = 1{,}000$). In both cases, these maximum Type I error rates are well below Angoff's (1974) proposed 1 in 10,000 criterion, which remains the most conservative copying criterion published.

In light of the conservative Type I error rates, it is not surprising that the power of INTEGRITY is less than adequate. Table 3 shows the percentage of examinees copying 40% of the items who were correctly identified (along with their source examinee) by the INTEGRITY indices for each of the three confidence levels. The best-performing index was PAIR2, which averaged detecting just more

Table 3
Detection Rates[a] for Examinees[b] Copying 40% of the Items

Sample Size	Confidence Level	B-Index	PAIR1	PAIR2	MESA	g_2	Any Index
100	Low	.000	.140	.150	.045	.040	.210
	Moderate	.000	.020	.105	.005	.035	.130
	High	.000	.000	.050	.000	.005	.050
1,000	Low	.000	.210	.240	.030	.075	.300
	Moderate	.000	.025	.115	.010	.025	.110
	High	.000	.000	.040	.000	.010	.035
10,000	Low	.000	.210	.220	.025	.070	.275
	Moderate	.000	.040	.125	.000	.040	.125
	High	.000	.005	.065	.000	.000	.060
Average	Low	.000	.187	.203	.033	.062	.262
	Moderate	.000	.028	.115	.005	.033	.122
	High	.000	.002	.052	.000	.005	.048

Note. MESA = modified error similarity analysis.
a. All detection rates are based on 200 true copier-source pairs.
b. And the correct simulated source examinee.

than 20% of the most flagrant copiers at the low confidence level. The B-index did not identify a single true copier in any of the data sets analyzed. Considering all indices together, an average of only 26.2% of the examinees copying 40% of the items was detected with low confidence.

By contrast, when the ω index (Wollack, 1997) was applied to these same data, using theoretical CVs, based on a normal (0,1) distribution, corresponding to more conventional α levels of .01, .001, and .0005, the empirical false-positive rates were close to nominal (i.e., .0074, .0007, and .0003, respectively), and power to detect 40% copiers was strong (.90, .77, and .72). Based on previous research, results from other indices, such as S_1 and S_2 (Sotaridona & Meijer, 2003) should be similar to those of ω. Therefore, by selecting better behaved indices, INTEGRITY could have used theoretical CVs, thereby enabling easier interpretations of the confidence levels and simultaneously improving detection rates.

The Problem of Data Overload

By computing five indices simultaneously, INTEGRITY gives users a tremendous amount of information. However, aside from saying that an examinee should be scrutinized if any of the five indices is significant (Castle Rock Research Corporation, 2005a), INTEGRITY provides little insight into how to use all this information.

There is one clear advantage to performing multiple copying detection indices on a data set. Different indices may be differentially sensitive to different types or amounts of copying, so computing multiple indices may provide better coverage against all types and amounts of copying.

However, there are also reasons not to use multiple indices. Applying multiple indices tends to inflate the probability of incorrectly identifying an examinee as copying, unless the indices' CVs are adjusted. However, because detection rates decrease precipitously with extreme CVs, it is often possible to detect more individuals by using a single index with a less extreme CV than multiple indices with more extreme CVs, particularly if the indices are largely redundant (Wollack, 2006).

One perceived advantage is that computing multiple indices provides the opportunity to collect corroborating (and therefore stronger) evidence of copying. However, it is not clear that this is

true. Rejection of H_0 communicates information about the probability of the data given the null $P(D|H_0)$, not $P(H_0|D)$ (Cohen, 1994). Rejecting the H_0 twice, with two (perhaps substantially) dependent tests, still does not convey information about $P(H_0|D)$.

Finally, although the manual states that sufficient evidence of copying is provided if even a single index is statistically significant, it is not clear how this would be viewed by a court of law. INTEGRITY does not provide information on how the various indices differ with respect to the types or amounts of copying to which they are most sensitive. Without this information, it might be very difficult to convince a judge that the evidence for copying from the one significant index warrants stronger consideration than the evidence against copying from the four nonsignificant indices.

In INTEGRITY's defense, the service does afford the advanced user with the opportunity to select the indices to compute or provide alternative CVs for each of the indices. However, because the selected indices do not follow well-known distributions, users have little or no basis on which to select appropriate values. Instead of allowing users to specify CVs, INTEGRITY should allow users to specify desired false-positive rates, with the exact CVs to be determined by the program. This feature could then be used for one of two purposes. First, users could select a subset of indices with correctly adjusted CVs (which would be aided greatly by including documentation on the statistical characteristics of the indices under different types and amounts of copying). Alternatively, users could decide for themselves what criterion is acceptable. As it is, INTEGRITY provides only three levels of confidence, and users have no input into what constitutes low, moderate, or high. It would be better—and more in keeping with good science—for users to specify their criteria based on the amount of statistical evidence necessary for them to be comfortable concluding that copying occurred. In situations where the visual evidence of cheating is very strong, it may not be necessary for the statistical evidence to be overwhelming. In fact, if enough nonstatistical evidence exists, one may be willing to accept statistical findings with Type I error rates as high as .05 or .01.

Purchasing INTEGRITY

INTEGRITY is not a piece of software that can be loaded on a computer; it is a service offered through the Castle Rock Research Corporation Web site. Users can choose between a variety of licenses, depending on the number of data sets needing to be analyzed and the maximum number of examinees per data set. A standard 1-year single-user license, which limits users to 30 submissions and a maximum of 500 examinees per data set, costs $70. The annual per-user cost is reduced when purchasing licenses for multiple users or multiple years. Professional licenses are also available, offering up to 300 submissions and a maximum of 50,000 examinees per submission. Professional licenses are priced based on the total number of examinees for all submissions. Discounts are offered when a multiple-year license is purchased. Given that computing collusion detection indices requires either writing a computer program or purchasing specialized software, INTEGRITY is a good value and an important resource for anyone concerned about classroom cheating.

Conclusion

INTEGRITY is an innovative new service for analyzing test performance, highlighted by a component that investigates examinee score patterns for possible answer copying. The strengths of the INTEGRITY service are its speed, ease of use, and its interactive displays of summary data, which offers users numerous hot links on which to click for more specific breakdowns, additional

statistics, or explanations of the summary statistics. The item analysis component, which provides prescriptive comments to help test developers interpret the statistics and guide item revision, is extremely well designed and helpful.

What distinguishes INTEGRITY from all other item analysis packages is its answer copying detection capabilities. Like all of the INTEGRITY components, answer copying detection analyses are easy and quick to run and provide a nice set of clear and easy to navigate output tables. However, some aspects of INTEGRITY's collusion detection component could be improved. INTEGRITY would benefit by adopting procedures to standardize and simplify the interpretations of its evidence standards (i.e., CVs) and/or by utilizing indices that offer good power and have well-behaved sampling distributions. Other potential improvements include providing more detailed explanations of the indices and advice on best practices for using multiple indices in concert.

Answer copying detection services should be a staple of all high schools and university testing offices, for purposes of both preventing and detecting classroom cheating. At present, INTEGRITY is the best commercially available software for answer copying detection and offers a monumental leap forward in the never-ending quest to eliminate cheating on classroom exams.

References

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, *69*, 44-49.

Assessment Systems Corporation. (1993). Scrutiny!: Software to identify test misconduct [Computer software]. San Antonio, TX: Advanced Psychometrics.

Bellezza, F. S., & Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology*, *16*(3), 151-155.

Castle Rock Research Corporation. (2005a, January). *How INTEGRITY can help evaluate and address issues of academic integrity*. Edmonton, AB, Canada: Author.

Castle Rock Research Corporation. (2005b). *INTEGRITY* [Computer software]. Edmonton, AB, Canada: Author.

Castle Rock Research Corporation. (2005c, January). *Quick start guide to interpreting collusion detection results from INTEGRITY*. Edmonton, AB, Canada: Author.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997-1003.

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* (Research Rep. Series No. 87-15). Iowa City, IA: American College Testing.

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.

Wollack, J. A. (1997). A nominal response model approach to detect answer coping. *Applied Psychological Measurement*, *21*(4), 307-320.

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

Wollack, J. A. (2006). Simultaneous use of multiple answer copying indices to improve detection rates. *Applied Measurement in Education*, *19*, 265-288.

Wollack, J. A., Cohen, A. S., & Serlin, R. C. (2001). Defining error rates and power for detecting answer copying. *Applied Psychological Measurement*, *25*, 385-404.

Author's Address

Address correspondence to James A. Wollack, University of Wisconsin, 1025 W. Johnson St., #373, Madison, WI 53706; e-mail: jwollack@wisc.edu.