

# DETECTING ANSWER COPYING ON HIGH-STAKES TESTS

*by James A. Wollack*

Cheating is a potentially serious problem on high-stakes tests, such as the Multistate Bar Exam (MBE), because it can result in under-qualified individuals receiving their Board's certification to practice, thereby undermining the certification process. Because cheating on exams has such serious ramifications, it is important for test administrators to minimize cheating opportunities and detect it when it occurs. Statistical indices such as the  $\omega$  index can be a useful tool in determining whether an applicant copied answers.

In high-stakes testing programs, candidates are required to register ahead of time for exams. They may be required to provide a signature, picture, or thumbprint identification to receive a test, and possibly again to hand one in. They are assigned to spaced-out seats according to a seating chart, told precisely what materials, if any, they are allowed to use or have available, and instructed to remove all hats and sunglasses while testing. Exam materials are kept in a locked cabinet prior to testing, and are returned there after testing is completed. Exam booklets and answer sheets are often numbered. With some testing programs, test forms will be distributed so that every other candidate receives an alternate form of the test containing different items or items in scrambled locations. The exams are administered under standardized conditions, with strict time

limits, in a heavily monitored environment, possibly including the use of video surveillance equipment. And yet, some examinees still manage to cheat.

Many types of cheating are difficult to detect. Candidates who somehow obtained a copy of an exam booklet or an answer key in advance cannot be identified during the examination without additional information. Candidates using notes or other unauthorized materials will hopefully be spotted by a proctor, but if they are sufficiently discreet, may well escape notice. Impersonators, examinees who pretend to be the registered candidates for purposes of taking an exam, can be very hard to identify if the forged identification is well done.

At first glance, answer copying may appear to be a very difficult form of cheating to conclusively identify. Unlike other methods of cheating, answer copiers don't have illegal materials that could be confiscated and used as evidence of cheating. Even if a proctor reports that the examinee was clearly looking at a neighboring candidate's answers, if the accused examinee simply denies having cheated, it would appear to boil down to a case of "he said-she said," where the credibility of the examinee must be weighed against the accuracy of the proctor. Such visual evidence, absent any other evidence, is often not strong enough to justify charging a candidate with something as serious as cheating.

However, unlike most forms of cheating, examinees who copy answers to test questions from neighboring examinees may leave behind on their answer sheets a collection of evidence indicating misconduct. This evidence appears in the form of unusual answer similarities between a pair of examinees seated beside one another. Uncovering and interpreting this evidence is not an easy task. Over the past several decades, researchers have developed a number of psychometric models and statistical indices for analyzing patterns of unusual answer similarities to determine whether a particular pattern is sufficiently unlikely to conclude that misconduct occurred.

In the vast majority of cases in which statistical indices to detect answer copying are used, their purpose is to corroborate an independent belief that a particular examinee copied from a specific source examinee. The decision to analyze response patterns statistically is the result of a detailed investigation into the facts of the case, including whether the examinees were seated sufficiently close to one another and whether anyone actually witnessed the alleged copying. In most cases, response patterns for the suspected copier and source examinees will be visually compared to see whether the two examinees appear to share a greater-than-expected number of item responses in common. If the result of this investigation is that it was unlikely or impossible for copying to have occurred, regardless of the amount of overlap in responses, an answer copying detection index will not be computed. If, however, the conclusion is that copying was possible or probable, then an answer copying detection index may be computed between the pair of examinees in question.

The rest of this article is devoted to a detailed discussion of copying detection analyses and how performing such analyses can bolster an allegation of

answer copying. This article will not concentrate on the mathematics of copying detection indices, as that information is readily available in the educational measurement literature. Instead, the focus of this discussion will be on interpreting answer copying indices, describing what they communicate and what they do not communicate, and understanding how such indices should be used in practice.

## THE LOGIC OF COPYING DETECTION INDICES

Any two independently working examinees will produce item response patterns with some amount of overlap. Sharing of correct answers is quite common, particularly among two examinees who both know the material well, as might be expected on a certification test. Even among examinees with less knowledge, it is quite likely that they will both produce correct answers to many of the easier questions. Similarly, any two examinees are expected to provide identical incorrect responses to certain items. Well-written items often have distractors (i.e., incorrect alternatives) that are attractive to people with partial or limited knowledge. Sometimes, for hard items, particular distractors are selected more frequently than the key itself. Therefore, two examinees who select the same distractor for an item would not be uncommon.

However, although matching answers on some items is to be anticipated, certain types of answer matches remain unlikely. Two high-achieving examinees would be unlikely to both select the same low frequency distractor. Similarly, low-achieving examinees would be unlikely to both correctly answer a very difficult item. Isolated occurrences of unlikely matches is not cause for concern, but finding many such matches is uncommon.

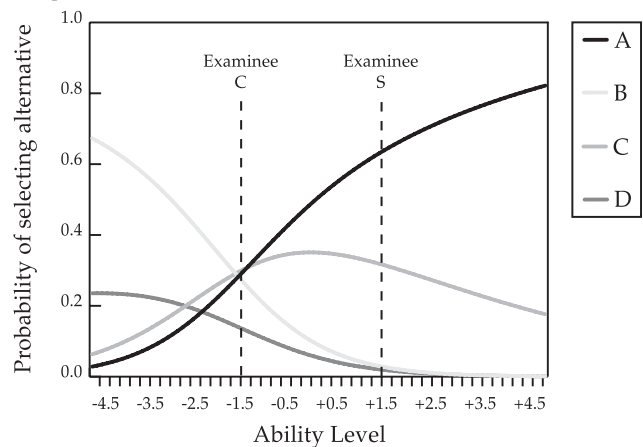
Several statistical indices have been proposed for detecting cases where the amount of answer overlap between two particular examinees is unusually large. Statistical indices essentially work by comparing the amount of overlap between two examinees to the normal amount that would be expected if the two examinees were known to have answered independently of each other. One of the most flexible indices available for copying detection is the  $\omega$  index (Wollack, 1997).

The  $\omega$  index compares the number of answer matches between a pair of examinees with the number of matches expected due to chance alone. The probability of an answer match on any given item is taken as the probability of an examinee with the alleged copier's ability<sup>1</sup> selecting the answer provided by the alleged source. This probability is estimated using a complex measurement model known as the nominal response model (NRM; Bock, 1972). The NRM yields probabilities of examinees of different ability levels<sup>2</sup> selecting each of the choices for a multiple-choice item. An example of the probability function for one four-alternative item is shown in Figure 1. The NRM is used to estimate examinee ability level and the probability curves for all examinees and all items.

The correct answer for the item shown in Figure 1 is (A). As should be true of all good multiple choice items, the probability of answering this item correctly increases with examinee ability. For examinees with abilities above about -1.5, (A) is the most likely choice, followed by (C), (B), and (D). For examinees with abilities below -1.5, however, the patterns are very different. For examinees with very low ability, less than -3.0, (B) is easily the most likely choice, followed by (D), (C), and then (A). Examinees with abilities between -3.0 and -1.5 have a different pattern

yet. As an example of how these probabilities are used, according to the model, an examinee S (source), with an ability level of 1.5 on the sample question would have about a 65% chance of selecting (A) and about a 30% chance of selecting (C). The probabilities that Examinee S would select (B) or (D) are both very low. For examinee C (copier), with an ability of -1.5, on the other hand, alternatives (A), (B), and (C) are roughly equally probable—each has about a 30% chance of being selected—while the probability of selecting (D) is only about 10%.

**Figure 1.** Nominal Response Model Probabilities of Correct Response for an Item



In computing  $\omega$ , these probability curves based on ability are estimated for every item. The probability of an examinee with the alleged copier's ability selecting each of the different alternatives for an item may be found directly from the probability curves. However, answer copying indices are sensitive to the number of times, relative to the expected number of times, that the alleged copier produced an answer that matched the alleged source's answer. Therefore, for any given item, the probability of an answer match is found from the particular curve corresponding to the alternative selected by the alleged source. The point on that curve, at the ability level of the alleged copier,

is taken as the probability of an answer match. Therefore, if C is the alleged copier and S the alleged source, and S selected alternative (A) to this item, although the probability of S selecting (A) is .65, the probability of an answer match, which depends on C's ability, is just .30.

## HOW THE $\omega$ INDEX IS COMPUTED

Regardless of the ability level of the alleged copier and the answer given by the alleged source, there will be some probability of an answer match for every item. For some items, it could be high, for some items it could be low. Prior to computing  $\omega$ , it is necessary to determine the probability of C selecting S's answer for all items on the test. The sum of these probabilities across all items equals the expected number of answer matches, EM(expected number of matches). To compute  $\omega$ , EM is subtracted from the actual number of answer matches between the two examinees, and the difference is divided by the standard error, which provides a measure of the amount of expected variability in observed number of answer matches. Therefore, the statistic is computed as follows:

$$\omega = \frac{\text{(number of matches)} - \text{EM(matches)}}{\text{standard error}}$$

The  $\omega$  index between any two individuals is computed by a computer program, based on the item strings for all examinees in a given test administration. For individuals who answer independently,  $\omega$  yields a value that is approximately normally distributed with a mean of 0 and a standard deviation of 1. It is advantageous that  $\omega$  follows this well known distribution because the probability of observing any particular  $\omega$  value, for independently working

examinees, is readily available from tables in the back of any basic statistics textbook. Large positive values will be associated with small probabilities that the amount of answer overlap occurred due to chance alone. Barring evidence to the contrary, the most likely reason for the overlap is that one examinee was copying from the other.

## INTERPRETING $\omega$ VALUES

As previously mentioned, tabled probability values from the normal distribution can be used to interpret the extremity of different  $\omega$  values. Obviously, smaller probabilities of observing an  $\omega$  value as extreme as the one observed are associated with more confidence that the amount of answer similarity was due to something other than chance. Because accusing someone of copying on a test is very serious and could have severe ramifications, only  $\omega$  values that are very unlikely to have occurred by chance should be considered as evidence of copying. Unfortunately,  $\omega$  will not be large for all examinees who copy, so by regarding only large  $\omega$  values as evidence of copying, it is possible that some examinees who actually copied will not be identified statistically.

A useful way to interpret the magnitude of  $\omega$  is to consider the percentage of cases in which independently working examinees would produce  $\omega$  values that are greater than or equal to a specified value. This percentage, which is available from statistical tables, is called the false positive rate because it identifies the percentage of  $\omega$  indices that can be expected to be flagged as high, even though no copying occurred between the examinees. Higher values of  $\omega$  coincide with smaller false positive rates. An example of how different  $\omega$  values and their associated false positive rates can be used to establish evidence of answer copying is provided in Table 1.

**Table 1.** Interpretations of different values of  $\Omega$

<b>Level</b>	<b>False Positive Rate</b>	<b><math>\Omega</math></b>	<b>Interpretation</b>
1	>50	<0	Evidence against answer copying
2	5%-50%	0-1.64	Weak evidence
3	1%-5%	1.65-2.32	Some evidence
4	.1%-1%	2.33-3.08	Good evidence
5	.01%-.1%	3.09-3.72	Strong evidence
6	<.01%	>3.72	Very strong evidence

The cutoffs for the various levels in Table 1 coincide with increasing levels of confidence that the amount of answer similarity did not arise by chance.  $\Omega$  values less than 0 (Level 1) have a false positive rate of greater than 50%. That is, on average, at least half of the honest, noncopying examinee pairs will produce values above zero (i.e., Level 2 or above), so a Level 1 criterion will result in incorrectly identifying many examinees as copiers. Positive values of  $\Omega$ , on the other hand, correspond with false positive rates less than 50% and indicate that there was more observed answer similarity than was expected. As an example,  $\Omega$  values between 1.65 and 2.32 (Level 3) produce false positive rates of between one and five percent.  $\Omega$  values as large as 1.65 are expected to occur by chance approximately 5% of the time  $\Omega$  is computed between noncopying pairs, whereas  $\Omega$  values as large as 2.32 occur by chance just 1% of the time.  $\Omega$  values less than 1.65 (but greater than 0) correspond to false positive rates in excess of 5% (Level 2). Level 4 corresponds to false positive rates between 0.1% and 1%, Level 5 corresponds to false positive rates between 0.01% and 0.1%, and Level 6 corresponds to false positive rates less than 0.01%.

Assuming that the examinees in question were seated within copying distance and other possible explanations for the high similarity have been eliminated,

higher levels of  $\Omega$  correspond to stronger evidence that copying occurred.  $\Omega$  values in Level 1 suggest fewer answer matches than expected due to chance, and actually provide evidence that the examinee did not copy.  $\Omega$  values in Level 2 indicate that the two examinees share slightly more answers in common than would be expected on the basis of chance, but not so many more as to indicate that they could not reasonably be attributed to chance.  $\Omega$  values in Levels 2, 3, or 4 generally do not correspond to sufficiently high levels of copying evidence to pursue a copying allegation absent other evidence for copying, such as proctor reports. However, depending on the strength of the additional evidence,  $\Omega$  values at these levels may provide useful information in pursuing a copying allegation.  $\Omega$  values at Levels 5 and 6 provide very good evidence of copying and should definitely be useful in supporting a charge of answer copying.

Because the estimated number of answer matches is based on the alleged copier's ability estimate, it is important to realize that values of  $\Omega$  computed between a pair of examinees will be different, depending upon which examinee is treated as the copier and which as the source. Unfortunately, comparison of the two  $\Omega$  values in itself is often not useful for identifying which of the two examinees copied from the other. When one examinee copies from the other, both examinees will have an unusually large number of answer matches. The information from the non-matched items is often not sufficient to clearly distinguish which examinee was copying. The identification of which candidate is the source and which is the potential copier usually relies on observations by proctors.

For situations where the examinees completed different forms of the test, however, it may prove

fruitful to analyze examinees' scores under the different test keys to try to distinguish which examinee was the copier. In particular, it is often helpful to compare the difference between the percentage correct and the expected percentage correct scores (when scored with the correct key) for two different sets of items: (a) items which have same keyed responses in both forms, and (b) items which have different keyed responses in the two forms. Non-copying examinees would be expected to perform similarly (and close to expectation) on the two sets of items. Examinees who copied would likely score much better on set (a) than on set (b).

If there is no visual evidence by proctors and no means to collect other statistical evidence against the candidate, it is very difficult to assemble a compelling case for copying. As previously mentioned, although  $\Omega$  yields different values depending on which examinee is treated as the copier and which is treated as the source,  $\Omega$  by itself does not allow for identifying which examinee is culpable. Also, without a proctor having been witness to the copying, one cannot rule out the possibility of a false positive. Still, high  $\Omega$  values do indicate that the item responses were not generated independently and suggest that one of the candidates was copying from the other. These examinees may be flagged in a database and monitored more closely during future exams.

## STATISTICAL PROPERTIES OF $\Omega$

$\Omega$  is a well-researched index, which has repeatedly been shown to have good statistical properties (Sotaridona & Meijer, 2002, 2003; Wollack, 1997, 2003; Wollack & Cohen, 1998). Most notably, the percentage of examinee pairs falsely detected by  $\Omega$  is never greater than the pre-specified false positive rate. Stated differently, the theoretical probability of a non-copying pair producing a particular  $\Omega$  value

(found from normal distribution probability tables) has been shown to be a conservative estimate of the actual probability. So, as an example, an  $\Omega$  of 2.65, which has a tabled probability of 0.004, will occur between honest, independently working examinees less than 0.4% of the time in practice. Good control of the false positive rate is essential in a detection index because it suggests that the tabled probability values are appropriate.

It is very difficult, in practice, to know how often an index is successful at detecting a copier. However, this information is available through simulation studies, in which certain examinees are simulated to have copied a fixed percentage of items from another examinee. Results of simulation studies have shown that four variables impact the detection rates of copying indices: sample size, test length, percentage of items copied, and false positive rate.

Sample size, the number of individuals who take the tests and whose data are used to estimate parameters in the statistical models, appears to have a small impact on detection rates. For the index, sample sizes ranging from 50 to 20,000 have been studied and detection rates remain fairly stable for tests administered to at least 100 examinees. When fewer than 100 examinees are tested, the model parameters are poorly estimated and the detection rates drop off slightly. Fortunately, most high-stakes testing programs have the luxury of rather large sample sizes.

Test length plays an important role in copying detection. In general, copying detection indices are more successful at identifying true copiers on longer tests than on shorter tests. Research on  $\Omega$  has focused on tests with between 20 and 80 items.

Another variable that contributes strongly to the detection rate is the percentage of items copied.

Research on  $\omega$  has focused on copying between 10 and 40 percent of the test items. Examinees who copy relatively few items, e.g., 10 or 20 percent, will often go undetected, particularly on shorter tests. Examinees who copy at least 30% of the items, however, are usually identified. It is worth mentioning that examinees who copy 100% of the items from another examinee will almost assuredly be identified by  $\omega$ . The lone exception to this is for copier and source examinees who have nearly perfect scores. For such individuals, estimates of their abilities would be very high and the probability of matching correct answers would similarly be high.

The final variable affecting the detection rate is the false positive rate. Far fewer simulated copiers are detected at Levels 5 or 6 than at Levels 1, 2, or 3, regardless of sample size, test length, or percentage of items copied.

To illustrate how these variables affect the detection rates of  $\omega$ , consider the data shown in Table 2. This table, which is adapted from Figures 3 - 5 in Wollack (2003), shows the proportions of simulated copiers who were detected at Levels 3 - 6 for three test lengths and four percentages of items copied, averaged across all sample sizes considered. Note that the data within a column of Table 2 are cumulative. That is, examinees detected at a Level 6 standard would also be detected if a Level 3, 4, or 5 standard had

been set. Similarly, examinees detected at a Level 4 standard would also be detected at a Level 3 standard, though not necessarily at a Level 5 standard.

Another way to learn how well  $\omega$  works in practice is by examining case studies in which  $\omega$  has been used to make or support copying allegations.  $\omega$  has been used to identify copying on law exams in over a dozen instances. Of the ten cases that have come to completion,  $\omega$  provided Level 5 or 6 evidence in seven of these cases. The statistical evidence and the copying charges were upheld in nine of these ten cases; in the remaining case, the copying charges were eventually dropped. In three other cases,  $\omega$  produced Level 6 evidence of copying between a pair of examinees, but charges were never filed because the copying was not observed by proctors and no additional evidence of copying existed, so it was not possible to determine which examinee was the copier.

The intricacies of these individual cases are fascinating and nicely illustrate the variety of ways in which statistical evidence has been used to evaluate answer copying charges. Five such case studies are provided below. The studies shown here are not intended to be representative of all the cases involving  $\omega$  or of all cases for which  $\omega$  could be used; rather, they are presented because they help illustrate the types of problems for which statistical evidence can be an asset.

**Table 2.** Proportions of Simulated Copiers Detected by  $\omega$

	<i>20 Items</i>				<i>40 Items</i>				<i>80 Items</i>			
	<i>Percentage of Items Copied</i>				<i>Percentage of Items Copied</i>				<i>Percentage of Items Copied</i>			
	<i>10%</i>	<i>20%</i>	<i>30%</i>	<i>40%</i>	<i>10%</i>	<i>20%</i>	<i>30%</i>	<i>40%</i>	<i>10%</i>	<i>20%</i>	<i>30%</i>	<i>40%</i>
<b>Level 3</b>	.12	.32	.55	.71	.23	.53	.80	.93	.26	.63	.86	.96
<b>Level 4</b>	.03	.13	.30	.52	.08	.30	.58	.82	.10	.38	.72	.89
<b>Level 5</b>	.00	.02	.11	.27	.02	.10	.32	.60	.03	.17	.49	.76
<b>Level 6</b>	.00	.01	.02	.11	.00	.03	.13	.38	.01	.06	.30	.61

## CASE STUDY 1

After two proctors observed one examinee copying from another examinee on the Multistate Professional Responsibility Examination (MPRE), one of the proctors intervened by pushing the source's answer sheet under her test booklet so that her answers would no longer be visible to the copier. Copying indices were expected to be high for the items early in the test but low for the items late in the test. It was not known how many items the alleged copier had completed at the time of the intervention, but the intervention occurred roughly half-way through the testing period. Therefore, three different sets of  $\omega$  values were computed, one corresponding to the first 20 items and the last 30 items, one corresponding to the first 25 items and the last 25 items, and one corresponding to the first 30 items and the last 20 items. Overall, the examinee shared 30 of 50 responses in common with the source, including matching answers on 22 of the first 25 items. The three sets of  $\omega$  values are shown below:

$$\omega_{1-20} = 2.95 \text{ (Level 4)} \quad \omega_{21-50} = 0.08 \text{ (Level 2)}$$

$$\omega_{1-25} = 3.37 \text{ (Level 5)} \quad \omega_{26-50} = -0.84 \text{ (Level 1)}$$

$$\omega_{1-30} = 1.65 \text{ (Level 3)} \quad \omega_{31-50} = 0.09 \text{ (Level 2)}$$

The above three pairs represented good-to-strong evidence that the suspected examinee copied early in the test, but not later. This evidence is consistent with the proctor's report and her actions to prevent copying.

In addition, whereas the suspected source performed very similarly on the first and second halves of the test, scoring at the 56th and 54th percentiles respectively, the alleged copier performed much better on the first half of the exam than on the second half,

scoring at the 67th and 1st percentiles respectively. This provided further support for the proctor's observations. This candidate was charged with answer copying. When the candidate did not answer those charges, the score on the test was canceled.

## CASE STUDY 2

Multiple proctors observed an examinee copying on two separate exams, the MBE and a state-specific multiple choice test. For the MBE, the copying was only witnessed during the morning session. On the state test, the examinee produced responses matching those of the source on 38 of 50 items. The examinees matched responses on 94 of 200 items on the MBE, including matching on 69 of 100 items during the a.m. section.  $\omega$  indices were run for both the state test and MBE, as well as separate indices based on only the a.m. and p.m. sections of the MBE. The overall  $\omega$  value for the state test was 2.34, providing Level 4 evidence of copying. On the MBE, the overall  $\omega$  value was 8.00, providing Level 6 evidence of copying. Interestingly, the separate  $\omega$  indices for the a.m. and p.m. sections were 10.66 and 0.52 respectively. The value for the state test combined with the large discrepancy between  $\omega$  values based on the a.m. and p.m. tests corroborated the proctors' report. This candidate was charged with answer copying. When the candidate did not answer those charges, this charge was upheld.

## CASE STUDY 3

An examinee was witnessed copying on the MBE. The alleged copier shared 115 of 197 items in common with the alleged source examinee. The corresponding  $\omega$  statistic was 10.21;  $\omega$  statistics for the a.m. and p.m. sections were also at Level 6. Because the MBE uses scrambled forms, it was also possible to examine candidate's performance when the test was scored with the two different answer keys.



Discounting the items that had the same key on both forms, the alleged copier was more than twice as likely to get an item right when it was scored with the answer key for the alleged source's exam than when scored with the correct key. This candidate was charged with answer copying and later admitted having copied answers on the exam.

#### CASE STUDY 4

Multiple proctors observed an examinee copying answers on the MPRE from two different sources. The suspected copier shared 34 of 50 answers in common with one suspected source and 30 of 50 answers in common with the other. The  $\omega$  indices between the suspected copier and each of the suspected source examinees were 2.03 and 1.90, both providing Level 3 evidence, even after making a statistical correction for having performed multiple comparisons on the same examinee (Wollack, Cohen, & Serlin, 2001). However, because it was suspected that the examinee copied from two individuals, a separate  $\omega$  was computed that investigated whether the amount of similarity between the copier and either of the suspected sources was unusually large. The suspected copier shared 41 of 50 answers in common with either of the two source examinees. The  $\omega$  for this comparison was 2.31. Although this is also Level 3 evidence, it is interesting that the  $\omega$  index for this comparison—the one that most closely models what the proctors observed—is higher than either of the two individual indices. This candidate was charged with answer copying. When the candidate did not answer those charges, the candidate's scores were canceled.

#### CASE STUDY 5

Two examinees seated next to each other during both a state-specific multiple-choice test and the MBE produced what appeared to be an unusual amount of

answer similarity, raising suspicions of one examinee having copied from the other. Neither examinee was observed copying on either exam.

The candidates had identical responses on 48 of 50 items on the state test. Also, the two examinees matched answers on 139 of 200 questions on the MBE, including matching on 92 of 100 questions during the a.m. session. Separate  $\omega$  statistics were computed treating each candidate as a possible copier and the other as the source. For the state test, the two values were very similar and both provided Level 6 evidence. Similarly, for the MBE,  $\omega$  values provided Level 6 evidence for both examinees. These results indicated that it was extremely likely that one of the examinees copied from the other. However, as indicated previously, values alone are not useful for distinguishing which examinee is the copier.

Fortunately, the two candidates in question completed different forms of the MBE. A follow-up analysis of the candidates' performances under the two test keys revealed an interesting pattern. One candidate (Examinee A) answered 63% of the items correct using the appropriate test key, but only 18% correct using the wrong (i.e., scrambled) test key. The other candidate (Examinee B), however, answered 30% of the items correctly using the appropriate key and 41% correct using the wrong key. Therefore, Examinee B was more likely to get an item correct when it was scored with the answer key for Examinee A's exam. Furthermore, comparison of the percentages correct with the expected percentages correct (based on their overall ability levels) showed that Examinee A was performing as expected, but Examinee B performed much better than expected for those items that were keyed the same on both forms, and much worse than expected for those items that were keyed differently on the two forms. These data are shown in Table 3.

**Table 3.** Comparison of Examinees' Scores Under Two Test Keys

	<b>Observed % correct (same keys)</b>	<b>Expected % correct (same keys)</b>	<b>Observed % correct (different keys)</b>	<b>Expected % correct (different keys)</b>
<b>Examinee A</b>	.66	.61	.63	.63
<b>Examinee B</b>	.59	.32	.30	.39

The combination of the statistical results indicating that one of these examinees almost certainly copied from the other and the patterns of scores from subsequent analyses of the MBE produced sufficiently strong evidence to charge Examinee B with copying answers from Examinee A on both the state test and the MBE. Examinee B defaulted by not answering the charges in time, but subsequently admitted to the state Board that she had, in fact, copied answers from a neighboring examinee on the exams in question. Examinee B's test scores were canceled.


## CONCLUSION

Answer copying on high-stakes exams can seriously compromise the validity of candidates' test scores. Fortunately, statistical methods have been developed to identify whether a suspected examinee shares an unusual amount of similarity with a neighboring examinee. Statistical evidence can be a powerful tool in successfully charging candidates with copying because it establishes that the amount of similarity between the alleged copier and source is anomalous for independently working examinees, and then quantifies how unusual it is.

Copying indices can be very valuable tools, but one must proceed cautiously. It is important to remember that copying indices, like all statistical tools, are probabilistic in nature. That is, it is possible

that even very unlikely results could have occurred due to chance. Therefore, before charging a candidate with copying, a testing organization should consider any alternative explanations for any answer similarity (e.g., examinee being a source). Witnessing the copying is the best way to eliminate alternative explanations for the answer similarity.

One must balance whatever evidence exists that copying occurred against any counter-evidence that copying did not occur. In collecting counter-evidence, one must answer the following questions: Did the alleged copier omit many items that were not omitted by the source or answer incorrectly many items that were answered correctly by the source? Were the copier and source seated too far away for copying to have occurred? If multiple forms exist, did the alleged copier do much better when the test was scored with the correct key than with the incorrect key? If the counter-evidence is strong, one cannot have confidence that the high  $\omega$  value was caused by answer copying.

Copying indices are a last resort—they can be used only after the test has been administered and the testing agency suspects that someone's score is spurious. Having available statistical tools to detect suspected copying is important, but exam developers and administrators must continue to proactively address the copying problem by creating a testing environment that will, to as large an extent as possible, discourage and prevent copying. Examples of essential proactive measures include an adequate number of diligent proctors or a surveillance system, assigned seating, check-in security, well-spaced seating, and alternate or scrambled test forms. 

## ACKNOWLEDGEMENTS

Portions of this paper were presented at the 2004 National Conference of Bar Examiners Spring Seminar in New Orleans.

The author would like to thank Mike Kane, John McAlary, and the Editor for many helpful comments and suggestions on early drafts of this paper.

## ENDNOTES

1. Ability is a generic word referring to the amount of mastery of the trait of interest, based on an examinee's overall performance on the test of interest. It is not intended to refer to general aptitude.
2. The convention with the NRM is to assume that ability of all examinees is normally distributed with a mean of 0.0 and a standard deviation of 1.0. Therefore, most of the examinees will have ability levels between -3 and +3.

## REFERENCES

- Bock, R. D. (1972). "Estimating item parameters and latent ability when responses are scored in two or more nominal categories" *PSYCHOMETRIKA*: 46, 443-459.
- Sotaridona, L. S., & Meijer, R. R. (2002). "Statistical properties of the K-index for detecting answer copying in a multiple-choice test" *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 copying" *APPLIED PSYCHOLOGICAL MEASUREMENT*: 21, 307-320.

- Wollack, J. A. (2003). "Comparison of answer copying indices with real data" *JOURNAL OF EDUCATIONAL MEASUREMENT*: 40, 189-205.
- Wollack, J. A., & Cohen, A. S. (1998). "Detection of answer copying with unknown item and trait parameters" *APPLIED PSYCHOLOGICAL MEASUREMENT*: 22, 144-152.
- 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.



JAMES WOLLACK is an associate scientist in the Department of Testing & Evaluation Services at the University of Wisconsin-Madison. His research interests are in detection of answer copying, score scale stability and test construction. He has served as a consultant or expert witness on numerous cases involving suspected cheating on high-stakes exams, and currently serves as a hearing examiner, presiding over cases of student academic misconduct at the University of Wisconsin. He earned his Ph.D. and M.S. in educational psychology from the University of Wisconsin and his B.S. in psychology from the University of California at Davis.