DOCUMENT RESUME

ED 351 358 TM 019 168

AUTHOR Frary, Robert B.
TITLE Statistical Detection of Multiple-Choice Test Answer Copying: State of the Art.
PUB DATE Apr 92
NOTE 13p.; Paper presented at the Annual Meeting of the Measurement Services Association (San Francisco, CA, April 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cheating; College Entrance Examinations; Ethics; *Evaluation Methods; Higher Education; Identification; Licensing Examinations (Professions); *Multiple Choice Tests; *Plagiarism; *Responses; Statistical Analysis; *Testing Problems
IDENTIFIERS *Large Scale Programs

ABSTRACT

Practical and effective methods for detecting copying of multiple-choice test responses have been available for many years. These methods have been used routinely by large admissions and licensing testing programs. However, these methods are seldom applied in the areas of standardized or classroom testing in schools or colleges, and knowledge about them is not widespread. This paper reviews the various methods proposed by nine researchers or teams with respect to their effectiveness and applicability and discusses possible reasons that they are not more widely used. Not the least of these may be discomfort with how conclusively this form of academic dishonesty can be detected in the absence of more direct evidence. The methods recommend by W. H. Angoff (1974), the paired indicator methods of B. A. Hanson and others (1987), and the adjacent-non-adjacent method of C. E. Stegman and B. M. Barnhill (1981, 1982) are feasible for use in a large-scale testing environment. (Author/SLD)

Statistical Detection of Multiple-Choice Test Answer Copying: State of the Art

Robert B. Frary
Office of Measurement and Research Services
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061-0438
Tel. (703)231-7285 Bitnet: FRARY@VTVM1

Presented at the annual meeting of the
Measurement Services Association
San Francisco, April, 1992

Abstract

Practical and effective methods for detecting copying of multiple-choice test responses have been available for many years. These methods have been used routinely by large admissions and licensing testing programs. However, they are seldom applied in the areas of standardized or classroom testing in schools or colleges, and knowledge concerning them is not widespread. This paper reviews the various methods with respect to effectiveness and applicability and discusses possible reasons that they are not more widely used. Not the least of these may be discomfort with how conclusively this prevalent form of academic dishonesty can be detected in the absence of more direct evidence.

Statistical Detection of Multiple-Choice Test Answer Copying: State of the Art

Robert B. Frary
Virginia Polytechnic Institute and State University

Not unlike what happened in the Garden of Eden, answer copying may well have occurred among the very first students with regular exposure to multiple-choice tests. In any case, the combination of the ease of committing this form of academic dishonesty and the ostensible difficulty of detecting it after the fact makes it highly unlikely that the practice was generally avoided early in the history of multiple-choice testing. Nor is it likely that those responsible for the authenticity of the test scores were long unaware of the phenomenon, even in the absence of sophisticated methods for detecting it.

Early Efforts

An early response to what must have been perceived as a problem was that of Bird (1927, 1929) who proposed three empirical approaches to detection, all based on inspection of observed distributions of the numbers of identical wrong responses for pairs of examinees. Crawford (1930) described a similar method based on the percentage of a pair's wrong responses that were identical. Dickenson (1945) derived a ratio described as the "probable percentage of identical errors." It was based only on the number of options per item, and no distributional characteristics for the observed percentages of identical errors were derived. Anikeef (1954) compared the number of identical wrong responses to Np, where N is the total number of wrong responses of the suspected examinee and p is the reciprocal of the number of options per item. This statistic is distributed binomially with a standard deviation of $\sqrt{Np(1-p)}$, if it is assumed that examinees guess at random among all options when the answer is not known. Anikeef acknowledged the inadequacy of this assumption but claimed his method was effective for identifying cheaters. All of these methods were hampered by the lack of computational resources.

The first report of the use of a computer in the detection of cheating was by Saupe (1960), who analyzed responses processed by an optical mark reader. Recognizing that the number of identical right or wrong responses for any pair of examinees depended on the total number of right and wrong responses of each of the two examinees, Saupe used linear regression to predict for each examinee pair the number of identical right or identical wrong answers from respectively the product of the pair's numbers of right answers and their number of items jointly answered incorrectly. Then, based on the standard error of prediction, two probabilities could be estimated, namely, the probabilities that the differences between the observed and the predicted numbers of right- and wrong-answer correspondences occurred due to chance. Theoretically, over all pairs of examinees, these two estimates should be statistically independent, and Saupe's results confirmed this expectation, thus making significant outcomes for both right- and wrong-answer correspondences all the more convincing. This approach appeared to be effective in identifying pairs who had copied but was cumbersome to apply due to the need to establish regression equations for every test analyzed.

Methods in Current Use

The first investigation comparing a variety of indicators of answer copying was reported by Angoff (1974). Like Saupe (1960), Angoff used the number of identical right responses and the number of identical wrong responses. Also investigated were the number of common omissions, the number of identical wrong responses *and* common omissions, and the number of items in the longest "run" of identical wrong responses and common omissions. Angoff also adopted "independent variables" to be used in conjunction with the indicators just mentioned. Two of these independent variables were the same as Saupe's predictor variables. The others were, for an examinee pair, the product of their numbers of wrong answers, the product of their numbers of omissions, and, finally, the sum of the number of omissions and wrong responses for the pair member with the smaller number of wrong responses. These independent variables were paired appropriately with the indicators to yield eight of what Angoff then referred to (jointly) as indices. However, rather than using regression to predict the indica · from the independent variables as Saupe had done, Angoff stratified his data on the independent variables and simply used the distribution of the indicators within a stratum to evaluate an indicator pertaining to a specific pair of examinees.

To validate the eight indices just described, Angoff used three samples of responses to the verbal and mathematical subtests of the Scholastic Aptitude Tests (SAT). One consisted of examinee pairs who could not have copied from each other due to being in different geographical locations. The second contained examinee pairs who were tested at the same location but who generated no reports of apparent copying. The third was like the first except for use of a different form of the SAT. Extremely similar results were found for all indices across the three samples. Then 50 pairs of responses involving known copiers were introduced, and the indices were evaluated for their effectiveness in detecting these cases. Also considered in this process was the extent to which the indices were intercorrelated. Based on these considerations, two indices were judged best. One involved identical wrong answers and the product of the numbers of wrong answers, and the other the number in the longest run (as defined above) and the number of wrong responses and omissions for the examinee with the smaller number of wrong responses. Interestingly, an index based on the number of identical right answers and the product of the numbers of right answers performed almost as well as the one based on identical wrong answers and had the advantage of statistical independence. (These same two variables were used by Saupe, 1960.) However, this index was rejected by Angoff due to concern that nonstatisticians might attribute right-answer correspondences to jointly held knowledge no matter how small the probability that an actual instance occurred in the absence of copying.

Frary, Tideman, and Watts (1977) developed and investigated two indices of answer copying that were fundamentally different from any reported earlier. Central to their method was the estimation of the probability that each (individual) examinee would make each available response on the test including the response of omitting. These probabilities were estimated from the examinee's score and the proportions of examinees making each response. Both indices were based on the number of identical responses (right, wrong, or omission) for an

examinee pair. The two indices differed, however, with respect to the underlying hypotheses to be evaluated. The first index evaluated the hypothesis simply that the observed number of identical responses occurred in the absence of copying. The second considered one examinee as the potential copier and the other as the source. This led to evaluation of two hypotheses, namely, that first examinee made responses identical to those of the second in the absence of copying and vice-versa. The index values corresponding to these two hypotheses would not be the same unless the two examinees had completely identical responses, because, otherwise, not all of the probabilities of each making the other's (respective) responses would be the same. Expected values and variances for the number of identical responses occurring by chance were derived under both types of hypothesis. It was then possible to create standardized statistics, whose distributions were approximately normal. Thus, given any pair of examinees, (the first index) or one examinee who might have copied from another (the second index), it was possible to estimate the probability that the identical responses occurred in the absence of copying.

Frary et al. validated their indices using responses from a test with two forms containing the same items in different orders administered simultaneously in two separate rooms. Distributions of the indices were studied for pairs who had different forms of the test *and* were in different rooms and for pairs who had the same form *and* were in the same room. The same-form/same room distributions had an excessive number of very high values for both indices, while the different-room/different-form distributions had only the numbers of very high values that might be expected by chance. The first index did not identify as many instances of likely copying as the second and was not recommended for further use.

Schumacher (1980) reported a detection method that required knowledge of the seating locations of two examinees suspected of cheating. It also required that they be seated together for one segment of an examination and apart for another. Then a 2x2 chi-square statistic could be computed to evaluate the independence of their number of identical and nonidentical responses in the two locations. This approach would be practical only for examinations with multiple seatings, such as those used for professional licensing. Stegman and Barnhill (1981, 1982) reported extensive use of Schumacher's method in connection with data from the National Board of Medical Examiners. They used only identical and nonidentical wrong responses, citing the same concern as Angoff (1974) about difficulty of convincing nonstatisticians that excessive right-answer correspondence could identify copiers. A large part of this investigation was devoted to verifying the distribution of chi-square under the circumstances just described. Some evidence was reported to the effect that the numbers of common wrong answers were slightly inflated for examinees that had common academic backgrounds.

Cody (1985), in an article with no literature citations, derived what was essentially a crude reinvention of the second index of Frary et al. For an estimate of the probability that an examinee would select a particular option, Cody used only the proportion of examinees selecting that option, while Frary et al. also took into consideration the score of the examinee. In addition, Cody used only identical wrong responses, while Frary et al. used all

responses of the suspected copier. Given these two "shortcuts," one would expect Cody's method to be less sensitive in detecting copying than the second index of Frary et al.

In a study mainly devoted to evaluating methods of detecting copying, Hanson, Harris, and Brennan (1987) also introduced two new indices. Based on extensive data analyses, they noted that for certain pairs of indicators of copying, a high value for both pair elements coincided with high values for single indicators adjusted for the score levels of the examinees. One such pair of indicators was the number of identical incorrect responses and the length of the longest string of identical responses or omissions. The second pair was the number of incorrect responses in the longest string of identical responses and the number of identical incorrect responses expressed as a percentage of the maximum possible number of identical incorrect responses.

Hanson et al. also proposed and investigated an improvement on Cody's index. They estimated the probability that a suspected copier would select a specific option as the proportion of examinees selecting that option within a score stratum including the suspect. However, they did not use identical right responses and omissions in computing this modified index. Had they done so, the result would have been extremely similar to the second index of Frary et al.

Roberts (1987) evaluated an ad hoc method of detecting cheating reportedly used by some university instructors when multiple forms of a test were administered in the same room unbeknownst to the examinees. Then crossform copying supposedly could be detected by scoring an examinee's responses with all of the keys and looking for a large negative difference between the score based on the applicable key and a score based on one of the other keys. Roberts showed conclusively that this method had a high potential for generating false positive outcomes.

Bellezza and Bellezza (1989) reported a detection method based on the binomial distribution, which is roughly analogous to the first method of Frary et al. (whose article they did not cite). Not only was the first index of Frary et al. found to be less effective than their second index, the index of Bellezza and Bellezza fails to utilize the information from right-answer correspondences or to consider the score levels of the examinees and the popularity of the common wrong options involved in any particular computation of their index.

Evaluation and Use of Indices

The only reported study comparing methods of detection proposed by different investigators is that by Hanson et al. (1987), which was referred to earlier. It compared the methods they had developed and those of Angoff (1974), Frary, et al. (1977), and Cody (1985) using data from a 100-item, four-choice test involving over 19,000 examinees. Insofar as possible, each examinee in this data set was paired with another who took the test in a different physical location, to yield 9,143 distinct pairs. Of these, 8,643 were used to produce "benchmark" data reflecting false positive rates for each of the indices investigated. The second examinee

in each of the remaining 500 pairs was designated as a copier. This strategy potentially reduced the realism of the study, because the size of the scores of the pair were not taken into consideration; there may well be somewhat more copying by lower scorers from higher scorers than vice-versa. This concern aside, five levels of copying (10% through 50% of the source's responses copied) and five types of copying were simulated to replace some of the designated copier's responses with those of the source. The five types were: 1) copying at random, 2) copying more difficult items, 3) copying a run of items at the beginning and 4) at the end of the test, and 5) copying runs of five items selected randomly.

The results of the study by Hanson et al. are comprehensive and complex. The false positive rates were varied as well as the number of items copied and the type of copying. No method performed well when only 10% of the items were copied. For more extensive copying, the methods performed differentially well according mainly to the type of copying. Methods based on runs were by far the best for detecting copying of types 3 and 4 listed above and by far the worst for types 1 and 2. Evidence both internal and external to the study suggested that copying type 5, involving short runs, may most closely correspond to copying that occurs in actual cases. Surprisingly, the indices based on runs performed relatively poorly for this type of copying. For type 5 copying, with a false positive rate of .001 and 50% of the items copied, the indices *not* involving runs identified from 89% to 96% of the copiers, while the indices involving runs identified from 82% to 90%. Overall or for detecting type 5 copying, the only relatively poor performer among the indices not involving runs was (not unexpectedly) the unmodified index of Cody (1985), and none was consistently the best.

Subsidiary findings of Hanson et al. are of interest. They found that the theoretical significance levels of the indices of Angoff, Frary et al. and Cody did not coincide well with the false positive rates of their benchmark data. Specifically, the theoretical levels led to false positive findings in excess of expectation. Thus, empirical establishment of significance levels was recommended. This finding complements that of Stegman and Barnhill (1982) to the effect that phenomena other than copying may cause answer similarities to a minimal degree and suggests that all theoretical levels of significance for copying indices should be discounted to some extent. Hanson et al. also recommended caution in the use of multiple indices due to potential inflation of the false positive rate and the difficulty of empirical or theoretical evaluation of this phenomenon as additional indices are used.

While the study by Hanson et al. was thorough, it had limitations. It was apparently conceived largely with respect to the detection of copiers in a large-scale professional licensing or admissions testing environment. Presumably, such findings might lead to invalidation of scores and ensuing legal actions. This focus is evident in their discussion of which indices to recommend; the extent to which an index was likely to be understood by nonstatisticians was viewed as an important characteristic. In any case, Hanson et al. were apparently not interested in the performance of the indices for relatively short tests or for tests with relatively few examinees. Also, ease of computation was not considered, which would certainly be a factor in any effort to monitor the prevalence of copying for purposes of

prevention over large numbers of tests of varied length administered to relatively small groups.

Methods requiring stratification would be especially difficult to deal with in such cases and would probably not be feasible for examinee groups of, say, fewer than 100. The pair methods of Hanson et al. and the second index of Frary et al. are the only viable indices not requiring stratification. Further, the pair indices of Hanson et al. require empirical establishment of their significance levels for every test, a process also requiring large numbers of examinees. In contrast, the second index of Frary et al. has theory-based significance levels. While the findings of Hanson et al. suggest that these levels should be interpreted conservatively, they nevertheless provide a consistent way to evaluate the index across varied testing situations. Thus, it would appear that the second index of Frary et al. is the only practical index for use in a classroom or similar testing situation.[1]

Extensive experience with the second index of Frary et al. permits additional indirect evaluation of its effectiveness. This index, known as g_2, has been computed routinely for users of the test scoring and analysis system at Virginia Polytechnic Institute and State University for over 15 years. Over 4,000 classroom tests are processed annually, and g_2s are produced for all examinee pairs for a substantial proportion of these. A subroutine for the program generating test scores and item analyses will produce a histogram of the g_2s for all examinee pairs and print out cases generating high values. The additional computational burden is minimal. The purpose is more often to monitor the prevalence of likely copying in order to check on the effectiveness of preventive measures than to obtain evidence for use in judicial hearings.

Numerous instances, probably in excess of 100, of the following scenario have occurred over the years: An instructor gives a single-form test in a crowded room, and analysis yields a number of high g_2s far in excess of expectation; the instructor then administers future tests using multiple forms (scrambling the order of the items), and high g_2s occur only within expectation or (occasionally) within the same form when examinees with that form managed to sit where at least one could observe the other's responses. These results have been extremely consistent over considerable variation in class size (as low as 30) and test length (as few as 20 items) and over varied test content. In other instances, instructors have monitored later test behavior of student pairs with high g_2s and obtained visual confirmation of copying. A brochure of the Office of Measurement and Research Services (1990) outlines a number of aspects of the use of the g_2 index and is available upon request. This brochure makes it clear that statistical evidence alone should not be used to determine guilt, since false positives can occur at any level of significance regardless of the efficacy of the detection

[1]B. A. Hanson (personal communication, 24 January 1992) has used a log-linear modeling procedure to estimate within-stratum distributions for the index of Angoff (1974) involving the number of common wrong responses and reported that "Such a modification...may make [this index] more useful for small testing programs, perhaps even for classrooom testing." However, this approach would seem rather cumbersome and might well not be reliable for classes with, say, fewer than 50 examinees.

statistic used. However, when examinees generating very high g_2s are shown to have been seated where copying could occur, the likelihood of guilt has to be considered almost overwhelming.[2]

A study by Frary (1978) monitored a large number of classroom tests to identify 100 university students who were extremely likely to have copied based on their values of g_2. These and a random sample of 100 were sent a questionnaire asking their evaluation of the penalty that should be imposed for various instances of academic dishonesty, including answer copying. Based on a return rate of over 90%, the sample composed of likely copiers provided significantly more lenient answers overall ($p=.006$) and especially so with regard to instances in which the dishonest act did not directly or necessarily benefit its perpetrator, e.g., stealing a copy of an already administered test for delivery to a fraternity test file.

Frary and Olson (1985) reported use of g_2 to detect answer copying in a districtwide administration of a standardized test in elementary schools. Teacher salary adjustments depended on the scores earned by students, and it was feared that some teachers might encourage collusion as a means of elevating their students' scores. Apart from this concern, the district wished to monitor the extent to which lax test administration might have permitted copying. For most classes, only the expected number of higher g_2s were noted. However, within a few classes excessive numbers occurred. It was also possible to make an extensive evaluation of the extent to which excessive false positives might occur, as evidenced by cross-room pairs with high g_2s. Virtually none with associated probabilities of less than .0001 occurred except in a school where the actual locations of the students at the time of testing were in doubt.

Discussion

The varied methods used to detect answer copying by large testing organizations, though cumbersome, are finely honed and should be generally effective. The methods recommended by Angoff (1974), the paired indicator methods of Hanson et al. (1987), and the adjacent-nonadjacent method reported by Stegman and Barnhill (1981, 1982) arose in and are feasible for this environment. Indeed, the organizations in using these indices no doubt maintain a high level of vigilance over the authenticity of their testing programs and use the methods they have chosen with care and diligence. An extensive analysis of the legal matters arising when testing organizations identify likely answer copiers (Buss & Novick, 1980) attests to the need for such attention to detail.

---

[2]When the suspect pair were seated side-by-side and there is no evidence as to the direction of copying, the guilt that is apparent must go unassigned. This is true despite the fact that separate g_2s are computed for each potential direction of copying. A correspondence sufficient to make one of the g_2s very high will unavoidably make the other quite high as well. In such a situation, the g_2 reflecting the probability that the higher scorer did not copy from the lower scorer will usually be greater regardless of the actual direction of copying.

Buss and Novick did, however, note one aspect of the procedures just mentioned that could be improved. All of these methods depend on the number of identical wrong responses as the primary indicator of copying. As noted earlier, this practice probably reflects concern that nonstatisticians might not understand that the number of identical right answers might exceed expectation. This point aside, Buss and Novick noted a different possible effect of excluding right answer correspondences.

> Unfortunately, some testing programs now examine only the number of identical incorrect responses... When the statistical index of cheating is computed in this way, evidence that may be favorable to the [accused] examinee from the items that were answered correctly by one examinee and incorrectly by the other...is ignored. ...this is unfair...[and]...a fundamentally incorrect application of statistical methodology. (p. 12)

Thus, inclusion of identical right-answer correspondences not only increases the evidence of copying when these exceed expectation but may also decrease this evidence when they are less similar than expected. Of all the indices in use as reported above, only the g_2 index of Frary et al. uses right- as well as wrong-answer correspondences. However, all of the other approaches could be adapted to include information from right-answer correspondences. This would seem to be the correct and fair thing to do to maximize detection and minimize false positives. At the same time, the use of indices based only on common wrong responses could still be justified under restricted circumstances. Consider the following scenario: A testing organization uses both right and wrong common responses for its detection index, the index and additional evidence strongly suggest that a specific examinee copied, the examinee's score is invalidated, the examinee sues the testing organization, and lawyers advise that the index based on right and wrong responses would be difficult to explain in court. In such a case, it would seem justifiable to present as evidence only a statistic based on common wrong responses. Of course, this line of reasoning assumes that the copying index based on common wrong responses, like the index based on common right *and* wrong responses, would be "highly significant."

While statistical detection of answer copying is well established for licensure examinations and for large-scale admissions testing programs, there are areas where it is little applied. One of these is standardized testing in elementary and secondary schools. As noted earlier, an easy way to increase the scores for a school or class is to precipitate answer copying, perhaps simply by lax attention to student behavior or subtle encouragement. As was shown in the study reported by Frary and Olson (1985), there is no technical difficulty in monitoring the prevalence of copying through use of the g_2 index. That testing companies do not offer such analyses as an additional service may well reflect what they (possibly correctly) perceive to be the low level of demand for such a service. An eminent school administrator who was a discussant when the Frary and Olson paper was delivered noted somewhat casually (as if what he said was rather obvious) that, while the results were interesting, it was not the kind of thing that one would want to do on a regular basis (to avoid stirring up trouble?)!

The situation with respect to classroom testing is not very different from that for standardized testing. Very few university testing/measurement services provide any form of answer copying detection. Even 10 years ago, this absence might have reflected in part lack of computing resources. However, this should rarely be the case at the present time. This point aside, many testing/measurement services lack professional direction, and this may account for many instances of failure to provide this service. Persons in charge of measurement services but not professionally trained in this area may have no knowledge of the literature concerning detection of copying.

Another prevalent cause of failure to offer answer copying detection is lack of demand or apathy on the part of users of university measurement services. Even worse, there are faculty members and administrators who view the process as extremely distasteful and wish to avoid it at all costs. Such persons typically believe that actual incidents of answer copying are rare and tend to distrust statistics. A conclusion reached by an individual displaying both of these characteristics may be illustrative. The individual in question was a university administrator in charge of appeals of convictions by the university's honor system. A high g_2 value played a part in a conviction that had been appealed. The administrator became aware that an excess number of very high g_2s had been observed for the test in question but that the probable other cheaters had not been accused. His preliminary conclusion was that the statistic must be faulty, because he believed that so many examinees could not have cheated. Actually, only about 25 students in a class of 500 were implicated, and consultation with more than one statistician finally convinced the administrator that nothing was wrong with the statistical outcomes.

When circumstances such as those just mentioned exist, it is the responsibility of individuals in charge of testing/measurement services to provide the information that will make provision of answer copying detection well accepted. For example, the literature is replete with evidence of the prevalence of answer copying in higher education. Various surveys have indicated that large proportions of college students will copy answers whenever they think it will be of benefit and that the probability of getting caught is very low. The popular press has also reported many circumstances consistent with this conclusion. In addition, testing/measurement personnel need to be effective promoters of proper use of copying detection statistics. While these statistics often tend to be underused, overzealous or vengeful faculty members or administrators may misuse them in an arena where many of the participants are at best seminumerate. Continual monitoring of the use of copying detection statistics is a *preliminary* requirement for offering them.

Because of the potential for misuse of any copying detection statistic, the developers of the g_2 statistic have not freely provided copies of the program used for its production at Virginia Polytechnic Institute and State University, though, of course, anyone familiar with statistical programming could produce one. However, the program is available free-of-charge to qualified directors of university measurement services provided that they agree: 1) Not to distribute it further, and 2) To adapt for their local situations the explanatory brochure

described earlier (Office of Measurement and Research Services, 1990) and make it available to all recipients of analyses involving g_2s.

References

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

Anikeef, A. M. (1954). Index of collaboration for test administrators. *Journal of Applied Psychology, 38*, 174-177.

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.

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

Bird, C. (1929). An improved method of detecting cheating in objective examinations. *Journal of Educational Research, 19*, 341-348.

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.

Crawford, C. C. (1930). Dishonesty in objective tests. *School Review, 38*, 776-781.

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

Frary, R. B. (1978). *Academic dishonesty as viewed by a random sample of college students and a sample identified as cheaters.* Paper presented at the annual meeting of the American Educational Research Association, Toronto, Ontario.

Frary, R. B., & Olson, G. H. (1985). *Statistical detection of answer copying and coaching.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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.* ACT Research Report Series 87-15. Iowa City, IA: American College Testing Program.

Office of Measurement and Research Services (1990). *Detection of answer copying on multiple-choice tests and interpretation of* g_2 *statistics.* Blacksburg, VA: Virginia Polytechnic Institute and State University.

Roberts, D. M. (1987). Limitations of the score-difference method in detecting cheating in recognition test situations. *Journal of Educational Measurement, 24*, 77-81.

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

Schumacher, C. F. (1980). *A method for detection or confirmation of collaborative behavior.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Stegman, C. E., & Barnhill, B. M. (1981). *Empirical sampling distribution of the adjacent/ nonadjacent statistical procedure for detecting copying.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Stegman, C. E., & Barnhill, B. M. (1982). *Sampling distribution of χ^2 for detecting copying on multiple-choice examinations.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.