DOCUMENT RESUME

ED 421 530 TM 028 856

AUTHOR	Bay, Luz
TITLE	Detection of Cheating on Multiple-Choice Examinations.
PUB DATE	1995-04-21
NOTE	47p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Cheating; Class Size; *Identification; *Multiple Choice Tests; Simulation; Test Length; *Testing Problems
IDENTIFIERS	*Binomial Distribution

ABSTRACT

An index is proposed to detect cheating on multiple-choice examinations, and its use is evaluated through simulations. The proposed index is based on the compound binomial distribution. In total, 360 simulated data sets reflecting 12 different cheating (copying) situations were obtained and used for the study of the sensitivity of the index in detecting cheating and the conditions that affect its effectiveness. A computer program in C language was written to analyze each data set. The simulated data sets were also used to compare an index developed by R. Frary and others (1977) and error-similarity analysis (F. Belleza and S. Belleza, 1989). In general, the new index was effective in detecting cheaters as long as enough items were copied. It was sensitive enough to detect cheating when between 25 and 50% of the items were copied in a 50-item test, but was less sensitive when the test was shorter. It was also less sensitive when there were fewer cheaters in a class. Although effectiveness is influenced by test length, it is not influenced by class size. Similarities and differences among the three indexes are discussed. (Contains 2 tables, 11 figures, and 14 references.) (SLD)

Detection of Cheating

on

Multiple-Choice Examinations

Luz Bay

American College Testing

Paper presented at the annual meeting of the

American Educational Research Association

21 April 1995

Detection of Copying on Multiple-Choice Examinations

Luz G. Bay
American College Testing (ACT)
21 April 1995

To cheat is to "act in a dishonest way to win an advantage or profit" (Webster, 1987, p.381). But no matter how ethically or morally wrong it seems, cheating is part of college life. Researches done on the prevalence of cheating in American colleges and universities indicate that many college students cheat their way through college. There are different forms of cheating that college students employ, but the all time favorite is copying during examinations (Collison, 1990).

At first consideration, cheating on examinations seems a problem of ethics of individual students. This is true if cheating is an isolated phenomenon. But large scale cheating could be a threat to the validity of the test. It is not hard to see that simply permitting a large proportion of students to copy answers or consult each other on test questions is an extremely effective way of increasing scores (Frary & Olson, 1985). Cheating could result in an inaccurate evaluation of achievement, and in the long run could invalidate college degrees. Moreover, answer copying during multiple-choice examinations can be a threat to the equity of testing procedures (Houston, 1983).

A Method to Examine Allegations of Copying

Suppose that students C and S took a multiple-choice test with n items. If out of n items they gave the same responses to m items, can we infer that they matched those m items by chance? Or did C copy from S?

The assumption of local independence, also made in Item Response Theory (IRT), will be made here. To quote Hambleton, Swaminathan, &, Rogers (1991),

> The property of local independence means that for a given examinee (or all examinees of a given ability value) the probability of a response pattern is equal to the product of probabilities associated with the examinee's responses to the individual items (p. 10-11).

Suppose $p_1, p_2,..., p_n$ are the probabilities of C choosing the responses that S made to n items. Assuming that responses of S are considered fixed, the distribution of the number of items matched with S is the same as that of the number of white balls obtained when one ball is drawn at random from each of n urns, the probability of drawing a white ball from urn i being p_i, $i = 1, 2,..., n$ (Kendall & Stuart, 1958 cited in Lord and Novick, 1968). Thus, the distribution of the number of matches with S is the compound binomial (Lord & Novick, 1968)

$$b_n(m) = \sum_{\Sigma u_i = m} \left(\prod_{i=1}^{n} p_i^{u_i} q_i^{1-u_i} \right) \qquad (1)$$

where $u_i = 1$ if the response of C to item i is the response chosen by S, and 0 otherwise, p_i is the probability that $u_i = 1$, and $q_i = 1 - p_i$. The large summation is over all patterns of response containing exactly m matches. Note that if all p_i's are equal, $b_n(m)$ would be reduced to the binomial distribution and the right side of equation 1 would be

$$\binom{n}{m} p^{m} q^{n-m}$$

From equation 1, the probability of matching at least m items with S is

$$B_m = \sum_{j=m}^{n} b_n(j) \qquad (2)$$

To make a decision whether or not C and S have unusually high similarity of responses a cutoff value of the number of matches has to be established. To do this two types of potential errors must be considered. False positive errors (also referred to as type I error) occur when a pair of examinees are accused of copying when copying did not take place. False negative errors (also referred to a type II error) occur when a pair of examinees who copied are not accused of doing so. A higher cutoff on the number of matches yields higher false negative error rate and lower false positive error rate. The strategy in setting the cutoff involves specifying a false positive error rate and choosing a cutoff value that corresponds to it.

Suppose one is willing to take the risk of wrongly accusing a student of cheating at a rate of α. That is, Prob(false positive error) = α. In the process of cheating detection, the paper of a student is compared with a number of students who are in close proximity during the exam. If there were k students that C could have copied from, then C's responses would be compared to these k students, thus making k comparisons. Let α^* be the probability of making a wrong decision about C (*i.e.*, accusing him/her of copying if in fact he/she did not) in one comparison. Assuming that the comparisons are independent, then the probability of at least one wrong decision about C in k comparisons is $1 - (1 - \alpha^*)^k$. Letting $\alpha = 1 - (1 - \alpha^*)^k$, we get $\alpha^* = 1 - \sqrt[k]{1 - \alpha}$. In this case, α^* is the actual false positive error rate for each comparison. The cutoff score then should be m^* , where m^* is the smallest value such that the probability of m^* or more matches is less than or equal to α^*. Thus, if C and S matched m items, a case should be made against C for unusually high similarity of responses with S if m is greater than or equal to m^*. In other words, C should be accused of copying from S if $B_m \leq \alpha^*$.

In this study, k is called the proximity index; *i.e.*, the number of examinees from whom C could have copied. The value k would be small if security measures were taken; *e.g.*, alternate forms were used, seating was assigned, and several proctors were present. A situation where the value of k would be large is a take-home exam. Moreover, a larger value of k gives a more conservative cutoff.

The binomial distribution is the one that is very often used in cheating indices (*e.g.*, Belleza & Belleza, 1989; Cody, 1985). In this study, however, the compound binomial distribution is being employed. The main reason is that the compound binomial is more realistic in that it allows the p_i's to be different for different items. The difficulty of computing compound binomial probabilities might turn one away from it. However, using the recursive procedure introduced by Lord and Wingersky (1984) as shown below, a computer can readily determine $b_n(m)$ even for several hundred items.

It is important to reiterate that all the compound binomial probabilities that have been discussed are conditioned on the responses of the source. For the purpose of computing $b_n(m)$ only the responses of C are considered random. The responses of S are considered fixed. The probability of matching S's response on item i (p_i) is estimated by the "response 'difficulty'" (Hanson, Harris, & Brennan, 1987, p. 10); that is, dividing the number of students that gave the same response as S on item i divided by the number of students that responded to item i.

Method

Data. Three hundred and sixty (360) simulated data sets reflecting twelve different cheating situations were obtained and used for this study. From an ACT Assessment mathematics data set that consisted of over 3,000 examinees' responses to 50 items, four different types of data sets were obtained. The first type, DATA3, are responses to the first 20 items of 100 examinees who were sampled randomly. The second type, DATA4, are the responses to the first 50 items of a different sample of 100 examinees. The last two types, DATA5 and DATA6, are responses of samples of 200 examinees to the first 20 and first 50 items, respectively. All the samples were obtained using random sampling. For each type of data set 90 samples were drawn. (Please refer to Figure 1 for a graphical summary of the sampling, cheating simulation, and analyses of the simulated data sets.)

Within each sample, examinees were ordered so that any pair coming from the same test center would not be within close proximity of each other when the cheating simulation was performed. This made sure that in cheating detection pairs of examinees compared do not include those from the same test center.

For each of DATA3 to DATA6 the 90 data sets were divided into three groups of thirty. The groups were denoted DATAn.1, DATAn.3, and DATAn.5, for n = 3, 4, 5, 6. For each group of 30 data sets a cheating situation was simulated. All of these data sets were described in Table 1. For DATAn.m, n = 3, 4, 5, 6, m = 1, 3, 5, and N = class size, .1mN students were randomly selected to be cheaters. Of the (copier and source) pairs created, .02mN or one-fifth of the cheaters copied 10, 25, 50, 75, & 90 percent of the source's answers, respectively. That is, the cheaters' answers for selected items were replaced with those of the sources' answers. The selection of the copied items was random. For example, in DATA3.5 there were 50 cheaters. For each of the first ten cheaters answers to two (10%) randomly selected items were changed to those of the source; for each of the second ten cheaters answers to five (25%) randomly selected items were changed to those of the source; 10 items (50%) for each of the third ten; 15 items (75%) for each of the fourth ten; and, 18 (90%) items for each of the last ten. Note that each of DATA3.1 through DATA6.5 had 30 replications.

Research Questions. The research questions to be addressed in this study are the following.

1. How sensitive is B_m in detecting copying?
2. How is the effectiveness of B_m affected by the percent of cheaters?
3. How is the effectiveness of B_m affected by the length of test?

4. How is the effectiveness of B_m affected by the class size?

5. What are the similarities and differences among B_m, g_2, and ESA?

6. How do the effectiveness of the three indices compare?

7. Are the effects of percent of cheaters on the effectiveness the same for all indices?

8. Are the effects of test length on the effectiveness the same for all indices?

9. Are the effects of the class size on the effectiveness the same for all indices?

The false positive rate (α) chosen for all cheating detection analyses in this study was .002. The above value is much lower than what is more commonly used in scientific studies (*i.e.*, .01 and .05). The rationale for choosing such a low significance level is consistent with the judgement that in investigation of allegation of cheating "false positive errors have more serious consequences than false negative errors." (Brennan, 1993)

Analyses. To evaluate the effectiveness of the new index B_m an analysis was done on each data set. A computer program in C language was written for this purpose. The input needed to run the program are: (1) the number of examinees who took the test, (2) the number of items on the test, (3) identification of each examinee, (4) the number of options per item (assumed to be uniform), (5) the answer key, (6) the item responses of the examinees, (7) the proximity index k, and (8) the level of significance α. The identification of examinees are codes used only for matching. The printout of the results was in tabular form with eight columns consisting of identification of copier and source and their respective scores on the test, the number of matches (r, w, and m), and the computed value of B_m. Due to the large number of all possible pairs not all of them were included in the printout. Only the pairs such that $B_m \leq \alpha^*$ were included in the printout. For each data set, only the simulated cheaters were considered cheaters. If a noncheater was detected by an index that was considered a false positive or type I error. A cheater not detected was counted as false negative or type II error.

According to Frary *et al.* (1977), "A single way to test the effectiveness of [a cheating index] is to count the number of cheaters 'caught'.... (p. 244)." Thus, to answer the questions enumerated above (except for question number 5) the data sets were analyzed using the three different methods and the number of cheaters "caught" were counted and tabulated. The number of noncheaters falsely caught were also reported. To analyze the data sets using ESA (Belleza & Belleza, 1989) and g_2 (Frary *et al.*, 1977), the computer programs used by the respective authors for their studies were used for the present study. Both computer programs were written in FORTRAN and are available from the original authors. For all three cheating detection analyses, the value of k was 4.

Each replication of each cheating situation was analyzed using each of the three cheating detection procedures. To tabulate the results of the 1080 analyses performed, the following algorithm was used:

Let N = the number of students

n = the number of cheaters[1]

Any pair of students whose cheating index is ≤ .0005[2] and the difference of the last three digits of the student numbers (copier's minus source's) is positive but less than or equal to four (4) were flagged. Each of the flagged pairs belonged to exactly one of seven categories. The number of pairs belonging to each category is as follows:

a_2 = the number of pairs who were flagged such that:

a) The copier copied 10% of the source's answers (*i.e.*, the first digit of the copier's student number (SN) is 2).

b) The last three (3) digits of the source's SN is exactly one (1) less than the copier's SN.

a_3 = the number of pairs who were flagged such that:

a) The copier copied 25% of the source's answers (*i.e.*, the first digit of the copier's student number (SN) is 3).

b) The last three (3) digits of the source's SN is exactly one (1) less than the copier's SN.

[1] The cheating simulation program was written so that each examinee has a four-digit ID. The last three digits of the ID is the virtual seat number. That is, the examinees are theoretically seated in a single file facing front such that anybody in front of any student has a lower three-digit code. The first digit of the ID is the indicator of whether the examinee is a simulated copier or not. An examinee whose first digit ID is 1 is not a copier. A first digit of 2, 3, 4, 5, or 6 indicates that the examinee copied 10, 25, 50, 75, or 90 percent of the responses of the examinee in front of him/her.

[2] This is equivalent to $g_2 \geq 3.29$.

a_4 = the number of pairs who were flagged such that:

a) The copier copied 50% of the source's answers (*i.e.*, the first digit of the copier's student number (SN) is 4).

b) The last three (3) digits of the source's SN is exactly one (1) less than the copier's SN.

a_5 = the number of pairs who were flagged such that:

a) The copier copied 75% of the source's answers (*i.e.*, the first digit of the copier's student number (SN) is 5).

b) The last three (3) digits of the source's SN is exactly one (1) less than the copier's SN.

a_6 = the number of pairs who were flagged such that:

a) The copier copied 90% of the source's answers (*i.e.*, the first digit of the copier's student number (SN) is 6).

b) The last three (3) digits of the source's SN is exactly one (1) less than the copier's SN.

b = the number of pairs who were flagged such that:

a) The copier copied (*i.e.*, the first digit of the SN $\neq$ 1);

b) (But) the difference between the last three (3) digits of the SNs is > 1.

c = the number of pairs who were flagged such that:

a) The copier did not copy (*i.e.*, the first digit of the SN = 1).

From the above values the following eight quantities were computed:

$$TP_i = \frac{5a_i}{n} \quad , \quad i = 2, 3, 4, 5, 6 \tag{3}$$

$$TP = \frac{\sum_{i=2}^{6} a_i}{n} \tag{4}$$

$$FP_1 = \frac{c}{N - n} \tag{5}$$

$$FP_2 = \frac{c + b}{N - n + b} \tag{6}$$

The values of TP_i, for i = 2, 3, 4, 5, and 6 are the proportions of cheaters who copied 10, 25, 50, 75, and 90 percent of their sources' answers, respectively, who were caught for each replication of each cheating situation. TP is the overall true-positive rate, and FP_1 and FP_2 are false positive rates. Note that b is the number of copiers who were caught but with the wrong source; and that if b = 0, FP1 = FP2. To answer questions one through four, each replication of DATA3.1 through DATA6.5 was analyzed using the three computer programs mentioned above. The number of cheaters detected and the number of noncheaters falsely detected for each replication were counted. The mean, standard deviation of the hit rates, and observed false positive rates across replications for each of DATA3.1 through DATA6.5 were computed.

Results

The original data set has a KR20 reliability of .88. (Please see Table 2.) The scores range from 6 to 50, with mean and standard deviation of 23.05 and 10.04, respectively. The item difficulties range from .22 to .86, with mean and standard deviation of .54 and .14, respectively. The truncated version of the original data set (*i.e.*, including only the first 20 items) has a KR20 reliability of .75. The scores range from two to 20, with mean and standard deviation of 13.40 and 4.13, respectively. The item difficulties range from .53 to .86, with mean and standard deviation of .67 and .09.

Simulated data sets were used to evaluate the effectiveness of the new index B_m. The simulated data sets mentioned above were also used to make empirical comparisons between B_m, g_2 (Frary *et al.*, 1977), and error-similarity analysis (ESA) (Belleza & Belleza, 1989). In addition, theoretical comparisons were made among the three indices.

In general, it was found that B_m is effective in detecting cheaters as long as enough items have been copied. That is, cheaters do not usually get caught if they only copied a few items. More specific findings were reached regarding the sensitivity and effectiveness of B_m. These are discussed below as responses to the research questions enumerated earlier. The percent of cheaters that were used are averages across the 30 replications of each cheating situation.

(1) *B_m is sensitive enough so that cheating becomes detectable when the copier copied between 12 (25%) and 25 (50%) items in a 50-item test. It is, however,*

less sensitive in detecting cheating when the test is shorter. The copier would have to copy between 50% (10) and 75% (15) of the source's responses to a 20-item test before cheating is detectable.

Question one asks how many items need to be copied in order for cheating to be detectable. To answer this question, it was necessary to operationalize the term "detectable". For the purpose of this study "detectable" means having more than 50% likelihood of catching a cheater. Notice in Figure 2 that in each cheating situation the percent of cheaters caught (PCC) is monotonically increasing as the percent of items copied (PIC) increases. Although the PIC at which cheating becomes detectable cannot be determined, the interval of PIC within which it occurs can be determined. Using Figure 2 notice that for each cheating situation with 20 items (*i.e.*, DATA3.n and DATA5.n, for n = 1, 2, and 3), cheating becomes detectable between 50 and 75 PIC, irrespective of class size (CS). On the other hand, in situations with 50 items (*i.e.*, DATA4.n and DATA6.n, for n = 1, 2, and 3), cheating becomes detectable between 25 and 50 PIC.

(2) B_m is more effective in detecting cheating when there are fewer cheaters in the class.

In response to question two, cheating detection using B_m is more effective when the cheater to noncheater ratio is lower; *i.e.*, when the percent of cheaters (POC) in a cheating situation is lower. This is indicated by the general pattern of increasing PCC as POC increases from 10 to 30, and to 50 in Figure 2.

(3) B_m is more effective in detecting cheating in longer tests than in shorter tests.

Clearly in Figure 3, for PIC levels of 25, 50, and 75, within each level of POC, and within each level of CS, PCCs are higher for situations with 50 items (represented by □ and ◊) than for situations with 20 items (represented by ■ and ♦). Such observation is also true for PIC of 10, although it is not clear in the graphs due to small differences between respective PCCs. This is a strong indication that B_m is more effective in detecting cheating when the test is longer. At PIC level 90, however, there is an exception to this generalization when at POC of 50% and CS of 200, PCC is higher for DATA5.5 (with 20 items) than it is for DATA6.5 (with 50 items). Moreover, at PIC levels of 10, and 90, the PCCs are not very different in numerical values within each level of POC.

(4) The effectiveness of B_m is not affected by class size.

The comparisons made were between DATA3.r and DATA5.r, and between DATA4.r and DATA6.r, for r = 1, 3, and 5, for each PIC within each level of POC. Each DATA3 and DATA5 situation has 20 items, and each DATA4 and DATA6 situation has 50 items. In terms of symbols used in the graphs, comparisons were made between shaded symbols (■ and ♦), and also between unshaded symbols (□ and ◊).

A very obvious observation is that, except in three (out of thirty) comparisons made, the comparable symbols are overlapping, if not coinciding. And that for each pair of non-overlapping comparable symbols, they are not very far apart. That is, for each level of PIC within each level of POC, in situations with the same TL, the differences between PCCs are very small. This is an indication that CS has no effect on the effectiveness of B_m.

To address question five it is deemed necessary to recall the following:

Suppose student C is suspected of copying from student S, and the observed number of items that they responded to in the same way is m.

- B_m is the compound binomial probability that C and S responded to at least m items by chance. The smaller the value of B_m the stronger the evidence that C copied from S.
- g_2 is the difference between the expected number of matches and the observed number of matches divided by the standard deviation of the number of matches. Assuming no cheating went on, it should have an approximately normal distribution. Thus, the larger g_2 is the stronger the evidence that C copied from S.
- In ESA, the binomial distribution of the number of identical responses given by C and S out of all the items to which they both responded incorrectly. The smaller the probability, the stronger the evidence that C and S collaborated.

(5) *The theoretical similarities and differences among the three cheating indices are the following:*

a) Both B_m and g_2 consider all items in the test when detecting cheating, whereas ESA uses only the items to which both the copier and the source responded incorrectly.

b) B_m uses the compound binomial distribution of the number of matches, while g_2, although derived from a compound binomial distribution is an approximation of the standard normal distribution. The ESA uses both the binomial distribution and the normal approximation depending on the number of items to which both the copier and the source responded incorrectly.

c) Both B_m and g_2 treat the source's responses as fixed and the copier's responses as random, whereas ESA treats both the source's and the copier's responses as random.

d) The estimation of the probability of a particular response differs for all three indices. However, g_2 is the only one that varies the probability estimate as a function of the estimated ability of the copier.

The differences discussed above are possible sources of the differences in performance and appropriateness of the three indices in cheating detection. The investigation of research questions six through nine focused on the performances of the three indices based on their effectiveness in different cheating situations.

(6) Based on true-positive rates, ESA is the least effective of the three indices in detecting copying, and the corresponding true-positive rates of B_m and g_2 do not differ by much. However, B_m has the highest observed false-positive rate. The observed false-positive rates for g_2 and ESA are very close to the nominal false-positive rate (i.e., α).

For each of DATA3.1 through DATA6.5, a figure similar to Figure 4 was produced to address question number six. For discussion, only the graphs for

DATAn.1 for n = 3, 4, 5, and 6 were included. Those are Figures 4, 5, 6, and 7. The graphs for DATAn.3 and DATAn.5 were very similar to DATAn.1 for each n = 3, 4, 5, or 6.

In each of Figures 4, 5, 6, and 7, the PCC for each of the cheating indices were compared within each level of PIC. The general observation is that for data sets with 20 items (Figures 4 and 6), at each of PIC levels 25 through 90, the PCC for ESA is much lower than those for g_2 and B_m, and that the PCCs for g_2 and B_m are not very different in values. For data sets with 50 items (Figures 5 and 7), the PCCs are still generally the lowest for ESA. That observation is not true, however, at PIC levels 10 and 90. At PIC level of 10, the PCC values are very similar. At PIC level 90, notice that PCC for g_2 is lower than those of the two indices, and that the it is lower than the PCC at PIC level of 75. This phenomenon will be discussed later in this section.

In general, therefore, based on the PCC, ESA is the least effective in catching cheaters, and that the effectiveness of g_2 and B_m are about the same. However, if we examine Figure 8, it tells us that B_m has the highest rate of catching noncheaters.

(7) *The effect of percent of cheaters on the effectiveness of the indices are generally the same for B_m and g_2, but different for ESA. Indices B_m and g_2 are more effective in detecting cheating when there are less cheaters in the class. The effectiveness of ESA is not very much affected by percent of cheaters.*

To address question number seven, graphs similar to those in Figure 9 were produced for each combination of number of items and number of students. For discussion the combination of 20 items and 100 students, and 50 items and 100 students were used. The graphs for data sets with 200 students were very similar to those with 100 students, for the same number of items.

Recall, that for B_m there was a general observation that PCC decreases as POC increases. Also, the decrease in PCC is more apparent when the POC increases from 10 to either 30 or 50. By examining Figures 9 and 10, notice that such general pattern can also be observe in g_2, but not as observable in ESA.

(8) The effect of test length on effectiveness are the same for all indices. All indices are more effective in detecting cheating in longer tests than in shorter tests.

To recall, it was found that B_m is more effective in detecting cheating when there are 50 items than when there are 20 items. This was indicated by the higher PCC for DATA4.r (represented by □) as compared to that for DATA3.r (represented by ■), and higher PCC for DATA6.r (represented by ◊) as compared to that for DATA5.r (represented by ♦), for each r = 1, 3, and 5, at PIC levels 25, 50, and 75. Those observations are also true for g_2 and ESA as can be seen in Figure 11. (A similar figure was made for each POC level, and a similar pattern was observed.) That is, for each index, within each level of POC and each level of CS, PCCs are higher for situations with 50 items than for situations with 20 items at PIC levels 25, 50, and 75.

At PIC level 10, the values of PCC are not very different within each level of POC, for all indices. At PIC level of 90, the values of PCC do not vary very much within each level of POC for B_m and g_2. For ESA, however, it is very clear that at PIC level 90, PCCs are higher for the situations with 50 items than they are for situations with 20 items. Moreover, for ESA, the values of PCC within each level of POC are not very different from each other at PIC level 25. Such is not true for the other two indices.

In general, the effect of test length on effectiveness is the same for all indices; *i.e.*, each index is more effective in detecting cheating when there are more items. However, this effect is more apparent at PIC levels 50, 75, and 90 for ESA, whereas it is more apparent at PIC levels 25, 50, and 75 for B_m and g_2. This is probably because of the already very high PCC for B_m and g_2 at PIC level 90 even when there are only 20 items.

(9) Class size does not affect the effectiveness of any of the three indices.

To recall the investigation of research question number four, it was found that CS has no effect on the effectiveness of B_m. This was indicated by the observation that at each level of PIC, within each level of POC, in situations with the same TL the pairs of PCCs are very close in numerical value. To address research question number nine it was investigated whether the same observation applies to g_2 and ESA as well. By examining Figure 11 and comparing DATA3.r to DATA5.r (■ *vs.* ♦) and DATA4.r to DATA6.r (□ *vs.* ◊), for r = 1, 2, and 3, it is clear that CS has no effect on the effectiveness of g_2 and ESA either. This is indicated by the very close numerical values of PCCs for DATA3.r and DATA5.r, and DATA4.r and DATA6.r, for r = 1, 3, and 5, for each level of PIC for all indices.

A Counter-Intuitive Result Using g_2. The question regarding the effectiveness of an index addresses the issue of how many of the source's responses need the copier copy so that cheating is detectable. Built into this concern is the intuitive notion that the more items that are copied the greater the likelihood that the cheater will be caught. However, such pattern was not observed when using g_2 in situations where there are 50 items. More specifically, for DATA4.r and DATA6.r, r = 1, 3, and 5, PCC increases as PIC increases from 10 through 75, but decreases as PIC increases from 75 to 90. Two questions come to mind:

(1) What goes on with g_2 between PIC levels of 75 and 90 (when there are 50 items)?

(2) Why doesn't it happen when there are only 20 items?

This concern was brought to the attention of Dr. Frary. His spontaneous answers were:

(1) It's possible that the approximation "breaks down" when there are too many items copied.

(2) An increase of PIC from 75 to 90 involves a difference of about eight items in a 50-item test, whereas it involves only a difference of three items in a 20-item test. (Personal communication, April 7, 1994)

A possible explanation that is consistent with Dr. Frary's comments follows.

It was mentioned earlier that cheating detection is hypothesis testing where the null hypothesis is that there was no cheating, and the alternative hypothesis is that student C copied from student S (in the case of g_2). The power of the test, *i.e.*, the probability of rejecting the null hypothesis when the null hypothesis is false, is analogous to the effectiveness of the index which is the likelihood that a cheater will be caught measured by the proportion of cheaters caught. The effect size is indicated by the number or percentage of items copied from the source. We know that an increase in effect size should increase the power of the test. That is, increasing PIC should increase effectiveness or PCC. However, such is only true if the distribution of the statistic when the null hypothesis is true has the same shape when the null hypothesis is false. There is a possibility that the distribution of g_2 becomes platykurtic as PIC increases. If this is true, it could be that the increase in power due to the increase in effect size (from 75 to 90 PIC) is overcome by the decrease in power due to the change in the shape of the distribution. If that is true, then why does it only happen on 50-item tests but not on 20-item tests? Because the change in effect size involving three items might not have caused as much change in the shape of the distribution of g_2 as the change in effect size involving eight items.

Implications of the Results

The findings of the present study have implications for teachers and test administrators who wish to use a cheating detection index whether it is in an investigation of an allegation of cheating, as a deterrent to cheating when paired with a threat of punishment for cheaters caught, or as a monitoring device to evaluate methods for preventing cheating. If the main concern is effectiveness in catching cheaters, other considerations aside, it will be a toss-up between B_m and g_2. These two indices have very comparable effectiveness as shown in Chapter IV. Furthermore, the effects of percent of cheaters, test length, and class size on their effectiveness are also the same. However, there are two important concerns in cheating detection regarding errors in decision making: (1) falsely exculpating cheaters; and (2) falsely accusing innocent students of cheating. Unfortunately, reducing the probability of one increases the probability of the other. And the obvious choice is to reduce the probability of falsely accusing innocent students for this type of error has more serious consequences. Thus, with this consideration in mind, g_2 is the better choice by far.

Recommendations for Further Research

Since B_m and g_2 are based on the same distribution it was expected that they would perform with similar effectiveness. In addition, it was expected that B_m would perform better since it does not employ an approximation of the standard normal distribution. Thus, there is a need to address why B_m performed so badly with respect to the observed false positive rates.

The only culprit could be the estimation of probability of selecting a particular response which is constant for all copiers for B_m, but is a function of the ability estimate for g_2. It is therefore recommended that a study comparing effectiveness of B_m and g_2 with respect to both true-positive and false-positive rates be done where the same estimation of probability of selecting a particular response will be used for both indices. This will settle the argument as to whether to use standard normal approximation as opposed to computation of compound binomial probabilities.[3] As far as addressing the concern regarding the computer time required in computation, it was observed in this study that there is not much difference. It is possible, however, that the difference would be amplified when the dimensions of the data sets are larger.

Another recommendation for further research is to investigate the breakdown of g_2 when there are too many items copied. One interesting aspect is to find out the percentage of items that need to be copied before the approximation breaks down. But more importantly there is a need to investigate the reason for the breakdown of the approximation in order to improve g_2. Another study that would be valuable is the investigation of whether g_2 could be improved by using other more sophisticated procedures for estimating the probability of a particular response. The possibilities suggested by Hanson (1994) include the use log-linear model or a latent variable model for polytomous items.

Lastly, a study comparing g_2 and B_m similar to the present one should be done with a little twist that could make a lot of difference. That is, "instead of interpreting probabilities produced by the compound binomial model directly, the

[3] Frary (1994) addressed the difference between the effectiveness of g_2 and a compound binomial version of g_2 (cb) with respect to the number of pairs identified as likely copiers, stratifying on the scores of the copying pairs.

distribution of the indices should be computed for benchmark data consisting of pairs of examinees who could not have copied and probability statements should only be made relative to the distribution of the index given by the benchmark data" (Hanson, 1994). Hanson *et al.* (1987) suggests that the probabilities computed from the compound binomial model should not be interpreted directly.

References

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

Brennan, R. L. (1993, April). *Addressing a testing taboo: Discussions on cheating.* Invited discussant for symposium at the annual meeting of the American Educational Research Association, Atlanta, GA.

Collison, M. N-K (1990, October 24). Survey at Rutgers suggest that cheating may be on the rise at large universities. *The Chronicle of Higher Education*, pp. A31-32.

Frary, R. B., & Olson, G. H. (1985, April). *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. (1994, April). *The evaluation of two indices of answer copying and the development of a spliced index.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics, 2*(4), 235-256.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Hanson, B. A. (1994, April). *Statistical indices of response similarity derived from the compound binomial distribution*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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.

Houston, J. P. (1983). Alternative test forms as means of reducing multiple-choice answer-copying in the classroom. *Journal of Educational Psychology, 75*(4), 572-575.

Kendall, M. G., & Stuart, A. (1958). *The advanced theory of statistics* (Vol. I). New York: Hafner.

Lord, F. M., & Novick, M. R. (1968). *Statistical theory of mental test scores.* Reading, MA: Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8*(4), 453-461.

Webster's third new international dictionary of the English language, unabridged, with seven language dictionary (Vol. 1). (1986). Springfield, MA: Merriam-Webster.

TABLE 1
DESCRIPTIONS OF SIMULATED DATA SETS

Data Set	Students	Cheaters	Items	% Copied
DATA3.1	100	10	20	10,25,50,75,90
DATA3.3	100	30	20	10,25,50,75,90
DATA3.5	100	50	20	10,25,50,75,90
DATA4.1	100	10	50	10,25,50,75,90
DATA4.3	100	30	50	10,25,50,75,90
DATA4.5	100	50	50	10,25,50,75,90
DATA5.1	200	20	20	10,25,50,75,90
DATA5.3	200	60	20	10,25,50,75,90
DATA5.5	200	100	20	10,25,50,75,90
DATA6.1	200	20	50	10,25,50,75,90
DATA6.3	200	60	50	10,25,50,75,90
DATA6.5	200	100	50	10,25,50,75,90

Notes:

1. Each fifth of the cheaters copied 10, 25, 50, 75, and 90 percent of the test items.
2. Each data set has 30 replications.

TABLE 2

DESCRIPTIVE STATISTICS OF THE ORIGINAL DATA SET

	Data Sets	
Statistics	DATA3 and DATA5	DATA4 and DATA6
Number of Items	20	50
Number of Examinees	3142	3142
Item Difficulty:		
Minimum	0.53	0.22
Maximum	0.86	0.86
Mean	0.67	0.54
Standard Deviation	0.09	0.14
Score:		
Minimum	2	6
Maximum	20	50
Mean	13.40	23.05
Standard Deviation	4.13	10.04
KR20	0.75	0.88

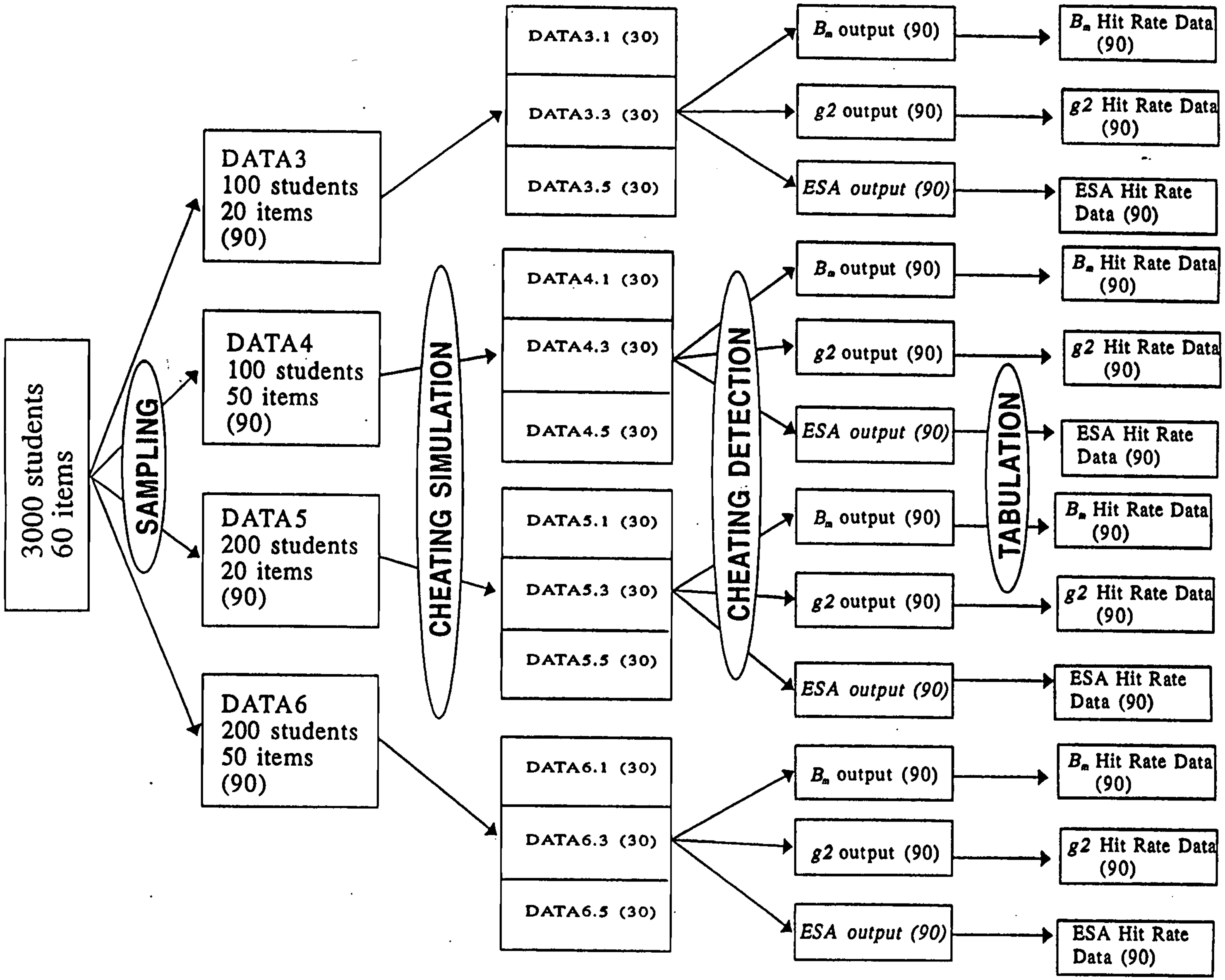

Figure 1. Summary of Procedures for Simulated Data

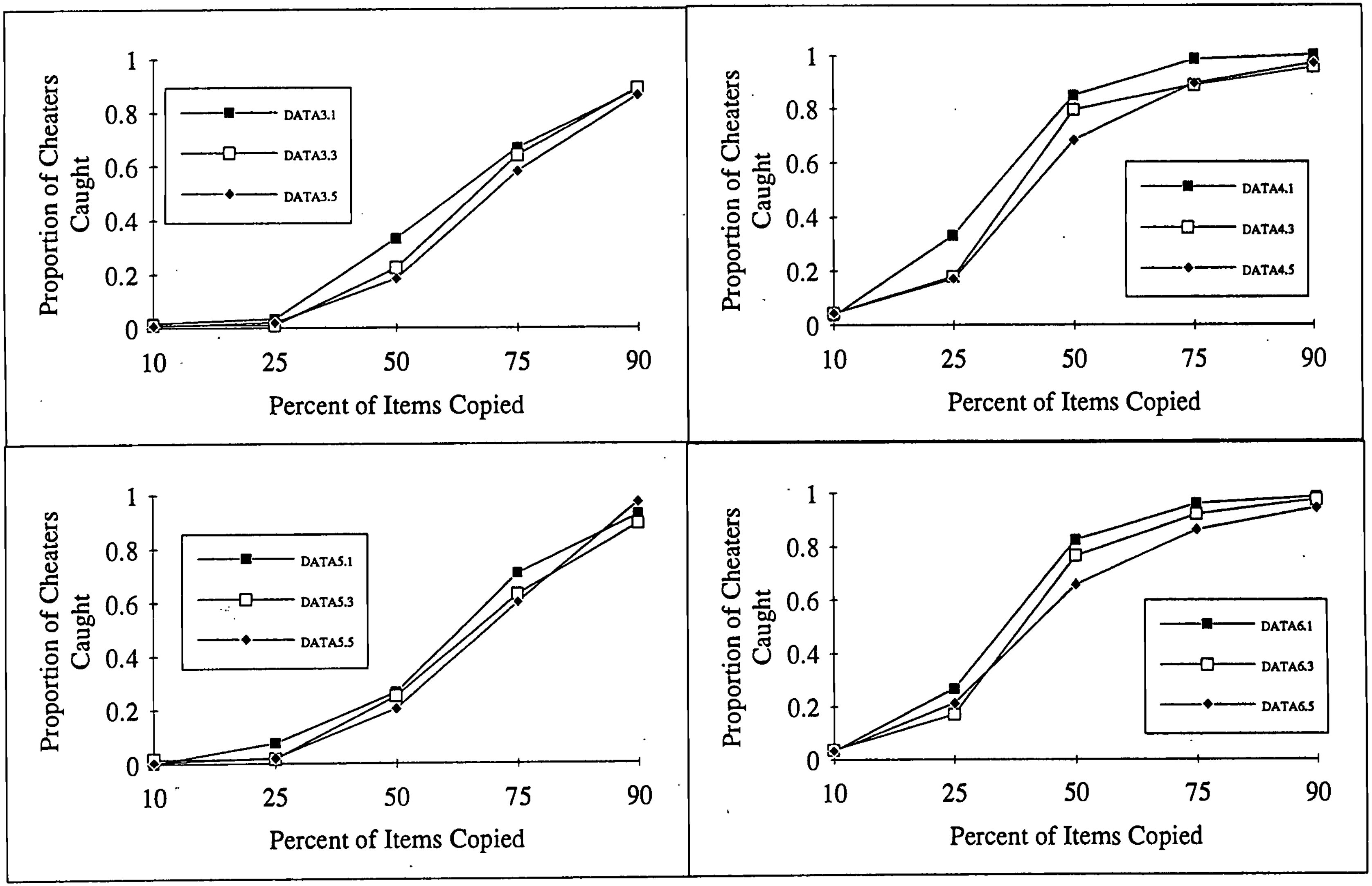

Figure 2. Graphical Results of the Compound Binomial Cheating Detection Analyses

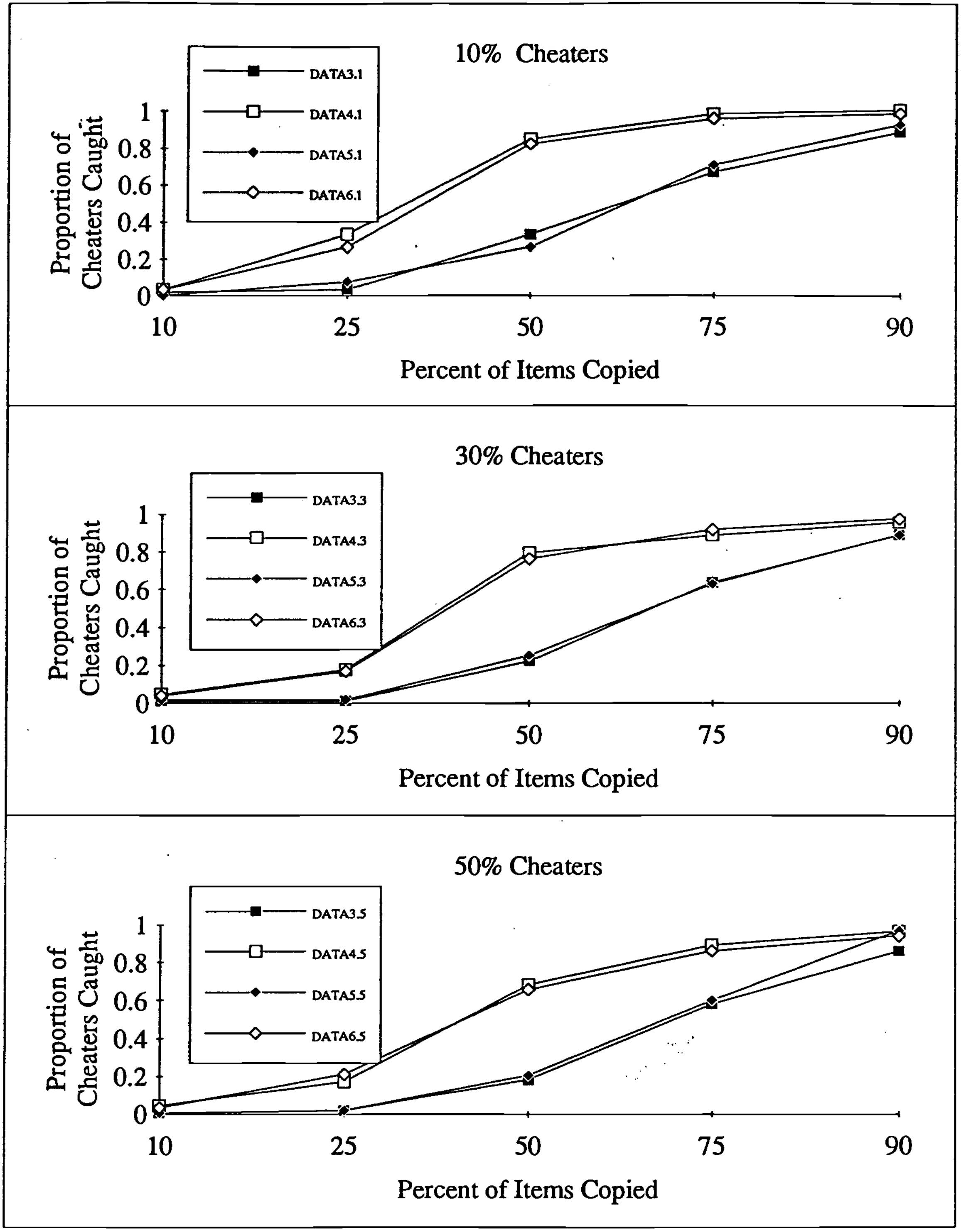

Figure 3. Results of Compound Binomial Analyses Within each Level of Percent of Cheaters

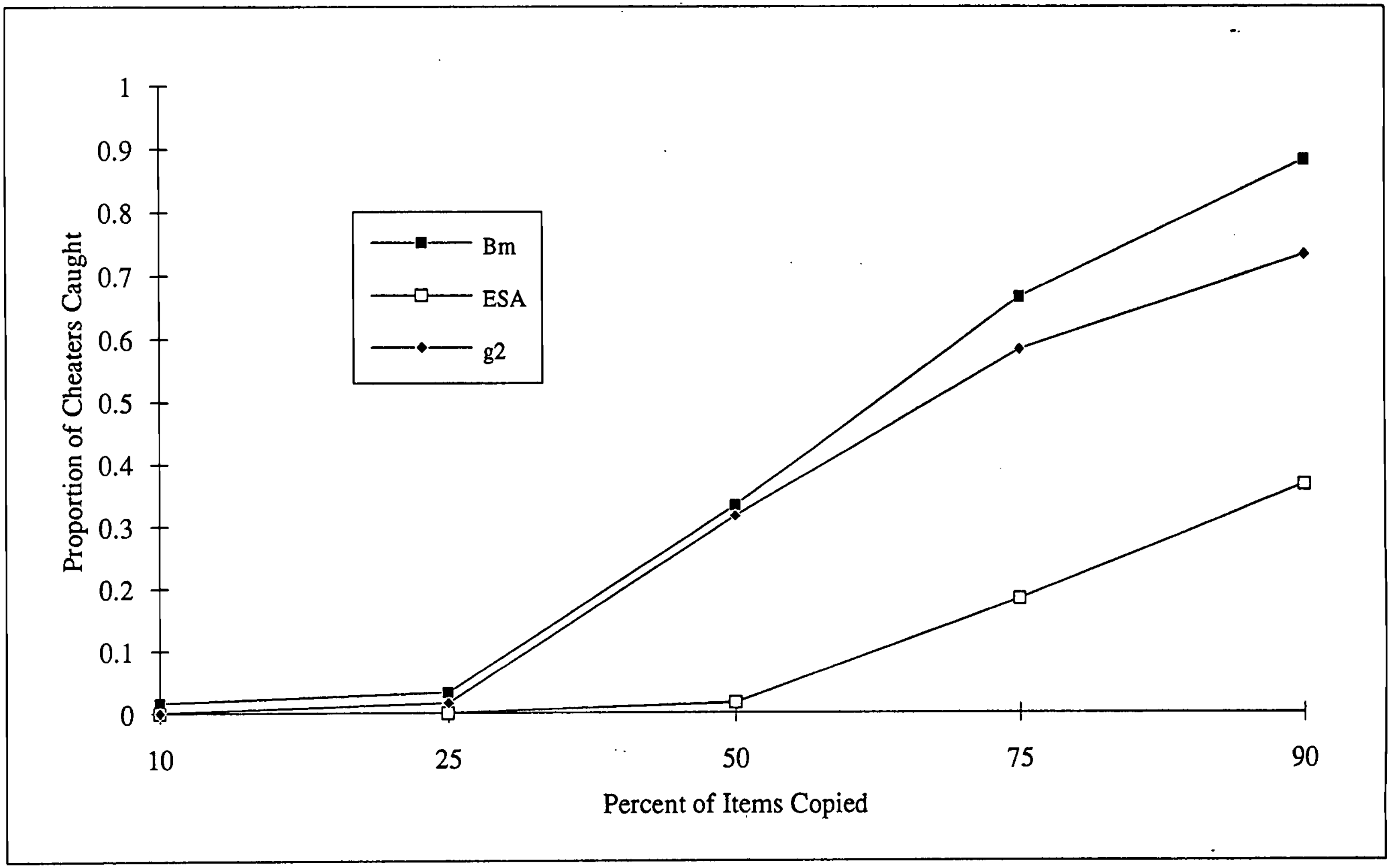

Figure 4. Comparing Performances of Cheating Detection Indices on DATA3.1

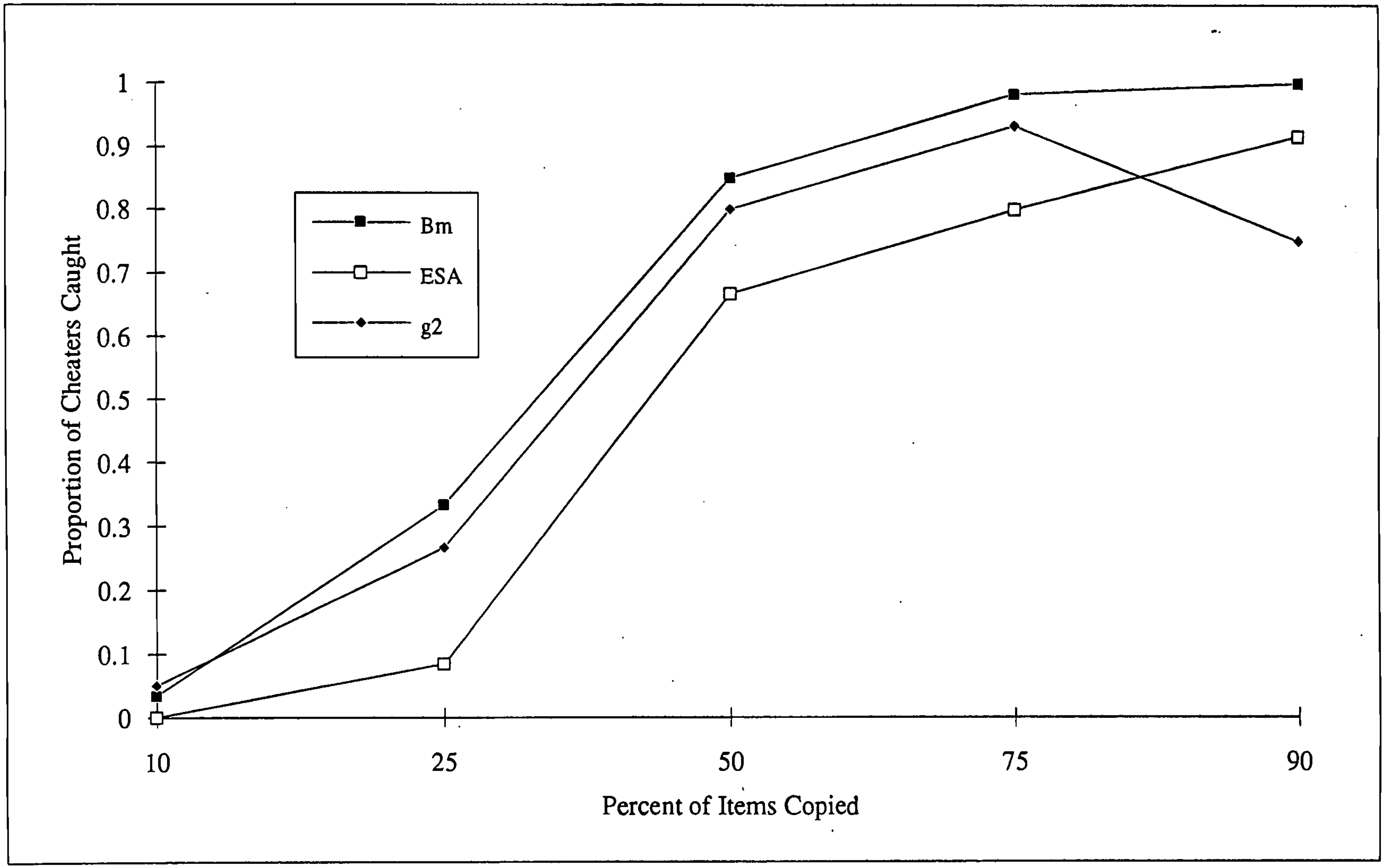

Figure 5. Comparing Performances of Cheating Detection Indices on DATA4.1

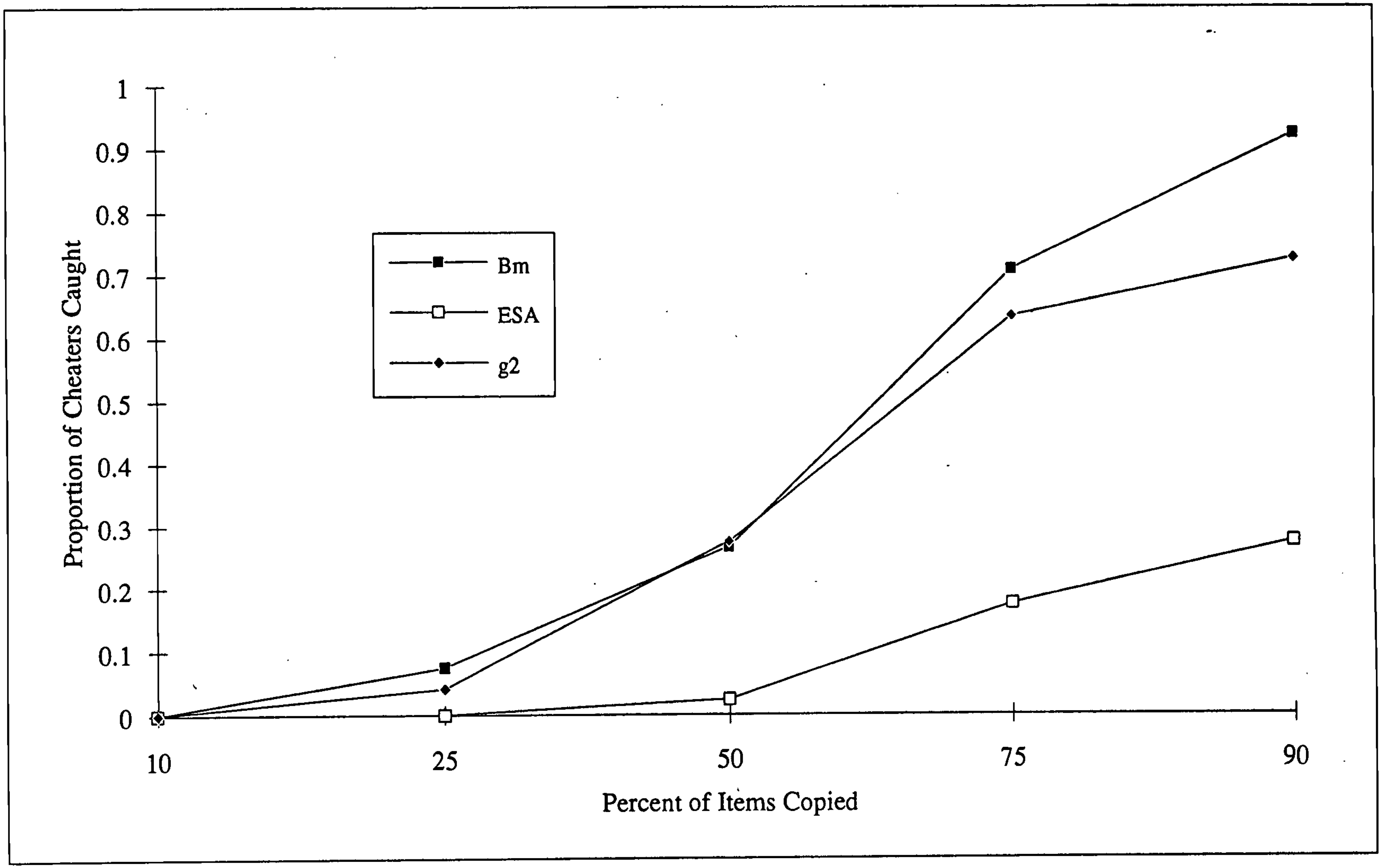

Figure 6. Comparing Performances of Cheating Detection Indices on DATA5.1

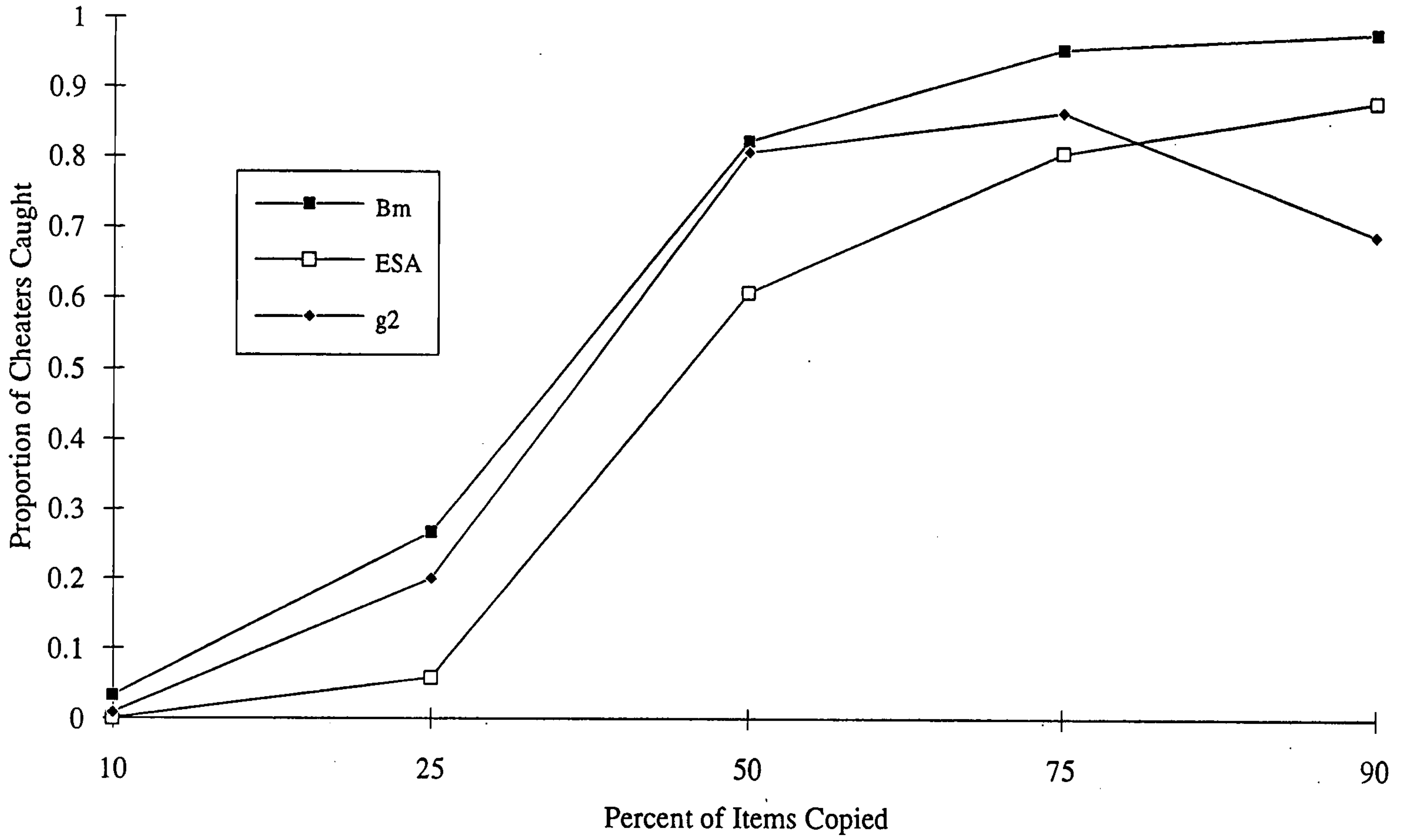

Figure 7. Comparing Performances of Cheating Detection Indices on DATA6.1

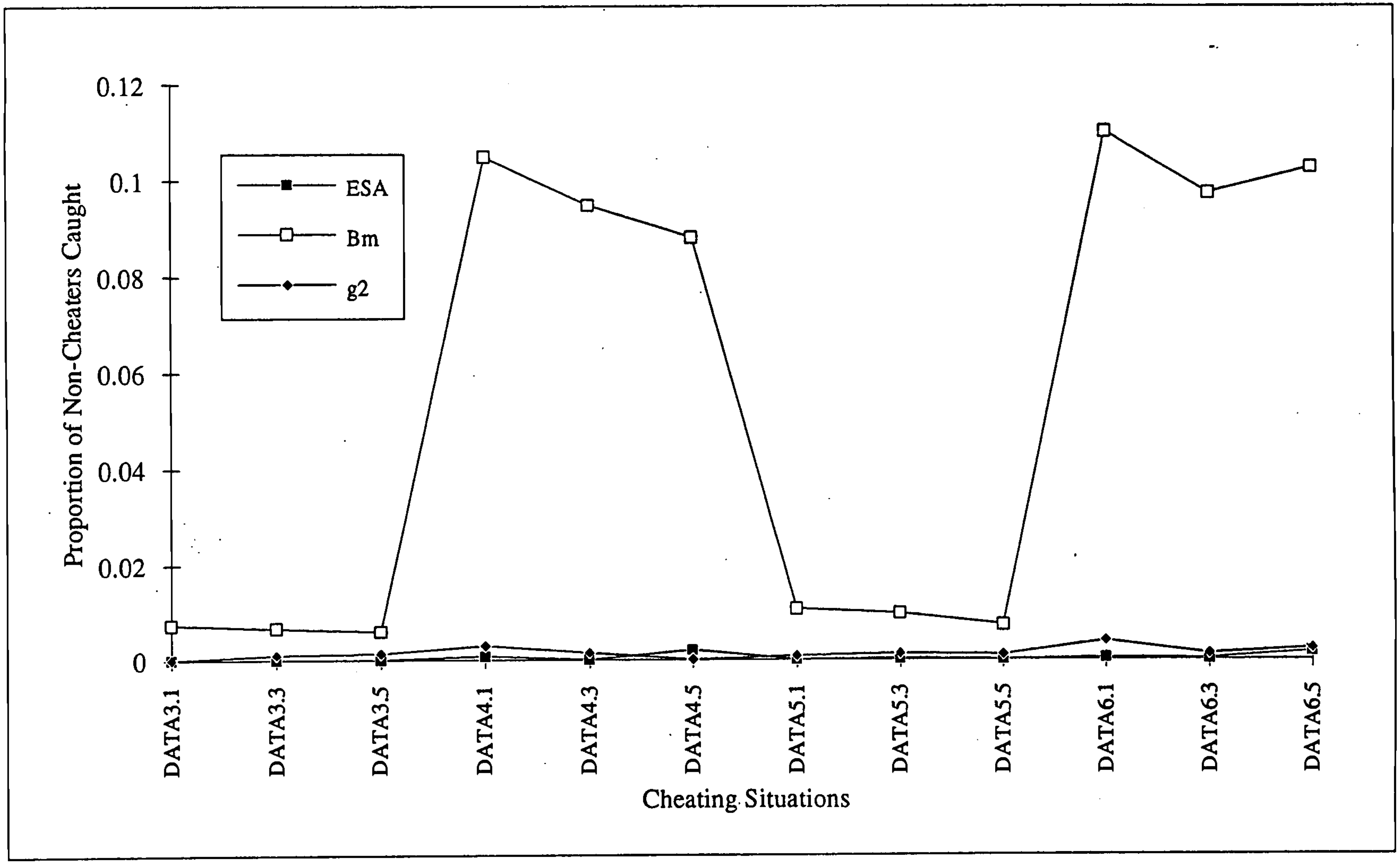

Figure8. False Positive Rates of the Three Indices for Different Cheating Situations

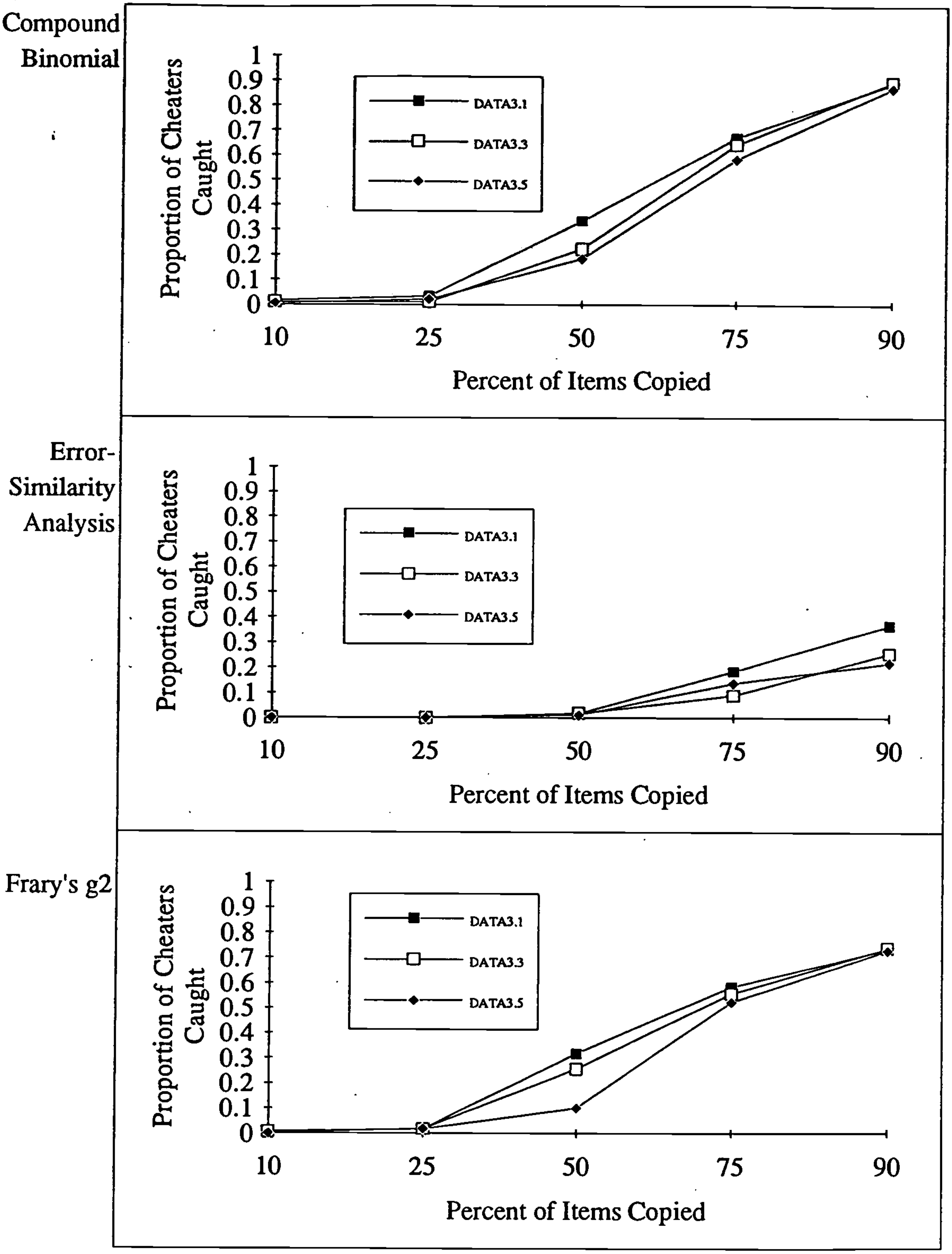

Figure 9. Comparing Cheating Detection Indices for Data Sets with 20 items and 100 Students

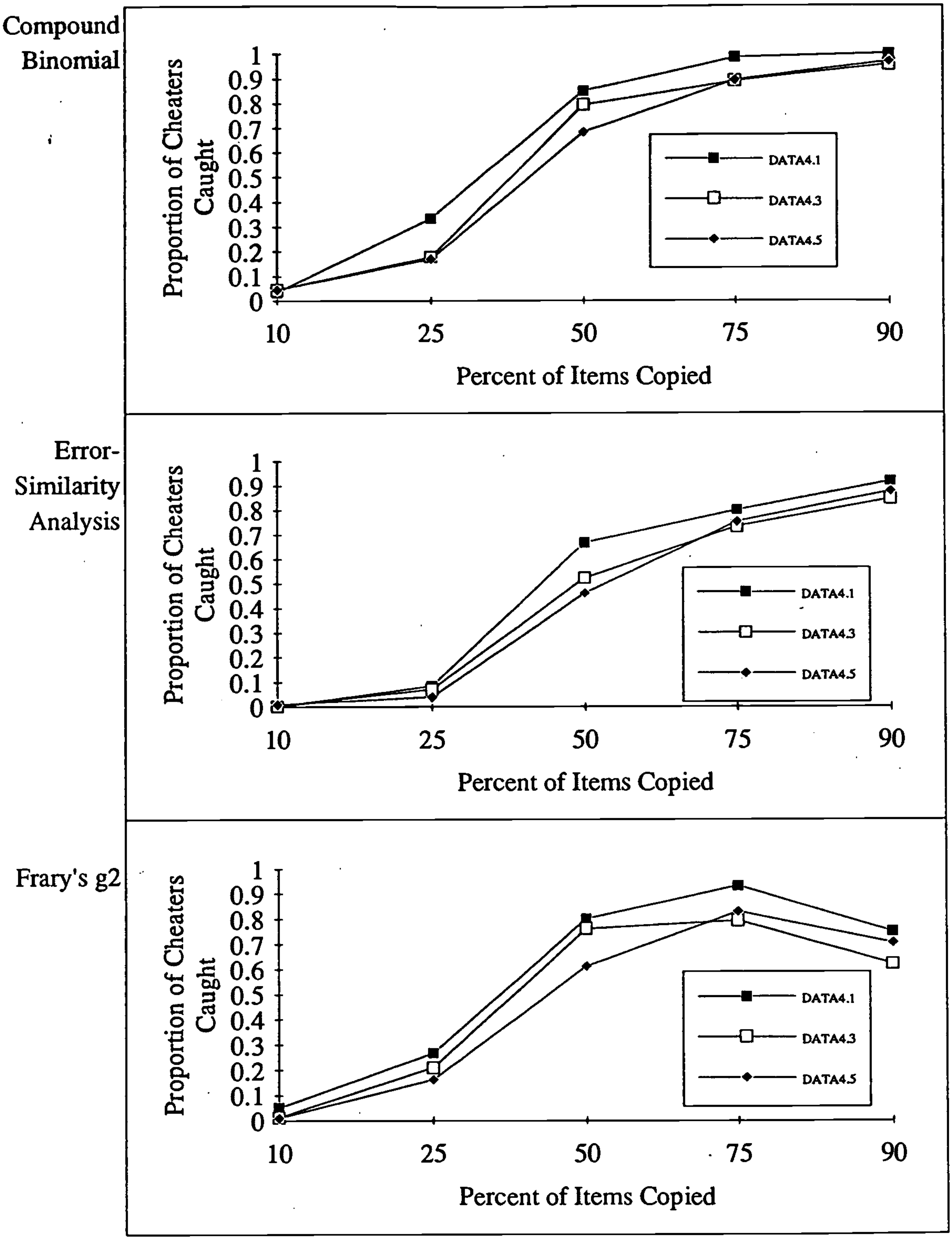

Figure 10. Comparing Cheating Detection Indices for Data Sets with 50 Items and 100 Students

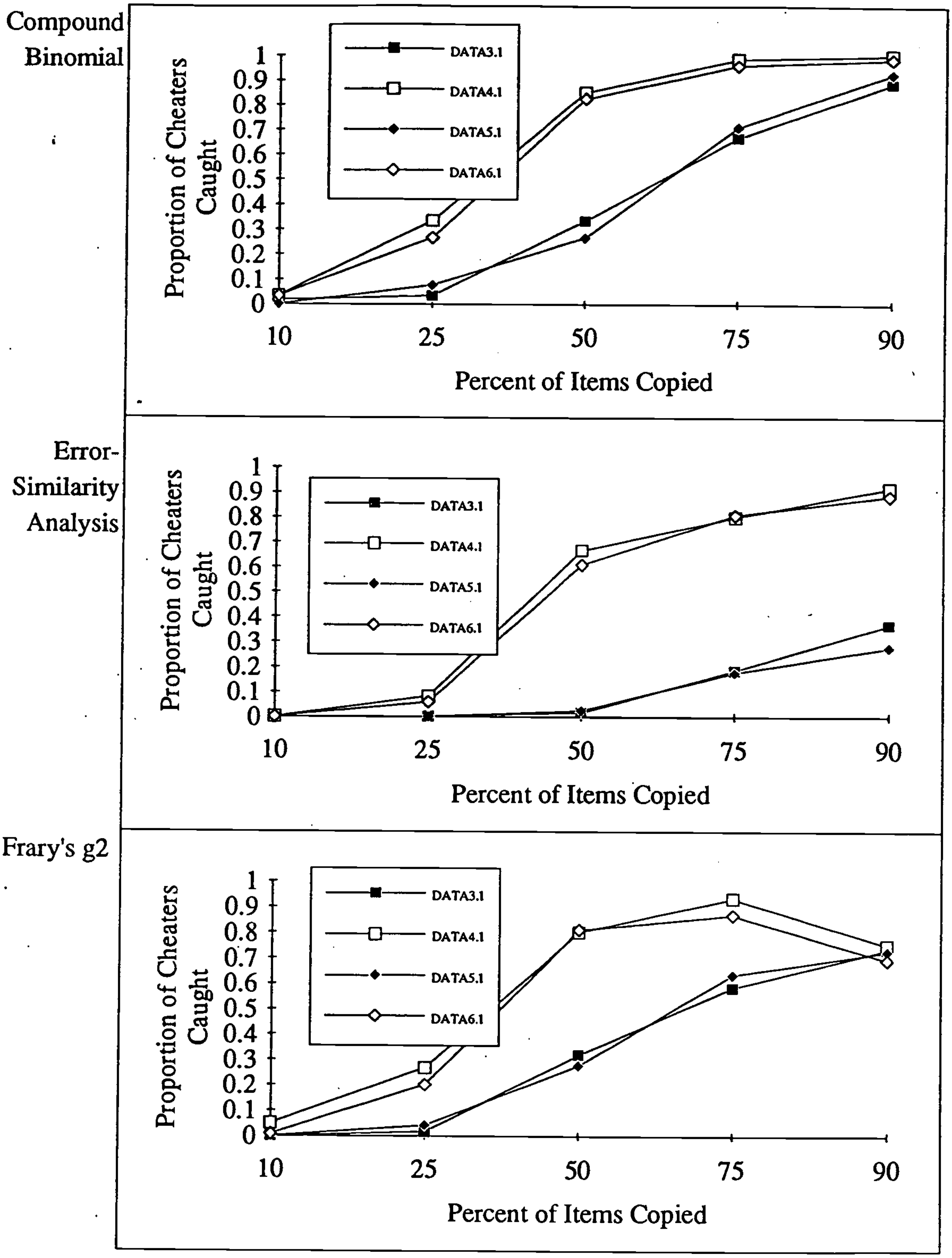

Figure 11. Comparing Cheating Detection Indices Within the 10% Level of Percent of Cheaters

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM028856

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Detection of Cheating on Multiple-Choice Examinations

Author(s): Luz Bay

Corporate Source: American College Testing

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
↑	↑	↑
☒	□	□
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature:	Printed Name/Position/Title: Luz Bay/Psychometrician	
Organization/Address: ACT, 2201 N. Dodge St., PO Box 168 ~~Iowa City, IA 52243-0168~~	Telephone: 319-337-1639	FAX: 319-337-1497
	E-Mail Address: Bay@act.org	Date: 5-11-98

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EVIOUS VERSIONS OF THIS FORM ARE OBSOLETE.