References

Block, J. H. (Ed.). (1971). *Mastery learning: Theory and practice.* New York: Holt, Rinehart & Winston.

Burdsal, C., & Halcomb, C. G. (1967). Unpublished technical document, Texas Tech University, Lubbock, TX.

Hobbs, S. H. (1987). PSI: Use, misuse, and abuse. *Teaching of Psychology, 14,* 106–107.

Johnson, J., & Chen, B. (1974). Communication. *Journal of Applied Behavioral Analysis, 7,* 353.

Kasschau, R. A., & Halpern, M. S. (1979, September). *Computer-managed instruction: Individualizing introductory psychology for 1,000 students.* Paper presented at the meeting of the American Psychological Association, New York.

Keller, F. S. (1968). "Good-bye, teacher. . . ." *Journal of Applied Behavior Analysis, 1,* 79–89.

Minke, K. A., & Carlson, J. G. (1972). *Psychology and life unit mastery system.* Glenview, IL: Scott, Foresman.

Minke, K. A., & Carlson, J. G. (1985). *Psychology and life unit mastery system.* Glenview, IL: Scott, Foresman.

Roll, J. H., & Pasen, R. M. (1977, June). *Computer managed instruction produces better learning in an introductory psychology course.* Paper presented at the Conference of Computers in Undergraduate Curriculum, East Lansing, MI.

Stokes, M. T., Halcomb, C. G., & Slovacek, C. (in press). Delaying user responses to computer-mediated test items. *Journal of Computer-Based Instruction.*

Wired for the future: COBA's computer network nears completion of phase. (1986, Winter). *Texas Tech Business,* pp. 13–14.

Notes

1. Preparation of this article was supported in part by an NSF-CAUSE Grant SER–7907702 to Douglas C. Chatfield (second author) in 1981.

2. Requests for reprints should be sent to Charles G. Halcomb, Box 4100, Department of Psychology, Texas Tech University, Lubbock, TX 79409.

Detection of Cheating on Multiple-Choice Tests by Using Error-Similarity Analysis

Francis S. Bellezza
Suzanne F. Bellezza
Ohio University

Cheating on multiple-choice examinations is a serious problem not easily overcome by using more test forms, more proctors, or larger testing rooms. A statistical procedure compares answers for pairs of students using those items on which both made errors. If the number of identical wrong answers is sufficiently greater than the number expected by chance and if the students were seated close together, then cheating is likely. Using this analysis with 90 examinations has suggested ways to discourage cheating and demonstrated some limitations of the procedure.

Cheating on tests is common in American colleges. About 75% of college undergraduates admit that they have cheated on tests (Baird, 1980). Baird found that obtaining information from another student about a test was the most common form of cheating. In a study of medical school students, 87% admitted to having cheated as undergraduates and 58% confessed to cheating while in medical school (Sierles, Hendricks, & Circle, 1980). Many students accept cheating without concern. Forty percent of the college students surveyed did not disapprove of cheating on tests, 29% did not feel guilty about cheating, and only 1% said that they would report cheating if they observed it (Baird, 1980). Certain groups are more likely to cheat: men, students with low grades, underclassmen, business majors, fraternity and sorority members, and students involved in few extracurricular activities (Baird, 1980). The two most common reasons for cheating are grade pressure and lack of time to complete class work (Barnett & Dalton, 1981).

Cheating by Copying on Multiple-Choice Examinations

We are specifically concerned with copying during multiple-choice examinations which, in large and crowded rooms with an inadequate number of proctors, favor cheating (Houston, 1976). Copying someone else's work during a test is the fifth most common form of cheating (Baird, 1980). One may assume that a large proportion of test copying involves multiple-choice tests because it is easy to copy the marks made on the answer sheet.

Instructors can discourage copying by taking precautions. Although use of multiple forms of the test is one such precaution, Houston (1983) found that rearranging the order of questions did not reduce copying, presumably because students copied only from someone with the same form. On the

other hand, Houston found that rearranging the order of questions on multiple test forms, rearranging the order of alternative answers, and spreading the students out in the testing room reduced copying.

Another helpful precaution is vigilant monitoring during the test by an adequate number of proctors. Although Barnett and Dalton (1981) found that intelligent students cheated less in high-risk situations (e.g., when there were several proctors) and Baird (1980) found that women were much less likely to cheat in high-risk environments, students typically perceive proctoring as less effective than do teachers. Forty-eight percent of faculty but only 21% of students agreed that proctors watch carefully and consistently throughout tests (Barnett & Dalton, 1981).

This difference between instructors' and students' evaluations of the effectiveness of proctoring suggests that copying is difficult to detect. Although instructors think that they are vigilant, students believe that copying is undetectable, a result they attribute to inadequate supervision. Our own experience is that instructors who performed the following analysis were surprised by the amount of cheating suggested by the procedure. In addition, the analysis indicates that some students engaging in suspicious behavior, such as looking around frequently, are not cheating.

A final precaution that discourages copying is to assign seats that vary from one examination to the next and differ from students' lecture seats. This procedure ensures that students cannot plan to copy from friends or from students who are performing well in the course.

Analysis of Multiple-Choice Items Using Error Similarity

Although these precautions are reasonable, instructors do not always have the resources to carry them out. To help discourage students from cheating on multiple-choice examinations, we developed a procedure that determines the similarity of errors for any pair of examinees. The procedure is implemented by a computer program that records the items both students got wrong and whether the number of same wrong answers was above a chance level. The analysis is restricted to errors because cheating is suggested if two students consistently choose identical wrong alternatives for the same items.

Cody (1985) described a similar procedure, and both Cody's method and the one described here are closely related to Angoff's (1974) Index B developed to detect cheaters. The advantage of our procedure is that it yields a critical value indicating whether cheating is likely.

To exemplify the procedure, assume that Student X made 25 errors and Student Y made 23 errors on a 60-item multiple-choice test with 5 alternative choices for each item. Also, assume that 20 of these errors involved the same items and that for 18 of these 20 items both students chose the same wrong alternative. Even though the answer sheets are not identical, an error-similarity analysis suggests that cheating is likely because the probability of choosing by chance 18 of the same wrong alternatives for 20 wrong items is on the order of 5 in a million. One must also use a seating chart to determine whether two students were sitting close enough to cheat and which student may be copying from the other.

Probability of Same Errors Using the Binomial Distribution

The error-similarity analysis works like this: Once all of the answer sheets have been scored, the computer program counts the number of times all pairs of students chose the same incorrect alternative for all items. If all four incorrect alternatives of a five-alternative item were equally attractive, then the probability of two students choosing the same incorrect alternative for an item by chance would be 4/16 or .25. Because it is unrealistic to expect that all incorrect alternatives will be equally likely, the probability (P) of an identical error by two students on an item is computed by the program from the students' responses and typically has a value around .40.[1] Using the binomial distribution (McNemar, 1962), when the probability that errors match on any item is P, then the probability for k of N item errors matching is:

$$\frac{N!}{k!(N-k)!} P^k(1-P)^{N-k}.$$

Using the previous example of Student X and Student Y, assume that the probability (P) of any two students choosing the same wrong answer is .40. Then the probability of choosing 18 or more answers the same out of 20 wrong items is the probability of choosing 18, 19, or 20 answers the same by chance. Using the binomial distribution with $P = .40$ and $N = 20$, this probability can be represented as $(20!/18!\ 2!)(.40)^{18}(.60)^2 + (20!/19!\ 1!)(.40)^{19}(.60)^1 + (20!/20!\ 0!)(.40)^{20}(.60)^0$, which is equal to .000004700 + .000000330 + .000000011 = .000005031, or about 5 millionths.

Because computing binomial probabilities when N is large can be time consuming, a normal approximation to the binomial may be used if $NP > 5$ and $N(1 - P) > 5$ (McNemar, 1962). The value for the mean of this normal approximation is NP and the value of the standard deviation is $\sqrt{NP(1-P)}$. For the binomial distribution from the just-cited example in which $P = .40$ and $N = 20$, the mean of the normal approximation is 8 and the standard deviation is 2.19. Because 18 identical answers were given for 20 wrong items, the standard normal value is $z = (X - M)/SD = (17.5 - 8)/2.19 = 4.34$. Using a standard normal table, the probability of obtaining 18 or more of the same wrong answers by chance is .000007 or about 7 millionths. This is close to the probability of 5 millionths computed previously using the binomial distribution.

The probability of choosing the same wrong answers by chance can be computed for every possible pair of students. If there are S students, then the number of comparisons is $S(S - 1)/2$. The probability computed for each pair of students indicates the probability with which their similar er-

[1] If there are only four alternative answers for each item, as is true of many items provided with introductory psychology texts, then P may be higher than .40. The program automatically takes the number of alternatives into account, and the procedure remains the same.

rors occur by chance. The error-similarity analysis is based on these probabilities. Although only probabilities need to be computed for the analysis, it is convenient to express the degree of similarity by using a z score computed using the normal approximation.

Is Cheating Occurring?

To illustrate how the detection procedure works, consider the following example. Figure 1 shows a distribution of error-similarity scores expressed as z scores for 39 students. Of the 741 possible pair-wise similarity scores, only 735 could be computed because some pairs of students had no errors in common. The 39 students were administered the same examination form and represent approximately half the students tested in an Introductory Management course at Ohio University.

The error-similarity scores form a unimodal distribution with a mean of $-.63$ and a standard deviation of 1.01. This distribution allows all the error-similarity scores to be inspected simultaneously; however, the exact shape of this distribution is not critical to the analysis. To generate these 735 scores, the N used in computing each score depended on the number of wrong items each pair of students had in common. The P of the same incorrect choice for any wrong item was estimated by the program from the item analysis, and for this group the value was .42.[2] Two similarity scores lie above the rest of the distribution: one at $z = 3.44$ and one at $z = 4.89$.

How can we decide whether these high error-similarity scores indicate cheating? Our procedure considers the computed z scores as z tests. If the z score is greater than some critical value, then most likely the score results from one student copying from another. In order to use this procedure, a proximity or closeness parameter (C) must be used. The parameter C represents the estimated number of other students with the same test form from whom a given student may be close enough to copy. To estimate C, we sat in a number of testing rooms and tried to read answer sheets placed on other desks. We estimated the value of C to be no larger than 3 when two forms of a test were used with a large number of test takers. That is, the average student can see the answer sheets of, at most, three other students using the same test form. So the maximum number of error-analysis scores that might actually involve cheating is $C \times S$, where C is the proximity parameter and S is the number of students using the same form of the test.

When using a z score as a z test, the alpha level represents the probability of a Type I error. When making a Type I error, one concludes that cheating is occurring when it really is not. If we wish to have only an alpha probability of making one or more Type I errors when performing $C \times S$ tests, then an error-similarity score considered as a z test has to be significant at the level of (alpha/C S). This adjustment in the alpha level is a Bonferroni correction (Kirk, 1982). In

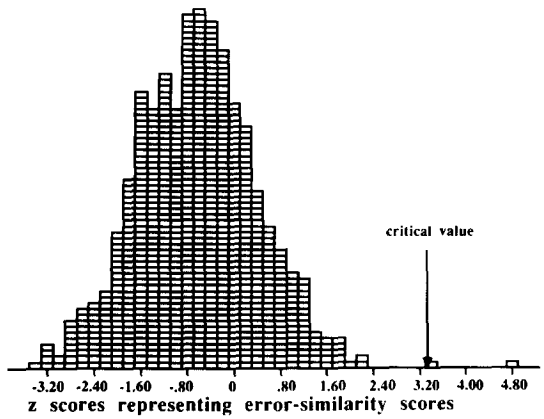

Figure 1. Distribution of error-similarity scores for all possible pairs of 39 students who took the same form of a 60-item multiple-choice examination.

the distribution presented in Figure 1, the alpha level was set at .05. Thus, the probability value used to test the z score for each pair of students was .05/3(39) or .00043. This probability can be looked up in a standard normal table and shows that a z score has to exceed the critical value of 3.33 for cheating to be suspected.

The first outlying score ($z = 3.44$) falls just above the critical value. However, the seating chart for the test revealed that the two students were not sitting near each other.[3] So, this z score is not one of the C S error-similarity scores that we need to test for significance. The second outlying score ($z = 4.89$) does represent two students sitting near one another. An analysis of the 60 items shows that Student X and Student Y got 31 of the same items correct. Also, Student X got 3 items correct that Student Y got wrong, and Student X got 1 item wrong that Student Y got correct. Finally, Student X and Student Y got 25 of the same items wrong. The z score is based on these last 25 items, of which the two students answered 24 the same. Because a Bonferroni correction has been made for the number of pairs of students tested, we may conclude that there is a probability of .95 that these two students were involved in cheating.

The next question is, which student was the cheater? We usually assume that collaboration does not occur because our seating involves random assignment. Inspection of the seating chart often shows the student suspected of cheating sits behind the other. In most instances, this student copied from the student in front. Inspection of nonmatching items may also provide some evidence. In this example, Student X got only two more items correct than Student Y, but sometimes this discrepancy is larger. The student with the lower score is likely to be the student copying.

As mentioned previously, the z scores are used as error-similarity scores to represent the degree of similarity between two students' sets of errors. These z scores are more convenient to use than probabilities because they do not take on

[2]For each pair of students, N is the number of items on which both students made an error. In addition, P can be computed using only those items on which both students made errors.

[3]Obtaining a z score this high and finding that the two students involved were not sitting close to each other is a rare event, as demonstrated in Figure 2.

extremely small values and seem to form a unimodal, symmetric distribution. To perform the cheating analysis, however, the z scores are not necessary. Only the binomial distribution is needed to compute a probability for each pair of students tested. If the computed probability is less than the value (alpha/C S), then one may suspect cheating.

Analysis of Extreme Scores From 90 Tests

Figure 2 shows 24 extreme error-similarity scores obtained by analyzing test results from 15 classes in Introductory Management at Ohio University. Each class was administered two midterm tests and a final examination, and each of these tests consisted of 60 five-alternative, multiple-choice items. Because two forms of each test were always used, there were a total of 90 analyses of the kind illustrated in Figure 1. From these 90 analyses, 24 pairs of students were found to be above the critical value computed for each analysis.

The mean number of students included in each analysis was 31.6. Note that in a class of 32 students, the number of pairs of students, regardless of proximity, is $[S(S - 1)/2] = (32 \times 31)/2 = 496$. Of these 496, $C \times S = 3 \times 32 = 96$ pairs are estimated to be sitting close to one another, and $496 - 96 = 400$ are not. Using comparable numbers from all the classes analyzed, a mean of 15% of the students were estimated to be sitting in proximity. Yet, Figure 2 shows that 19 of the 24, or 79%, of the extreme scores involved pairs of students sitting close enough to cheat. If these large error-similarity scores were due to chance factors, then only 15% rather than 79% should involve students sitting close together. Using a binomial test with $N = 24$ and $P = .15$, it can be shown that obtaining 19 or more cases of students sitting in proximity out of a total of 24 cases is a very unlikely event. The results shown in Figure 2 allow us to conclude that error-similarity scores larger than the critical value resulted largely from one student's copying answers from another.

The percentage of students in these 90 analyses suspected of copying was very low. For the first test in each course, this was .5%; for the second test, .3%; and for the final examination, 1.4%. In each of the classes, the students had been forewarned of the analysis and were cautioned against trying to cheat. However, data from classes prior to the use of the analysis suggest that the number of students copying on a test may have been as high as 5%.

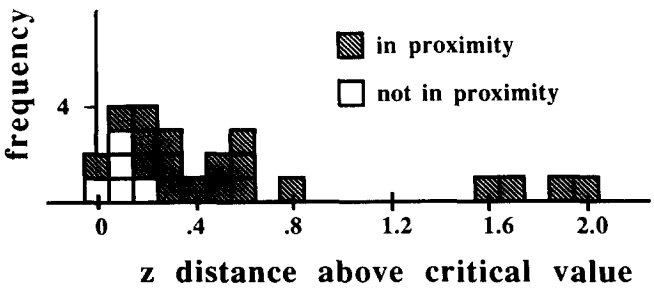

Figure 2. Distribution of extreme scores from 90 multiple-choice examinations involving students sitting in proximity versus those not in proximity.

Measures to be Taken Once Cheating is Suspected

Although error-similarity analysis should be used in a conservative manner, we believe that it can help control cheating. We list examples from the most to the least conservative.

1. General deterrent to copying. The instructor may announce in class that a computer program exists to detect cheating and that it is used whenever answer sheets are machine scored. Even if the instructor does not always perform the analysis, this policy is helpful because it discourages students from cheating. Such an announcement is most effective if students are convinced that such a computer program exists.
2. Evaluation of testing environment. Sometimes an instructor wants to determine, in general, if any copying is occurring on tests. No action against individual students is planned. After performing the analysis and inspecting the results, the instructor may decide to use more forms of a test, use random assignment of test seats, get more proctors, use a larger testing room, and so on. Also, the program can be used when doing research on the factors that influence cheating behavior.
3. Special seating assignments. As a result of the analysis, the instructor may decide to assign suspected cheaters test seats that can be closely monitored. If randomly assigning test seats is part of the course procedure, then students suspected of cheating should consider their assigned seats as resulting from chance.
4. Special counseling. The instructor may wish to speak to suspected cheaters. Such an interview provides the opportunity not only to inform the student of the instructor's suspicions but also to discuss problems the student may be having in the course. Special seating may also be assigned for the next test.
5. Disciplinary action. The instructor may wish to bring charges against the suspected cheater using the departmental, college, or university committee dealing with student misconduct. Sometimes an instructor will have corroborative evidence, such as other students who witnessed the cheating or the reports of proctors who noted suspicious behavior. Occasionally, the same individual will be repeatedly detected as cheating despite the precautions taken. In this case, the statistical evidence becomes compelling. We have observed the same students appearing as suspects in more than one course.

Limitations of the Error-Similarity Analysis

When using the error-similarity procedure, one must be aware of its limitations. First, it deals only with one kind of cheating: copying another's work on a multiple-choice examination. It cannot detect cheating by using crib sheets on tests or through out-of-class plagiarism. Furthermore, in order for copying to be detected, the test should be difficult enough for the mean number of errors to be 15 items or more when there are five alternatives for each item. Even under these optimal conditions, cheating will not be detected if a student copies only a few answers.

A second limitation is that there must be a record of where

students sit during the test. The procedure works best when students are randomly assigned to different seats for each test so that any cheating discovered is probably not the result of collaboration.

A third problem is that the procedure is statistical in nature. Even if there is a probability of only 1 in 1 million that two students could have the same pattern of errors, it is still possible that cheating did not occur.

Fourth, there are questions as to how statistical evidence is interpreted legally (Buss & Novick, 1980). This last point is important because disciplinary action taken against students involves judiciary committees that often use procedures and rules of evidence similar to those of a court of law.

Fifth, many college deans and judiciary committees, even when convinced of the value of the procedure, are not familiar with evaluating statistical evidence and its associated probability values. They usually deal with physical evidence and interview witnesses of cheating. A user of the error-similarity procedure who wishes to take disciplinary action must be prepared to educate the people involved in the case.

The Computer Program

The computer program used to implement the error-analysis procedure is written in FORTRAN IV and can be obtained from the authors. The program was developed on an IBM mainframe computer but can be used with any computer that has a FORTRAN compiler, including microcomputers. For the program to be of practical value, the user must be able to read multiple-choice answer sheets with an optical scanner or some such device. The file created by this device must be available to the computer running the program. Input–output modifications of the program might be necessary to make it conform to the particular computer system being used.

The output of the program corresponds to what has been described here. A distribution of z scores is printed as well as statistics for pairs of students who have high z scores. These statistics include the probability that the error overlap could occur by chance, the item numbers for which answers are the same, and information regarding which of the two students performed better on the test. Also printed are critical z values for a variety of values for the proximity parameter (C) combined with different alpha values. The user can estimate what value of C is appropriate for a particular testing room and decide on the value of alpha. The appropriate critical z value can then be found in the output.

References

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

Baird, J. S., Jr. (1980). Current trends in college cheating. *Psychology in the Schools, 17,* 515–522.

Barnett, D. C., & Dalton, J. C. (1981). Why college students cheat. *Journal of College Student Personnel, 22,* 545–551.

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.

Houston, J. P. (1976). Amount and loci of classroom answer copying, spaced seating, and alternate test forms. *Journal of Educational Psychology, 68,* 729–735.

Houston, J. P. (1983). Alternate test forms as a means of reducing multiple-choice answer copying in the classroom. *Journal of Educational Psychology, 75,* 572–575.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Wadsworth.

McNemar, Q. (1962). *Psychological statistics* (3rd ed.). New York: Wiley.

Sierles, F., Hendricks, I., & Circle, S. (1980). Cheating in medical school. *Journal of Medical Education, 55,* 124–125.

Note

A copy of the FORTRAN program in IBM MS DOS can be obtained from the authors by sending a blank 5-1/4 in. computer diskette to Francis S. Bellezza, Department of Psychology, Ohio University, Athens, OH 45701. If readers provide a 3-1/2 in. computer diskette, the FORTRAN program will be sent as a Macintosh text file.