ACT Research Report Series 87-15

A Comparison of Several Statistical Methods for Examining Allegations of Copying

Bradley A. Hanson
Deborah J. Harris
Robert L. Brennan

September 1987

ACT

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

A Comparison of Several Statistical Methods for Examining Allegations of Copying

Bradley A. Hanson
Deborah J. Harris
Robert L. Brennan

Table of Contents

Page

Abstract..iii

STATISTICAL METHODS...1
    Methods Based on Direct Adjustment........................3
    Methods Based on Indirect Adjustment......................3
    Model Based Methods.......................................4

METHOD..7
    Data..8
    Calculation of Statistics.................................8
    Determination of False Positive Rates....................10
    Types of Copying...12
    Conditional False Positive Rates.........................14

RESULTS..16

DISCUSSION...19
    Other Factors in Choosing a Method.......................21
    The Use of Benchmark Data................................23
    Possible Limitations.....................................24
    Cautions...25

CONCLUSIONS..26

References...28

Tables...29

Figures..48

ABSTRACT

Seven statistical methods of investigating an allegation that one examinee copied answers of an examination from another examinee are investigated. Benchmark distributions of the statistics used in each of the seven methods were obtained using data from a 100 item multiple choice licensure exam on 8643 pairs of examinees who could not have copied. The benchmark data were used to determine decision rules for each method corresponding to nine false positive rates. True positive rates corresponding to these decision rules were computed based on simulated copying for 500 pairs of examinees not included in the benchmark data. Five types of simulated copying and five levels of number of items copied (10%, 20%, 30%, 40%, and 50%) were crossed to produce 25 simulated copying conditions. The performance of the methods interacts with the type of simulated copying. For the type of simulated copying thought to be most realistic the methods do not differ greatly in performance, and approximately 5%, 20%, 50%, 85% and 95% of the simulated copiers who copied 10%, 20%, 30%, 40%, and 50% of the items, respectively, could be detected with a false positive rate of .001. Conditional false positive rates were found to be slightly higher than unconditional false positive rates for subgroups of examinees with either high or low test scores, although it is concluded this effect is not strong enough to necessitate use of conditional benchmark data. False positive rates based on theoretical assumptions were not found to agree well with the false positive rates produced using the benchmark data, and it is suggested that benchmark data be used when possible. The limitations in using statistical techniques to investigate allegations of copying are discussed.

A COMPARISON OF SEVERAL STATISTICAL METHODS FOR EXAMINING ALLEGATIONS OF COPYING

Several statistical methods have appeared over the years for investigating allegations of copying (e.g., Bird, 1927; Saupe, 1960; Angoff, 1974; Frary, Tideman, and Watts, 1977; Cody, 1985). This paper compares seven statistical methods for investigating allegations of copying which are based on a direct comparison of the responses of the examinee suspected of copying (suspected copier) and the examinee from whom this person is suspected of copying (source). Other possible statistical procedures for investigating allegations of copying (e.g., appropriateness measures, Levine & Rubin, 1979) are not examined.

Statistical methods such as those to be discussed only provide information on the degree of similarity in the responses of two examinees. They can not provide direct evidence that copying occurred. The perspective taken here is that the statistical methods are best employed when there is collateral information suggesting that copying may have occurred.

The seven statistical methods for investigating allegations of copying will be compared based on their performance in detecting simulated copying. Both the amount and type of simulated copying will be varied. Of main interest is the effect of the amount of copying on the ability of the methods to detect copying, and the relative performance of the different methods for detecting different types of simulated copying.

Statistical Methods

All of the statistical methods of investigating allegations of copying to be considered are based on one or more of five primary statistics, each of which is an indicator of the similarity of responses of a pair of examinees. The five primary statistics are: the total number of items the two examinees answered identically (TJOINT); the total number of items for which the suspected copier and source picked the same incorrect alternative (JII2); the number of items in the longest string (sequence of consecutive items) of identical responses (STRINGL); the number of incorrect responses in the longest string of identical responses (STRINGI1); and the maximum number of incorrect items in any string of identical responses (STRINGI2). STRINGI2 is always greater than or equal to STRINGI1. STRINGI2 can be greater than STRINGI1 when, for example, there is 1 incorrect response in the longest string of identical responses (which may contain 7 items), but there is a another string of 6 items which has 2 incorrect responses. In this case, the value of STRINGI1 will be 1 and the value of STRINGI2 will be 2 (STRINGL would be 7.)

There are two ways of dealing with jointly omitted items in these primary statistics: not counting them, or considering them the same as incorrectly answered items. The data to be used here were chosen such that there were no omitted items for any of the examinees. Therefore, the issue of how to treat jointly omitted items need not be considered for the purposes of this study. As with the above definitions, later definitions will also ignore the issue of jointly omitted items.

The distribution of these primary statistics for pairs of examinees with particular test scores (number of items correct on the test) who did not copy will depend on these particular test score values. For example, it would be expected that the distribution of identically incorrect responses for pairs of examinees for whom copying did not occur and who both had relatively low total test scores would be quite different from the distribution for pairs of examinees who had relatively high total test scores (primarily, the location of the distribution should be higher for lower scoring pairs). The opposite would be the case for the distribution of the longest string of identical responses.

Each of the statistical methods to be considered uses one or more of the primary statistics in a way that attempts to adjust for the relationship of the primary statistics to the test scores of the pair of examinees. To the extent that the adjustment is effective, this procedure allows decision rules to be used independent of the particular test scores of a pair of examinees. The process of adjusting primary statistics for the test scores of a particular pair of examinees is done in three ways: direct adjustment, indirect adjustment, and modeling of the probability distributions.

Methods Based on Direct Adjustment

Angoff (1974) examined eight indices of copying; the two Angoff reports most useful are considered here (Index B and Index H). Both indices are based on the conditional distributions of primary copying statistics conditioned on a function of the test scores of the two examinees. Hence, primary statistics are directly adjusted for the test scores of the pair of examinees. Index B is formed by conditioning the number of identically incorrect responses (JI1I2) on the product of the number of incorrect responses of each of the two examinees (I1I2). Index H is formed by conditioning the maximum number of incorrect items in any string of identical responses (STRINGI2) on the number of incorrect responses for the examinee with the higher test score (MINI).

Methods Based on Indirect Adjustment

Extensive data analysis identified two pairs of statistics based on primary statistics in which a decision about copying for a particular pair of examinees would, in almost all cases, be the same based on the conditional (on the test scores of the pair) or unconditional bivariate distribution of the pair of statistics. Hence, one bivariate distribution could be used to investigate copying for pairs of suspected examinees at any test score values, since considering the pair of statistics together provides an indirect adjustment for the test scores of the pair of examinees in question.

The first method (which will be referred to as PAIR1) is based on the following pair of statistics: the number of identically incorrect responses (JI1I2), and the length of the longest string of identical responses (STRINGL). The second method (which will be referred to as PAIR2) is based on number of incorrect responses in the longest string of identical responses (STRINGI1), and a function of several primary statistics which will be referred to as PJ:

$$PJ = 100 \frac{JI1I2}{NITEMS - (TJOINT - JI1I2)} ,$$

where NITEMS is the number of items on the test. The denominator in the expression for PJ is the number of items for which the pair of examinees in question do not have identical correct responses. Therefore, PJ can be interpreted as the percentage jointly incorrect of the maximum possible jointly incorrect responses. In this study PJ was rounded to the nearest integer.

Model Based Methods

The model based methods use one of two primary statistics: the number of identical responses (TJOINT), or the number of identical incorrect responses (JI1I2). These methods are based on an assumed probability distribution for TJOINT or JI1I2 for non-copying pairs of examinees. At the core of the methods is a model that assumes each person has a probability of responding to each alternative of a particular item when presented with that item (these probabilities will be referred to as the item response probabilities). Further, it is assumed that responses of the particular person of interest to all the items in the test are mutually independent. Given these assumptions, the probability distribution of the number of items the suspected copier will answer in common with any fixed set of responses to the items will be compound binomial. In particular, this holds for the set of responses the source gave to the items. If one knew the probabilities of the suspected copier choosing the same response as that chosen by the source for each item, then the probability of the suspected copier having at least as many identical (or identically incorrect) responses as the source can be computed from the compound binomial distribution. This is the basic idea behind copying indices proposed by Frary, Tideman, and Watts (1977), and Cody (1985). These indices differ in using all responses (Frary, Tideman, and Watts) or only incorrect responses (Cody), and on what they use as the item response probabilities.

Frary, Tideman, and Watts (1977) use two piecewise linear functions of total test score to obtain the item response probabilities. The slopes and intercepts of the linear functions that give the response probabilities for a particular item are functions of the marginal item response "difficulties" over the group of examinees of interest (i.e., the proportion choosing a particular response alternative of the item), number of items on the test, and the mean total test score (this is the point at which the linear pieces are joined). These two piecewise linear functions are given below.

Case 1: Response j to item i is correct.

$$\hat{P}(U_{ia} = j) = p_{ij}\frac{X_a}{\bar{X}} \ , \ 0 \le X_a \le \bar{X}$$

$$\hat{P}(U_{ia} = j) = 1 - (1 - p_{ij}) \left[\frac{N - X_a}{N - \bar{X}}\right], \ \bar{X} < X_a \le N \ , \qquad (1)$$

where U_{ia} is a random variable that takes on the response of suspected copier a to item i, p_{ij} is the observed proportion of examinees that give response j to item i, $\bar{X}$ is the observed mean total test score, X_a is the test score for suspected copier a, and N is the total number of questions on the test.

Case 2: Response j to item i is incorrect.

$$\hat{P}(U_{ia} = j) = p_{ij}\left[\frac{1 - \dfrac{p_{ic}X_a}{\overline{X}}}{1 - p_{ic}}\right], \; 0 \leq X_a \leq \overline{X}$$

$$\hat{P}(U_{ia} = j) = p_{ij}\left[\frac{N - X_a}{N - \overline{X}}\right], \; \overline{X} < X_a \leq N \; , \qquad (2)$$

where p_{ic} is the observed proportion of examinees that give the correct response to item i.

Instead of using the compound binomial distribution to calculate the probability that the number of responses the suspected copier answered identically to the source is at least as great as that observed, Frary, Tideman, and Watts (1977) use the mean and standard deviation of this distribution to calculate a standardized statistic they refer to as g_2. Specifically, the mean is subtracted from the observed number of identical responses and divided by the standard deviation.

Cody (1985) considers only the number of identically incorrect responses. Therefore, only items which the source answered incorrectly are included in computing the compound binomial probability distribution of the total number of identically incorrect responses. Cody uses the item response "difficulties" for the group of examinees of interest as the item response probabilities needed in the compound binomial calculation. Cody suggests using a binomial approximation based on the average item probabilities, although the compound binomial will be used here. The resulting probability of obtaining a value for the number of identical incorrect responses greater than or equal to that observed is taken as an index of copying. This index will be referred to as the P index.

An index similar to the P index can be defined by using, as the item response probabilities, item response "difficulties" for persons with test scores similar to the suspected copier. This requires partitioning the test score range into intervals and computing item response "difficulties" for all score intervals. The compound binomial probabilities for a particular person suspected of copying are then computed based on the item response "difficulties" in the score interval that the suspected copier's test score falls in. As with the P index, the probability of obtaining a value for the number of identical incorrect responses greater than or equal to that observed is taken as an index of copying. This index will be referred to as the CP index (for conditional P index).

Definitions of the primary statistics and how they are used in each method is summarized in Table 1.

Insert Table 1 about here

Method

The basis for comparing these seven statistical methods will be decision rules for deciding if a pair of examinees' responses indicate the presence of copying. For each statistical method the decision rules will be in the form of cutoffs on the statistic(s), derived from the primary statistics, that are the basis of the method (see Table 1). Data from a 100 item, four alternative, multiple choice lisensure test will be used to establish distributions of the statistic(s) used in each method for pairs of examinees who could not have copied from one another. These data will be referred to as the benchmark data. The benchmark data will be used to establish decision rules for each method corresponding to approximate false positive rates (the proportion of pairs of examinees identified as copiers) of: .0005, .001,

.0025, .005, .0075, .01, .05, .10, .25. For the decision rule corresponding to each false positive rate the proportion of 500 pairs of examinees (not included in the benchmark data) for which simulated copying takes place which are above the cutoff(s) of the decision rule (true positive rate) will be calculated.

Data

The data used in the study are from a single administration of a 100 item lisensure test. All examinees having no omitted items and raw scores greater than 0 (approximately 96% of the all examinees) were included, with the resulting data set having 19167 examinee records. The order of the examinees was permuted using a random permutation of integers. The first examinee in the permuted data set was paired with the next examinee who was tested in a different state. The observation following the second observation of the first pair was paired with the next observation that was tested in a different state from that observation. This process was continued until observations in the permuted data set were exhausted. This resulted in 9143 pairs of observations (18286 individuals). The first 8643 pairs of examinees were taken to be the benchmark data set. The last 500 pairs were used to generate copying pairs.

Calculation of Statistics

Each of the Angoff statistics consists of a primary statistic indicating the similarity of responses of two examinees, and a controlling variable that is a function of the test scores of the two examinees. In computing both B and H, the controlling statistic is partitioned into intervals and the mean and standard deviation of the primary statistic are computed for each interval. The value of B or H for a particular pair of examinees is the value of the primary statistic observed for that pair minus the mean of the primary statistic in the interval of the controlling statistic in which the pair's value on the controlling variable falls, divided by the standard deviation of the primary statistic in that interval. Table 2 gives the 20 intervals used for the controlling statistic for index B (the product of the number of incorrect responses for each examinee, I1I2). For each interval the number of pairs in the benchmark data that fall in that interval is given, along with the mean and standard deviation of the primary statistic for Index B (the number of jointly incorrect responses, JI1I2) in that interval. Table 3 gives the 16 intervals used for the controlling statistic of Index H (number of incorrect responses for the examinee of the pair with the highest score, MINI). For each interval, the number of pairs in the benchmark data, and the mean and standard deviation of the primary statistic (the largest number of incorrect responses in any string of identical responses, STRINGI2) is given.

Insert Tables 2 and 3 about here

For indices g_2 and P the proportions of persons responding to each alternative of each item (response "difficulties") are needed. These were computed using all 17286 examinees in the benchmark data. The response "difficulties" were used in Frary, Tideman, and Watts' (1977) formulas (equations 1 and 2 above) to produce the linear functions of total test score used to calculate the item response probabilities for the suspected copier needed in calculating g_2.

The response "difficulties" computed from the benchmark data are used to compute the P index in the following way. Let p_{ij} be the response "difficulty" for response j of item i. Suppose that the source has answered w items (items k_1, k_2, . . ., k_w) incorrectly and the suspected copier has identical responses to v of these w items. Then, the value of the P index for this pair is given by the probability that the random variable Y is greater to or equal to v, where Y has a compound binomial distribution with parameters

$$(w, p_{k_1 u_{k_1 a}}, p_{k_2 u_{k_2 a}}, \ldots, p_{k_w u_{k_w a}}) ,$$

and source a gives response u_{ia} to item i.

The calculation of the CP index is identical to that of the P index except that response "difficulties" are calculated for various intervals of total test score. The p_{ij} in the score interval of the suspected copier are used to calculate CP. The 11 test score intervals used in calculating response "difficulties", and the number of pairs in the benchmark data for each interval are given in Table 4.

Insert Table 4 about here

Determination of False Positive Rates

The indices B, H, g_2, P, CP, and the value of the four statistics used for PAIR1 and PAIR2 were computed for the 8643 pairs of examinees in the benchmark data. For the B, H, g_2, P and CP indices, cutoffs were determined from the distributions in the benchmark data such that proportions corresponding to approximately .0005, .001, .0025, .005, .0075, .01, .05, .10, and .25 of the benchmark data fell at or above the cutoffs. These cutoff points as well as the extreme values of each of the indices in the benchmark data are given in Table 5.

Insert Table 5 about here

For methods in which a single statistic is used, it is likely that the optimal decision rule for a particular false positive rate, in the sense of maximizing the true positive rate for detecting a particular type of copying, will be of the form of a cutoff on that statistic. For such a rule, pairs above or equal to the cutoff involve suspected copying and those below the cutoff do not. This type of decision rule is considered for the five methods based on one statistic. For the methods based on two statistics it is not clear that any simple set of decision rules will always include the optimal decision rule in any particular situation. Here, for simplicity, we consider decision rules for the PAIR1 and PAIR2 methods that take the form of cutoffs on each of the statistics. A pair of examinees must be equal to or above both cutoffs for copying to be suspected. This may result in a conservative assessment of the PAIR1 and PAIR2 methods relative to the methods based on one statistic.

For the PAIR1 and PAIR2 methods, cutoffs on both variables were chosen to produce approximately the false positive rates given above. There may be several pairs of cutoffs that approximately give a particular false positive rate. In cases where this occurred, the cutoffs for which the marginal proportions below each individual cutoff were most equal were chosen. Because the statistics used in the PAIR1 and PAIR2 methods have fewer distinct values than the statistics used in the other methods, the actual false positive rates attained by the cutoffs were in most cases not as near the target values as the cutoffs given in Table 5 (which were very near the target values). Table 6 gives the cutoffs for the PAIR1 and PAIR2 methods and the achieved false positive rates for each target false positive rate. It should be noted that in some cases these cutoffs are such that it is possible that the true positive rates would at some point decrease with an increase in false positive rates.

Insert Table 6 about here

In some cases one might want to assume copying occurred only if the value of the index used was well outside the extreme value of the index in the benchmark data. With this in mind, values were chosen beyond the extreme of each index for computing true positive rates. For the B, H, and g_2 indices these values were taken to be one interquartile range beyond the extreme value observed. Since P and CP are supposedly directly interpretable as probabilities, the "beyond" values of .00001 and .0001 were used for P and CP, respectively. These cutoff points beyond the extremes are given in Table 5 in the row labeled "Beyond".

For the PAIR1 and PAIR2 methods the cutoffs beyond the extremes of the data were determined based on visual inspection of a plot of the pair of variables for each method using the benchmark data. One cutoff was chosen such that observations beyond that point would, it was thought, appear clearly suspicious to most persons. Another cutoff was chosen beyond the first. Points beyond this second cutoff would, it was thought, be considered clear outliers by most persons. These two cutoffs for both methods based on pairs of statistics are given in Table 6, in rows labeled "Extreme" and "Beyond", respectively.

Types of Copying

Copying was simulated by changing responses of the second examinee of each of the 500 pairs of examinees in the "copying" data set to the responses of the first examinee of the pair. The second examinee of a pair had 10, 20, 30, 40, or 50 of his/her item responses overwritten with the responses of the first examinee of the pair (this may or may not change the response of the second examinee), corresponding to copying 10, 20, 30, 40, or 50 percent of the 100 item test. Each of these percentages of items copied was crossed with 5 methods of selecting which items to copy. Thus, there were 25 copying conditions to be examined.

The first method of selecting items to be copied (the random copying condition) represented a type of copying in which an examinee looks at the source's answer sheet at random intervals and copies the sources answer to the question he/she is working on. This was implemented by randomly selecting n unique integers from 1 to 100 and using these integers as the items to be copied (where n items are to copied).

The second method of selecting items (the difficulty copying condition) also involved random selection, but the items were weighted by their difficulty (i.e., proportion of examinees in the benchmark data who answered the item correctly). This represented a type of copying in which an examinee copies randomly, and is more likely to copy more difficult items. This was implemented by first subtracting the difficulty of each item from 1. This value for each item, divided by the sum of these values over all items, was taken as the probability of the item being chosen. Using these probabilities, items were "randomly" selected until n distinct items were chosen (where n items were to be copied).

The third, fourth, and fifth methods of selecting items to be copied were all based on the premise that examines tend to copy items in consecutive groups or strings. In the third method (the string end copying condition) the second examinee's responses to the last n items on the test were chosen to be changed to the responses of the first examinee (where n items were to be copied). This condition may represent the situation in which an examinee copies items at the end of the test due to running out of time or increased item difficulty. Note that the items in the test used for the data in this study are not ordered on the test in terms of difficulty.

The fourth method (the string beginning copying condition) had the responses of the second examinee to the first n items on the test changed to the responses of the first examinee (where n items were to be copied). This condition was included to determine if the specific placement of the string of items copied affected the ability to detect copying.

The fifth method of selecting items to be copied (the string 5 copying condition) represented a situation where several strings of items were copied. The 100 items were subdivided into 20 consecutive sets of 5 items each (i.e., the first set contained items 1 through 5, the second set contained items 6 through 10, etc.). For the 10, 20, 30 40 and 50 percent copying conditions, 2, 4, 6, 8 and 10 sets of the five items were randomly chosen as the items to be copied. This means that simulated-copying pairs may have strings of copied items longer than 5 if two consecutive item strings were chosen.

Conditional False Positive Rates

All of the statistical methods considered here for investigating allegations of copying are based on derived statistics using one or more of 5 primary statistics that are indicators of the similarity of responses of two examinees. The main reason for using the derived statistics instead of the primary statistics is that a decision rule based on the derived statistics should result in more approximately the same false positive rates for all levels of test score values for pairs of examinees. To check the extent to which this condition holds for the methods examined, three sets of conditional benchmark data were produced. For each set of conditional benchmark data the false positive rates corresponding to the cutoffs obtained for the unconditional benchmark data (given in Tables 5 and 6) were computed.

A conditional false positive rate significantly above the unconditional false positive rate for a particular method would make questionable the practice of using a single unconditional decision rule for all pairs of examinees for that method. In such a case, an examinee accused of copying based on the unconditional benchmark data might be able to rightfully argue that evidence would have suggested he/she did not copy if benchmark data similar to the suspected pair in terms of test scores had been used.

Two of the three sets of conditional benchmark data were for subgroups in which differences in the conditional and unconditional false positive rates were thought most likely to occur: examinees with high and low total test scores, respectively. The high group consisted of examinees in the original benchmark data with test scores above or equal to 86. There were 45 such examinees out of the 17286 examinees in the benchmark data. All possible pairs of these 45 examinees were formed, and the statistics used in each method of examining allegations of copying were computed for every pair who were not both tested in the same state. This produced a conditional high benchmark data set of 835 pairs of examinees.

To produce the low benchmark data, examinees were chosen at two score points (61 and 62) around the 37th and 41st percentiles of the test score distribution for all examinees in the unconditional benchmark data. This level of test score was chosen for the low condition due to it being, in our experience, the lowest score level observed for examinees suspected of copying, on this particular licensure exam. All examinees with raw scores of 61 and 62 in the unconditional benchmark data were selected. Pairs of examinees, one from each score level, were chosen until all examinees from the score level with fewer examinees (61) were used. Pairs who took the test in different states were kept, resulting in a conditional low benchmark data set of 628 pairs of examinees. For purposes of computing the statistics for each of the methods of examining allegations of copying, the examinee in each pair with a score of 61 was chosen to be the examinee suspected of copying.

An additional set of conditional benchmark data was produced in which the examinee suspected of copying had a relatively low score (60, around the 33rd percentile), and the other examinee had a relative high score (75, around the 90th percentile). All examinees with test scores of 60 or 75 were chosen from the unconditional benchmark data. Pairs of examinees, one from each score level, were chosen until all examinees at the score level with the greater number of examinees (60) were used. This resulted in some examinees at the score level of 75 being used in two of the pairs. All pairs of examinees who took the test in a different state were used, resulting in a mixed conditional benchmark data set of 583 pairs.

Results

The true positive rates for cutoffs corresponding to the fixed false positive rates are presented in Tables 7 through 11 for the random, difficulty, string end, string beginning, and string 5 copying conditions, respectively. Figures 1 through 5 graphically present a small subset of the information in Tables 7 through 11 that illustrate some of the major results. These figures plot the true positive rates corresponding to a false positive rate of .001 as a function of the method used, for each type of copying.

Insert Figures 1 through 5 about here

The effects of the method used and type of copying on the true positive rates interact. As would be expected, the methods based on strings of identical responses (H, PAIR1, PAIR2) perform better, in terms of true positive rate, than the other methods in the string end and string beginning copying conditions. On the other hand, the methods based on strings perform less well than the other methods in the random and difficulty copying conditions. In general, there are not great differences between the methods for the string 5 condition.

Considering the methods not based on strings (B, g_2, P, CP), P, in general, performs worse than the other methods. CP tends to perform best overall, but there is usually not a large difference in true positive rates between CP and B and g_2.

For the methods based on strings (H, PAIR1, PAIR2), the relative performance of the methods interacts with the type of copying. For the string end and string beginning conditions, H significantly out performs all other methods (including PAIR1 and PAIR2), especially for lower numbers of items copied (10 through 30). In the random and difficulty copying conditions H performs less well than all the other methods (including PAIR1 and PAIR2). When 40 or more items are copied PAIR1 performs better than PAIR2 in these two copying conditions. When 30 or fewer items are copied neither PAIR1 nor PAIR2 consistently performs better.

The number of items copied has a large effect on true positive rate. Copying only 10 items is generally not very detectable for false positive rates below .01. An exception to this is when copying of all the 10 items occurs in a string and one of the methods based on strings (especially H) is used. For example, in the string end condition with a false positive rate of .005, the H method detects 48.6% of the copiers and the PAIR1 method detects 31.8% of the copiers, but the most copiers detected by any method not using strings is 6.2%. When 20 items are copied all the types of copying are detected at least moderately well by at least one method. For example, for all types of copying, at least 14% of copiers are detected with a false positive rate of .001, and at least 33% of copiers are detected with a false positive rate of .005. In general, approximately half or more of the copiers are detected by at least one method when 30 items are copied for all the types of copying (e.g., at least 45% of copiers are detected with a false positive rate of .0005). When 40 or 50 items are copied, for all types of copying a large majority of copiers are detected by at least one method (e.g. at least 80% and 95% of copiers who copied 40 and 50 items, respectively, are detected with a false positive rate of .0005).

Conditional False Positive Rates

The conditional false positive rates for the three sets of conditional data (high, low, and mixed) are presented in Table 12. It should be remembered when examining these conditional false positive rates that they are based on less than 10% of the data that the unconditional false positive rates were based on. Therefore, in the following discussion, conditional false positive rates corresponding to unconditional false positive rates below .005 will not be interpreted.

In the high conditional data (examinees with scores above or equal to 86) the conditional false positive rates for g_2 are consistently higher than the corresponding unconditional false positive rates. Recall that g_2 is the only method based on the total number of identical responses for the pair, which is positively related to the test scores of the pair of examinees.

In the low conditional data (examinees with scores of 61 and 62 for the copier and source, respectively) all the methods which are based in some way on the number of identical incorrect responses (which is negatively related to the test scores of a pair of examines) have nontrivially higher conditional than unconditional false positive rates for at least some of the decision rules. CP, PAIR1 and PAIR2 seem to be most affected, although not severely so.

In the mixed conditional data (examinees with scores of 75 and 60 for the source and copier, respectively) only H shows consistent and/or much greater conditional than unconditional false positive rates. This could be partly due to the discreteness of H in the conditional data. H takes on only a small number of values, and so only a small number of false positive rates can be attained.

Insert Tables 7 through 12 about here

Discussion

A major finding is that different methods of investigating allegations of copying work differentially well in detecting different types of copying. In applying these results one could deal with this issue in two ways. First, one could use multiple methods so that no matter what type of copying was occurring at least one of the methods considered would probably work well. In this case, decision rules would be set for each method, and copying would be suspected if any of the methods indicated unusual response similarity. To keep the same false positive rate when considering two or more methods, the decision rules of each individual method would have to be such that they would have much lower false positive rates than the decision rules that would be used if the method was used alone. This might result in higher true positive rates than using a single method in cases in which the single method does not perform well for the type of copying that is occurring, and one of the additional methods to be considered does. On the other hand, using the method that performs well by itself may result in higher true positive rates than using it in combination with another method.

The second way to deal with the interaction of methods of investigating allegations of copying and types of copying is to determine the type of copying that is most likely to occur and use a method that is good at detecting that type of copying. There are a couple sources of evidence which indicate that the string 5 copying condition is the most realistic of the types of simulated copying considered here. First, some observational evidence we have seen indicates that at least some copying is done in strings. Second, Angoff (1971) computed both the B and H statistics for 50 examinees who were "known and admitted copiers" (p. 46). Angoff found that the B and H statistics worked almost equally well at identifying the copiers. The B statistic indicated evidence for copying existed for 47 of the 50 examinees on at least one of the examinations that were taken. The H statistic indicated evidence for copying existed for 49 of the 50 examinees on at least one of the examinations. These results coupled with the extremely poor performance of H in the random and difficulty conditions, strongly suggests that copying in Angoff's group of "known and admitted copiers" was occurring in a manner different from that simulated in the random and difficulty conditions. In particular, it suggests the examinees Angoff studied were copying at least some strings of responses.

Angoff used a cutoff of 3 as a decision rule for B and H in determining the performance of each for his data. In the present benchmark data these cutoffs would imply a false positive rate of slightly less than .0025 for B and between .01 and .0075 for H (see Table 5). The string 5 condition produces the most equal true positive rates for B and H corresponding the false positive rates of .0025 and .01, respectively, over all amounts of copying. Therefore, the results in the string 5 condition are most consistent with Angoff's result of near equal performance of B and H. This evidence, along with the assumption that it is unlikely all items copied will be in a single string (as in the string end and string beginning conditions), suggests that the string 5 copying condition is perhaps the most realistic of the conditions studied.

Given the evidence that the string 5 condition may most closely correspond to the type of copying that occurs in many cases, the results found here suggest that it may not make a great deal of difference which of the statistical methods of investigating copying considered here are used. This is due to the minor differences in the true positive rates of the methods in the string 5 condition.

Other Factors in Choosing a Method

There are other factors besides performance in terms of true positive rate that may be important in deciding on a statistical method of investigating allegations of copying to use in practice. Two of these factors are interpretability and consistency of conditional and unconditional false positive rates.

Interpretability may be important to consider when the statistical results of an investigation of an allegation of copying are to be reported to an audience not familiar with statistical arguments. The more easily the statistics used by a method can be related to one or more of the primary statistics of response similarity, the more interpretable the results of the method will be. By this criterion of interpretability, the methods based on direct adjustment (B and H) and indirect adjustment (PAIR1 and PAIR2) are easier to interpret than the methods based on probability distributions of primary statistics. Of the adjustment methods, PAIR1 and PAIR2 are most easily interpretable since the pairs of variables used in each method are direct measures of the response similarity of a pair of examinees.

Directly related to interpretability of a method is the number of assumptions made. The methods based on the probability distribution of a primary statistic make assumptions about examinees' responses to the test. These types of assumptions may need to be made in investigating theoretical psychometric properties of tests, but are not needed to investigate the straight-forward issue of whether an examinee copied answers from another examinee. Using such assumptions in investigating instances of copying is only be justified when it can be demonstrated that the use of the assumptions produces better results (e.g., higher true positive rates) than methods not using the assumptions. The finding here that under some conditions the CP method produced higher true positive rates than the other methods could be used to support the use of the assumptions made in the CP method in investigating allegations of copying. On the other hand, the P method was generally inferior to the B method (as well as to the CP and g_2 methods), suggesting that the assumptions used in computing the P index are not well supported for the purposes of investigating allegations of copying.

All the methods consider the statistics they use to have approximately equal interpretability whatever the test scores of the pair of examinees investigated. This is to avoid the situation in which it is found that using only benchmark data with scores similar to a pair of examinees in question the pair does not have unusually similar responses when they were found to have unusually similar responses with benchmark data using all examinees. This issue was investigated here by comparing conditional false positive rates for three sets of conditional data to unconditional false positive rates corresponding to specific decision rules. The results reported here indicated that conditional false positive rates were greater than unconditional false positive rates for certain methods of simulated copying, for types of conditional data in which the two examinees had both high or both low scores. If a method was based on a primary statistic that was positively related to the test scores of the two examinees, then the method tended to have higher conditional than unconditional false positive rates for conditional data in which both examinees had high scores. If a method was based on a primary statistic that was negatively related to the test scores of the two examinees, then the method tended to have higher conditional than unconditional false positive rates for conditional data in which both examinees had low scores.

Based on these results, it is recommended that in cases in which the pair examinees have either both low or both high scores a slightly conservative decision rule may be appropriate when the method to be used is one that was shown to be liberal in that situation. Based on the results found here it is not felt that this effect is large enough, even for those cases in which it is most severe, to necessitate using conditional benchmark data consisting of examinees with scores near the pair in question.

The Use of Benchmark Data

The statistics used in the B, H, g_2, P and CP methods are given in such a way that they could potentially be considered meaningful without reference to benchmark distributions of the statistics for non-copying pairs of examinees. P and CP should, by definition, be directly interpretable as probabilities. Frary, Tideman and Watts (1977) state that g_2 should be approximately normally distributed. Angoff (1974) uses the normal distribution as a reference in setting cutoffs on B and H, but acknowledges that the distributions of B and H may be in some cases distinctly non-normal. Using these assumptions theoretical cutoffs can be set for each of these methods without using benchmark data. For example, theoretical cutoffs, based on the above assumptions, which would produce a false positive rate of .001 would be: .001 for P and CP, and 3.09 (the 99.9 percentile point of a normal distribution) for B, H, and g_2.

Table 13 compares the theoretical cutoffs for methods B, H, g_2, C and CP based on the assumptions given above with the actual values based on the benchmark data from Table 5. In many instances there are large discrepancies. It is possible that large discrepancies would exist even if the theoretical cutoffs were correct because of sampling error, since the cutoffs are order statistics in the extremes of the sample. Still, the magnitude of the differences in the observed and theoretical values is in some cases large and consistent enough (over cutoffs) that the hypothesis that the differences are due to sampling variability is unlikely.

In cases in which the direction of the difference is in a liberal direction (i.e., the theoretical cutoffs result in higher false positive rates than the observed cutoffs, as for B, H, g_2, and P) using the benchmark data is definitely preferred over using the theoretical results, especially for smaller false positive rates where the differences in the benchmark and theoretical results appear to be larger. In cases in which the direction of the difference is clearly in a conservative direction (i.e., the theoretical cutoffs result in lower false positive rates that the observed cutoffs, as for CP) one might feel justified in using the theoretical values, although it would probably be best even in these cases to use benchmark data.

Insert Table 13 about here

Possible Limitations

Some issues involved with limitations of a study of this type should be kept in mind when interpreting and/or applying the results. One of these limitations is that it is very likely that none of the types of simulated copying studied here completely describes the copying that occurs by any real examinee. The variety of types of copying studied here were included to attempt to overcome this limitation to as great an extent as possible. It is possible, though, that none of the results reported here is very accurate for some types of copying that actually occur.

Only one test and one group of examinees were studied. It is possible that different results would be obtained using different tests and different examinees. The most likely source of such differences are changes in the tails of the empirical distributions studied. Even if the present test and group of examinees could be considered as a "random sample" of tests and examinees, there is much variability in the tails of an empirical distribution, and a different "sample" might produce large differences in the cutoffs obtained in the benchmark data, and perhaps differences in the results. Many of the results of this study, though, are logically consistent and would most likely be reproduced with different examinees and/or tests, although additional data would need to be studied to verify this.

Cautions

No statistical method of investigating allegations of copying can provide conclusive proof that copying occurred. The perspective taken here is that the statistical methods discussed are best employed in cases in which there is collateral information suggesting that copying may have occurred.

Alternative sources of response similarity should be taken into account in investigating an allegation of copying. Unusual response similarities may occur for reasons other than copying. Several researchers have noted that similarities in incorrect item responses may be partly the result of background characteristics, such as similar instruction, and experiences (Buss and Novick, 1980; Powell, 1968; Dickenson, 1945; Harnisch, 1983).

Along with evidence that copying occurred (e.g., statistical evidence of similarity of responses) evidence that indicates copying did not occur should also be considered (Buss and Novick, 1980). For example, statistics that provide evidence against the allegation of copying should be examined (e.g., number of questions the suspected copier answered correctly and the source answered incorrectly).

Conclusions

The performance of the methods of investigating allegations of copying studied here interact with the type of simulated copying that occurred. It appears that the string 5 copying condition approximates actual copying better than the other types of copying studied. For this type of copying there was not a great deal of difference in the performance of the methods. Thus, the choice among the methods studied may not be very important in terms of detecting copiers. One exception is that the P method seemed distinctly inferior to other methods, in particular to the CP method, and it is recommended that this method not be used. The decision, then, about which method to use should be based on other considerations, including the interpretability of the method.

The methods studied here were, in general, not able to detect copying when only 10 out of 100 items are copied. For the string 5 copying condition and a false positive rate of .001 approximately 20, 50, 85, and 95 percent of the copiers could be detected when 20, 30, 40, and 50 of the 100 items were copied. Thus, with approximately one non-copier identified as copying out of each 1000 investigated, a majority of actual copiers were detected if they copy 30 or more items on the test.

Conditional false positive rates may be slightly greater than unconditional false positive rates for cases in which the conditioning is done on a subgroup of examinees in which the two examinees have both high or both low scores. Based on this result, it is suggested that for cases in which the pair of examinees have both very high or both very low scores slightly conservative decision rules might be appropriate (depending on the method used).

The results of the study suggested that benchmark data on non-copying pairs of examinees, rather than theoretical assumptions, should be used in determining the false positive rates of the methods The extent to which a particular set of benchmark data can be used for different test forms and/or different groups of examinees is an empirical question.

References

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

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

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.

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

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

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. Journal of Educational Measurement, 20, 191-206.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. Journal of Educational Statistics, 4, 269-290.

Powell, J. C. (1968). The interpretation of wrong answers from a multiple choice test. Educational and Psychological Measurement, 28, 403-412.

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

TABLE 1

Primary Statistics and their use in the Statistical Methods of Investigating Copying

Primary Statistics	
Name	Definition
TJOINT	Number of items the two examinees answered identically.
JI1I2	Number of items for which the suspected copier and source picked the same incorrect alternative.
STRINGL	Number of items in the longest string (sequence of consecutive items) of identical responses.
STRINGI1	Number of incorrect response in the longest string of identical responses.
STRINGI2	Maximum number of incorrect items in any string of identical responses.

Table 1 (continued)

Primary Statistics and their use in the Statistical Methods of Investigating Copying

Statistical Methods	Definition
B	JI1I2 conditioned on the product of the number of incorrect responses of the two examinees (I1I2).
H	STRINGI2 conditioned on the number of incorrect responses for the examinee with the higher test score (MINI).
PAIR1	JI1I2 and STRINGL.
PAIR2	STRINGI1 and PJ = (100 JI1I2)/(NITEMS - (TJOINT - JI1I2)), where NITEM is the number of test items.
g_2	Probability of suspected copier having TJOINT or greater reponses identical to source.
P	Probability of suspected copier having JI1I2 or greater reponses identical to the incorrect responses of the source.
CP	Probability of suspected copier having JI1I2 or greater responses identical to the incorrect responses of the source (using item response "difficulties" of examinees with scores near the suspected copier).

TABLE 2

Intervals of I1I2 for Computing the B Index with Mean and Standard Deviation of JI1I2 in Each Interval

Interval of I1I2	Number in Benchmark Data	JI1I2 Mean	JI1I2 Standard Deviation
1 - 499	150	4.39	2.06
500 - 599	209	4.89	1.87
600 - 699	344	5.47	1.95
700 - 799	521	6.07	1.98
800 - 899	631	6.52	2.00
900 - 999	744	7.00	2.11
1000 - 1099	774	7.46	2.21
1100 - 1199	781	7.75	2.34
1200 - 1299	771	8.32	2.45
1300 - 1399	619	8.94	2.43
1400 - 1499	647	9.10	2.57
1500 - 1599	534	9.45	2.56
1600 - 1699	392	9.65	2.50
1700 - 1799	342	10.62	2.62
1800 - 1899	308	10.75	2.69
1900 - 1999	218	10.88	2.78
2000 - 2099	159	11.90	2.82
2100 - 2299	236	12.03	2.62
2300 - 2499	126	12.87	2.87
2500 - 4550	137	13.88	3.38

TABLE 3

Intervals of MINI for Computing the H Index with Mean and Standard Deviation of STRINGI2 in Each Interval

Interval of MINI	Number in Benchmark Data	STRINGI2	
		Mean	Standard Deviation
9 - 16	153	1.33	0.548
17 - 18	152	1.39	0.587
19 - 20	299	1.48	0.626
21 - 22	451	1.53	0.622
23 - 24	627	1.60	0.665
25 - 26	780	1.69	0.695
27 - 28	960	1.72	0.700
29 - 30	971	1.77	0.703
31 - 32	947	1.84	0.709
33 - 34	822	1.89	0.750
35 - 36	714	1.95	0.705
37 - 38	582	2.01	0.744
39 - 40	410	2.09	0.767
41 - 42	314	2.13	0.717
43 - 46	309	2.25	0.748
47 - 65	152	2.30	0.796

TABLE 4

Categories of Test Score Used in Computing Response "Difficulties" for Index CP

Interval of Test Score	Number in Benchmark Data
0 - 40	139
41 - 45	292
46 - 50	774
51 - 55	1729
56 - 60	2821
61 - 65	3496
66 - 70	3551
71 - 75	2703
76 - 80	1331
81 - 85	405
86 - 100	45

TABLE 5

Cutoffs in Benchmark Data For Methods Based on One Statistic

False Positive	Methods				
	B	H	g_2	P	CP
Beyond	5.223	6.459	5.363	.000010	.00010
Extreme	3.853	5.118	4.148	.000068	.00029
.0005	3.516	4.680	3.534	.000148	.00133
.0010	3.341	4.453	3.448	.000650	.00229
.0025	2.989	3.795	2.894	.001993	.00669
.0050	2.722	3.252	2.644	.003584	.01246
.0075	2.504	3.171	2.486	.005316	.02039
.0100	2.370	2.904	2.380	.007366	.02360
.0500	1.669	1.823	1.751	.051728	.09283
.1000	1.267	1.478	1.374	.112168	.15755
.2500	.696	.444	.810	.291021	.33528

TABLE 6

Cutoffs in Benchmark Data For Methods Based on Two Statistics

	PAIR1			PAIR2		
Target False Positive	Achieved False Positive	JI1I2	STRINGL	Achieved False Positive	PJ	STRINGI1
Beyond	Beyond	25	17	Beyond	40	8
Extreme	Extreme	20	12	Extreme	35	5
.0005	.00035	16	11	.00046	30	4
.0010	.00104	14	12	.00081	29	4
.0025	.00289	13	12	.00266	26	4
.0050	.00509	12	11	.00498	23	4
.0075	.00671	14	9	.00717	22	4
.0100	.00972	12	10	.01111	25	3
.0500	.05091	11	8	.04593	19	3
.1000	.10656	9	8	.09511	20	2
.2500	.22886	8	7	.25060	14	2

TABLE 7

True Positive Rates for Random Copying Condition

10 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.000	.002	.000	.002	.000	.000	.000
Extreme	.002	.004	.000	.006	.000	.000	.000
.0005	.010	.006	.004	.008	.008	.010	.010
.0010	.018	.008	.006	.020	.016	.004	.012
.0025	.030	.012	.028	.032	.042	.008	.028
.0050	.064	.014	.064	.046	.066	.030	.036
.0075	.080	.016	.078	.058	.082	.032	.042
.0100	.096	.032	.098	.076	.092	.042	.070
.0500	.246	.110	.260	.220	.274	.176	.132
.1000	.404	.216	.424	.346	.394	.290	.294
.2500	.620	.388	.654	.588	.650	.488	.442

20 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.002	.004	.000	.022	.024	.000	.000
Extreme	.052	.012	.026	.044	.054	.004	.008
.0005	.104	.012	.082	.066	.110	.042	.052
.0010	.128	.020	.098	.122	.140	.056	.066
.0025	.216	.036	.246	.170	.262	.068	.086
.0050	.294	.048	.332	.220	.328	.128	.096
.0075	.372	.054	.406	.258	.400	.180	.096
.0100	.416	.080	.452	.296	.430	.168	.222
.0500	.648	.204	.692	.574	.690	.422	.286
.1000	.780	.382	.810	.698	.788	.492	.550
.2500	.906	.580	.936	.860	.912	.688	.592

30 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.032	.012	.030	.134	.190	.000	.000
Extreme	.298	.034	.214	.212	.272	.030	.062
.0005	.426	.048	.410	.256	.456	.160	.204
.0010	.472	.066	.440	.398	.540	.158	.206
.0025	.604	.104	.682	.522	.682	.178	.226
.0050	.714	.144	.760	.596	.742	.284	.230
.0075	.776	.158	.790	.634	.804	.424	.232
.0100	.806	.208	.814	.672	.816	.378	.420
.0500	.920	.430	.934	.860	.938	.674	.428
.1000	.956	.616	.970	.920	.958	.710	.724
.2500	.986	.796	.996	.978	.988	.866	.738

Table 7 (continued)

True Positive Rates for Random Copying Condition

			40 Items Copied				
False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.234	.040	.214	.390	.568	.006	.018
Extreme	.708	.098	.596	.532	.656	.160	.170
.0005	.774	.118	.786	.600	.794	.376	.350
.0010	.822	.144	.810	.722	.838	.350	.354
.0025	.882	.250	.894	.782	.904	.382	.358
.0050	.916	.310	.924	.822	.924	.510	.360
.0075	.926	.336	.944	.848	.946	.648	.360
.0100	.934	.406	.956	.862	.946	.624	.580
.0500	.978	.622	.990	.958	.986	.856	.580
.1000	.992	.770	.998	.982	.988	.874	.816
.2500	.998	.896	.998	.992	1.000	.956	.818

			50 Items Copied				
False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.618	.090	.548	.704	.880	.036	.040
Extreme	.936	.214	.882	.794	.920	.412	.324
.0005	.948	.264	.940	.830	.956	.634	.518
.0010	.958	.320	.946	.898	.966	.626	.518
.0025	.966	.436	.984	.936	.972	.646	.518
.0050	.976	.506	.990	.956	.986	.732	.520
.0075	.978	.544	.992	.964	.988	.856	.520
.0100	.980	.612	.994	.972	.992	.814	.710
.0500	.996	.806	.994	.988	.996	.958	.712
.1000	.998	.904	.998	.992	.998	.976	.900
.2500	1.000	.976	1.000	.998	.998	.992	.900

TABLE 8

True Positive Rates for Difficulty Copying Condition

10 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.000	.000	.000	.000	.000	.000	.000
Extreme	.004	.006	.000	.004	.002	.002	.006
.0005	.016	.008	.016	.010	.008	.008	.018
.0010	.024	.010	.018	.018	.018	.006	.018
.0025	.040	.018	.044	.040	.048	.008	.026
.0050	.076	.028	.062	.052	.082	.038	.032
.0075	.106	.028	.090	.070	.118	.032	.032
.0100	.134	.034	.118	.086	.136	.046	.064
.0500	.338	.120	.304	.290	.352	.210	.136
.1000	.466	.222	.458	.438	.506	.316	.354
.2500	.700	.416	.718	.650	.732	.514	.468

20 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.002	.002	.004	.018	.034	.000	.000
Extreme	.094	.010	.038	.070	.066	.006	.022
.0005	.170	.014	.140	.112	.202	.062	.078
.0010	.230	.024	.156	.182	.274	.058	.084
.0025	.340	.050	.380	.284	.424	.076	.096
.0050	.442	.064	.466	.330	.508	.150	.112
.0075	.522	.078	.544	.380	.590	.224	.112
.0100	.568	.110	.588	.420	.608	.218	.282
.0500	.804	.250	.788	.704	.838	.496	.324
.1000	.868	.448	.896	.844	.912	.572	.606
.2500	.964	.678	.970	.948	.966	.748	.632

30 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.108	.014	.068	.272	.382	.000	.006
Extreme	.504	.048	.358	.392	.506	.078	.076
.0005	.614	.058	.586	.462	.676	.246	.278
.0010	.678	.074	.624	.594	.728	.232	.282
.0025	.782	.134	.816	.684	.824	.252	.288
.0050	.866	.174	.866	.750	.876	.380	.290
.0075	.894	.194	.904	.780	.916	.528	.290
.0100	.914	.260	.910	.810	.920	.492	.554
.0500	.976	.494	.976	.942	.986	.730	.564
.1000	.992	.698	.992	.980	.994	.764	.792
.2500	.998	.874	1.000	.996	.998	.880	.794

Table 8 (continued)

True Positive Rates for Difficulty Copying Condition

40 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.498	.058	.368	.622	.832	.032	.028
Extreme	.890	.170	.776	.760	.904	.280	.272
.0005	.930	.202	.908	.816	.950	.522	.464
.0010	.944	.244	.924	.888	.962	.490	.464
.0025	.968	.352	.970	.942	.982	.506	.464
.0050	.980	.434	.984	.954	.988	.614	.464
.0075	.980	.474	.990	.964	.990	.778	.464
.0100	.984	.536	.992	.970	.992	.730	.714
.0500	1.000	.748	1.000	.994	1.000	.912	.714
.1000	1.000	.878	1.000	.998	1.000	.918	.890
.2500	1.000	.970	1.000	1.000	1.000	.978	.890

50 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.852	.200	.760	.914	.980	.088	.084
Extreme	.974	.330	.954	.960	.986	.558	.460
.0005	.986	.380	.984	.966	.994	.760	.650
.0010	.986	.434	.984	.976	.994	.716	.650
.0025	.992	.592	.998	.992	.998	.724	.650
.0050	.994	.672	.998	.994	1.000	.822	.650
.0075	.994	.712	1.000	.994	1.000	.926	.650
.0100	.994	.766	1.000	.994	1.000	.892	.816
.0500	.998	.920	1.000	1.000	1.000	.980	.816
.1000	1.000	.976	1.000	1.000	1.000	.986	.940
.2500	1.000	.998	1.000	1.000	1.000	.996	.940

TABLE 9

True Positive Rates for String End Copying Condition

10 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.000	.092	.000	.002	.000	.000	.000
Extreme	.004	.212	.000	.006	.002	.000	.004
.0005	.010	.246	.006	.010	.016	.078	.032
.0010	.016	.300	.008	.018	.022	.082	.042
.0025	.032	.406	.028	.030	.046	.108	.130
.0050	.062	.486	.054	.046	.062	.318	.256
.0075	.080	.528	.074	.066	.104	.232	.306
.0100	.102	.570	.096	.078	.112	.408	.194
.0500	.306	.748	.252	.254	.314	.510	.536
.1000	.414	.844	.408	.386	.456	.748	.498
.2500	.646	.918	.666	.596	.678	.846	.848

20 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.002	.618	.000	.026	.024	.010	.010
Extreme	.066	.810	.036	.054	.052	.068	.068
.0005	.096	.842	.090	.076	.140	.332	.270
.0010	.138	.872	.108	.140	.182	.482	.316
.0025	.208	.908	.256	.224	.284	.584	.522
.0050	.316	.938	.328	.270	.364	.672	.686
.0075	.380	.942	.388	.306	.450	.482	.758
.0100	.438	.956	.438	.352	.480	.672	.592
.0500	.712	.984	.698	.568	.732	.758	.878
.1000	.816	.990	.820	.708	.832	.916	.872
.2500	.928	.994	.936	.886	.928	.946	.984

30 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.042	.968	.040	.180	.256	.038	.140
Extreme	.382	.984	.240	.290	.358	.266	.430
.0005	.508	.990	.462	.334	.518	.596	.754
.0010	.600	.990	.500	.460	.600	.734	.782
.0025	.714	.996	.714	.570	.748	.816	.886
.0050	.800	.996	.792	.638	.826	.888	.950
.0075	.832	.996	.828	.680	.868	.734	.964
.0100	.840	.998	.848	.714	.878	.888	.926
.0500	.950	.998	.958	.908	.960	.920	.990
.1000	.972	.998	.986	.952	.982	.966	.986
.2500	.996	1.000	.998	.990	.996	.986	1.000

Table 9 (continued)

True Positive Rates for String End Copying Condition

40 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.326	.996	.244	.482	.684	.166	.600
Extreme	.824	.998	.662	.612	.754	.482	.828
.0005	.856	.998	.826	.680	.876	.792	.934
.0010	.878	.998	.846	.774	.904	.888	.958
.0025	.916	.998	.934	.876	.946	.922	.990
.0050	.952	.998	.972	.908	.972	.954	.994
.0075	.956	.998	.982	.918	.980	.888	.996
.0100	.962	1.000	.984	.930	.988	.954	.992
.0500	.994	1.000	1.000	.984	1.000	.964	1.000
.1000	.996	1.000	1.000	.996	1.000	.990	1.000
.2500	1.000	1.000	1.000	1.000	1.000	.996	1.000

50 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.706	1.000	.592	.752	.920	.320	.886
Extreme	.946	1.000	.898	.856	.942	.680	.978
.0005	.964	1.000	.968	.890	.980	.884	.994
.0010	.964	1.000	.972	.948	.986	.952	.998
.0025	.980	1.000	.990	.966	.996	.964	1.000
.0050	.994	1.000	.996	.976	.996	.974	1.000
.0075	.994	1.000	.998	.980	1.000	.952	1.000
.0100	.994	1.000	.998	.986	1.000	.974	1.000
.0500	.996	1.000	1.000	.998	1.000	.980	1.000
.1000	.998	1.000	1.000	.998	1.000	.994	1.000
.2500	1.000	1.000	1.000	1.000	1.000	.996	1.000

TABLE 10

True Positive Rates for String Beginning Copying Condition

10 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.000	.022	.000	.000	.000	.000	.000
Extreme	.000	.076	.000	.000	.000	.004	.002
.0005	.004	.094	.002	.000	.000	.026	.020
.0010	.006	.134	.004	.012	.008	.054	.030
.0025	.020	.218	.024	.026	.028	.078	.072
.0050	.038	.278	.046	.034	.054	.146	.170
.0075	.068	.302	.066	.048	.082	.196	.210
.0100	.084	.358	.088	.062	.086	.368	.150
.0500	.244	.538	.256	.208	.270	.480	.422
.1000	.360	.664	.388	.338	.372	.704	.444
.2500	.582	.794	.632	.556	.584	.802	.764

20 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.000	.422	.000	.012	.004	.006	.006
Extreme	.036	.612	.018	.034	.022	.064	.030
.0005	.076	.650	.072	.050	.068	.248	.214
.0010	.106	.696	.082	.100	.106	.442	.270
.0025	.180	.760	.208	.150	.226	.542	.448
.0050	.258	.810	.314	.200	.296	.642	.622
.0075	.314	.828	.356	.234	.376	.442	.674
.0100	.348	.850	.398	.260	.388	.642	.524
.0500	.610	.926	.660	.546	.660	.724	.832
.1000	.746	.942	.794	.672	.756	.876	.810
.2500	.882	.978	.912	.830	.884	.918	.968

30 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.026	.824	.018	.118	.166	.036	.104
Extreme	.288	.910	.196	.214	.236	.224	.336
.0005	.374	.934	.370	.262	.394	.496	.586
.0010	.434	.942	.394	.382	.472	.660	.642
.0025	.508	.960	.610	.476	.590	.736	.784
.0050	.620	.974	.692	.538	.672	.800	.876
.0075	.682	.974	.756	.570	.728	.660	.896
.0100	.714	.974	.782	.600	.756	.800	.830
.0500	.866	.994	.914	.794	.882	.864	.964
.1000	.926	.996	.940	.860	.922	.936	.954
.2500	.984	.998	.976	.946	.974	.968	.992

Table 10 (continued)

True Positive Rates for String Beginning Copying Condition

40 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.194	.984	.168	.354	.492	.132	.450
Extreme	.612	.994	.556	.478	.592	.390	.692
.0005	.702	.994	.756	.538	.762	.692	.890
.0010	.764	.994	.774	.650	.810	.826	.906
.0025	.850	.996	.884	.746	.876	.866	.952
.0050	.904	.998	.922	.796	.912	.908	.980
.0075	.916	.998	.938	.818	.938	.826	.984
.0100	.932	1.000	.940	.836	.942	.908	.972
.0500	.980	1.000	.980	.948	.978	.952	.996
.1000	.992	1.000	.996	.978	.982	.984	.994

50 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.518	.996	.498	.650	.826	.260	.798
Extreme	.892	.998	.842	.758	.890	.586	.922
.0005	.926	1.000	.924	.802	.938	.824	.980
.0010	.952	1.000	.938	.876	.956	.916	.980
.0025	.972	1.000	.976	.924	.982	.942	.992
.0050	.980	1.000	.984	.942	.984	.968	1.000
.0075	.982	1.000	.988	.950	.986	.916	1.000
.0100	.984	1.000	.990	.966	.986	.968	.998
.0500	.998	1.000	.998	.988	.996	.984	1.000
.1000	.998	1.000	1.000	.994	1.000	.994	1.000
.2500	1.000	1.000	1.000	1.000	1.000	.998	1.000

TABLE 11

True Positive Rates for String 5 Copying Condition

10 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.000	.004	.000	.000	.000	.000	.000
Extreme	.006	.030	.000	.006	.002	.006	.004
.0005	.014	.044	.010	.010	.010	.018	.018
.0010	.018	.054	.010	.018	.016	.024	.028
.0025	.038	.088	.034	.038	.036	.048	.062
.0050	.066	.132	.064	.050	.064	.082	.120
.0075	.084	.144	.096	.066	.090	.086	.128
.0100	.096	.192	.112	.086	.100	.122	.126
.0500	.238	.344	.256	.238	.284	.304	.306
.1000	.380	.516	.406	.376	.388	.442	.420
.2500	.632	.726	.634	.572	.652	.700	.650

20 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.000	.058	.000	.016	.024	.000	.000
Extreme	.054	.126	.016	.048	.050	.022	.024
.0005	.096	.156	.080	.064	.104	.146	.142
.0010	.142	.178	.102	.118	.140	.194	.168
.0025	.206	.260	.252	.206	.258	.232	.268
.0050	.298	.322	.362	.256	.332	.392	.342
.0075	.362	.350	.420	.280	.416	.412	.358
.0100	.412	.434	.456	.316	.430	.504	.440
.0500	.688	.652	.700	.576	.700	.690	.594
.1000	.782	.812	.826	.736	.812	.820	.750
.2500	.912	.914	.928	.872	.930	.908	.840

30 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.044	.152	.038	.122	.182	.012	.034
Extreme	.304	.284	.214	.206	.280	.130	.214
.0005	.426	.332	.432	.262	.444	.438	.460
.0010	.500	.390	.474	.404	.524	.474	.492
.0025	.620	.510	.656	.518	.672	.540	.568
.0050	.716	.592	.742	.590	.752	.712	.592
.0075	.764	.622	.794	.628	.804	.684	.600
.0100	.782	.692	.814	.666	.824	.812	.714
.0500	.910	.846	.930	.868	.924	.878	.764
.1000	.946	.918	.962	.922	.956	.952	.900
.2500	.984	.976	.988	.964	.978	.974	.926

Table 11 (continued)

True Positive Rates for String 5 Copying Condition

40 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.234	.352	.196	.400	.568	.064	.136
Extreme	.708	.538	.590	.550	.674	.410	.522
.0005	.794	.606	.784	.608	.820	.714	.748
.0010	.834	.670	.814	.708	.856	.792	.756
.0025	.886	.774	.916	.810	.918	.826	.778
.0050	.928	.816	.940	.854	.944	.910	.784
.0075	.944	.848	.958	.874	.956	.872	.784
.0100	.946	.884	.962	.892	.962	.930	.878
.0500	.986	.970	.992	.968	.992	.956	.892
.1000	.998	.986	.998	.988	.992	.986	.960
.2500	1.000	1.000	1.000	.994	1.000	.998	.960

50 Items Copied

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.626	.554	.582	.704	.866	.224	.372
Extreme	.908	.740	.852	.798	.920	.622	.722
.0005	.936	.784	.948	.838	.944	.854	.844
.0010	.938	.824	.952	.892	.956	.898	.850
.0025	.966	.884	.968	.932	.976	.922	.854
.0050	.982	.922	.982	.950	.984	.956	.854
.0075	.982	.936	.994	.952	.988	.918	.854
.0100	.984	.962	.996	.966	.990	.960	.938
.0500	.998	.990	1.000	.994	.998	.980	.940
.1000	1.000	.996	1.000	.998	1.000	.996	.980
.2500	1.000	1.000	1.000	1.000	1.000	1.000	.980

TABLE 12

False Positive Rates for Conditional Benchmark Data

High Conditional Benchmark Data (≥ 86)

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.000	.000	.000	.000	.000	.000	.000
Extreme	.000	.000	.000	.000	.000	.000	.000
.0005	.000	.000	.001	.000	.002	.000	.000
.0010	.000	.000	.001	.000	.004	.000	.000
.0025	.000	.000	.007	.000	.005	.000	.000
.0050	.000	.000	.014	.000	.010	.000	.000
.0075	.000	.000	.022	.000	.012	.000	.000
.0100	.000	.010	.027	.000	.012	.000	.002
.0500	.000	.010	.104	.004	.050	.000	.002
.1000	.002	.010	.189	.011	.077	.000	.036
.2500	.008	.140	.386	.038	.150	.000	.064

Low Conditional Benchmark Data (62,61)

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.000	.000	.000	.000	.000	.000	.000
Extreme	.000	.000	.000	.000	.000	.000	.000
.0005	.000	.000	.000	.000	.002	.000	.002
.0010	.000	.000	.000	.002	.002	.000	.002
.0025	.006	.010	.002	.002	.003	.000	.002
.0050	.006	.010	.003	.003	.008	.010	.004
.0075	.010	.010	.006	.003	.016	.010	.006
.0100	.010	.010	.010	.006	.018	.014	.018
.0500	.065	.029	.043	.060	.084	.122	.088
.1000	.148	.029	.104	.131	.143	.184	.180
.2500	.387	.207	.274	.344	.331	.422	.404

Mixed Conditional Benchmark Data (75,60)

False Positive	B	H	g_2	P	CP	PAIR1	PAIR2
Beyond	.000	.000	.000	.000	.000	.000	.000
Extreme	.000	.000	.000	.000	.000	.000	.000
.0005	.000	.000	.000	.000	.000	.000	.000
.0010	.000	.000	.000	.000	.000	.000	.000
.0025	.000	.000	.000	.002	.002	.000	.000
.0050	.002	.009	.002	.002	.003	.002	.000
.0075	.002	.009	.002	.002	.003	.002	.002
.0100	.002	.009	.002	.003	.005	.006	.000
.0500	.021	.089	.012	.053	.053	.028	.022
.1000	.055	.089	.048	.100	.091	.128	.056
.2500	.259	.594	.166	.300	.238	.292	.206

TABLE 13

Comparison of Decision Rule Cutoff Based on Theory (T) and Observed Benchmark Data (O)

	Methods									
	B		H		g_2		P		CP	
False Positive	T	O	T	O	T	O	T	O	T	O
.0005	3.28	3.52	3.28	4.68	3.28	3.53	.0005	.00015	.0005	.00133
.0010	3.09	3.34	3.09	4.45	3.09	3.45	.0010	.00041	.0010	.00229
.0025	2.81	2.99	2.81	3.80	2.81	2.89	.0025	.00065	.0025	.00669
.0050	2.57	2.72	2.57	3.25	2.57	2.64	.0050	.00199	.0050	.01246
.0075	2.43	2.50	2.43	3.17	2.43	2.49	.0075	.00358	.0075	.02039
.0100	2.33	2.37	2.33	2.90	2.33	2.38	.0100	.00532	.0100	.02360
.0500	1.64	1.67	1.64	1.82	1.64	1.75	.0500	.00737	.0500	.09283
.1000	1.28	1.27	1.28	1.48	1.28	1.37	.1000	.05173	.1000	.15755
.2500	.67	.70	.67	.49	.67	.81	.2500	.29102	.2500	.33528

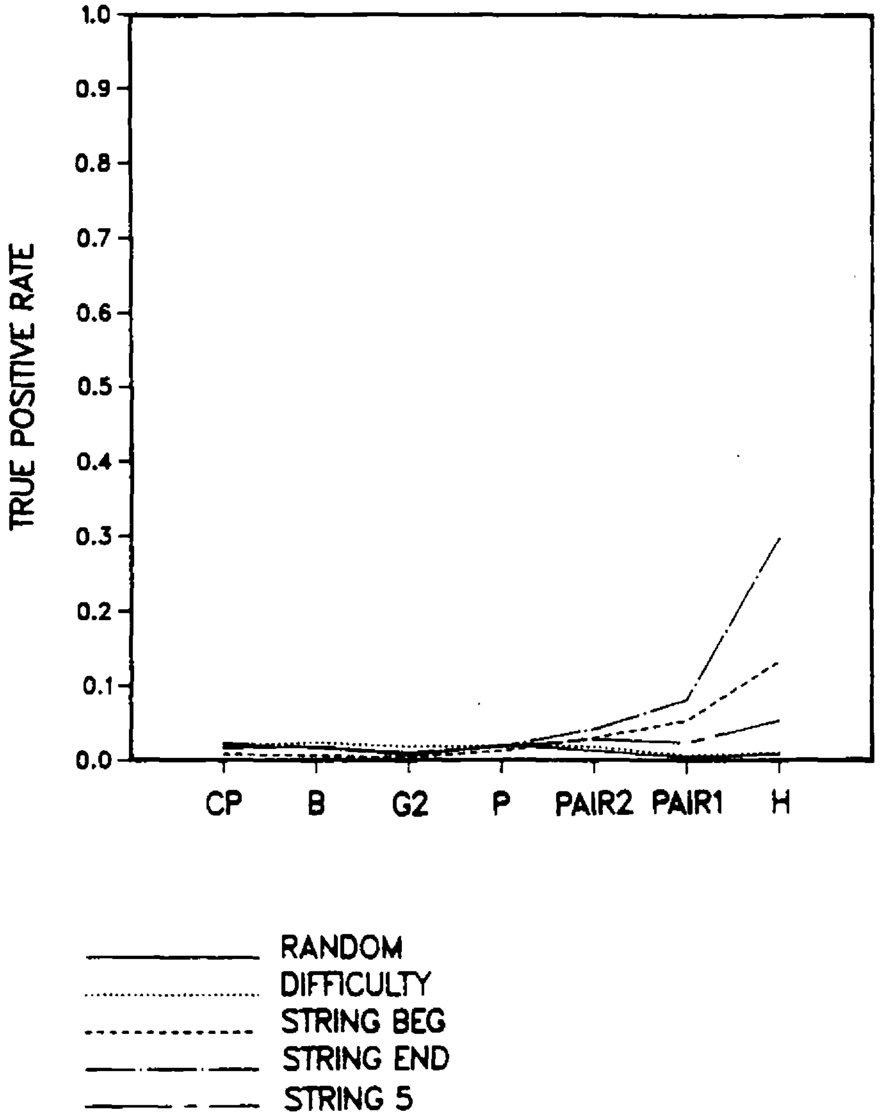

Figure 1. True positive rates corresponding to a false positive rate of .001 for 10 items copied.

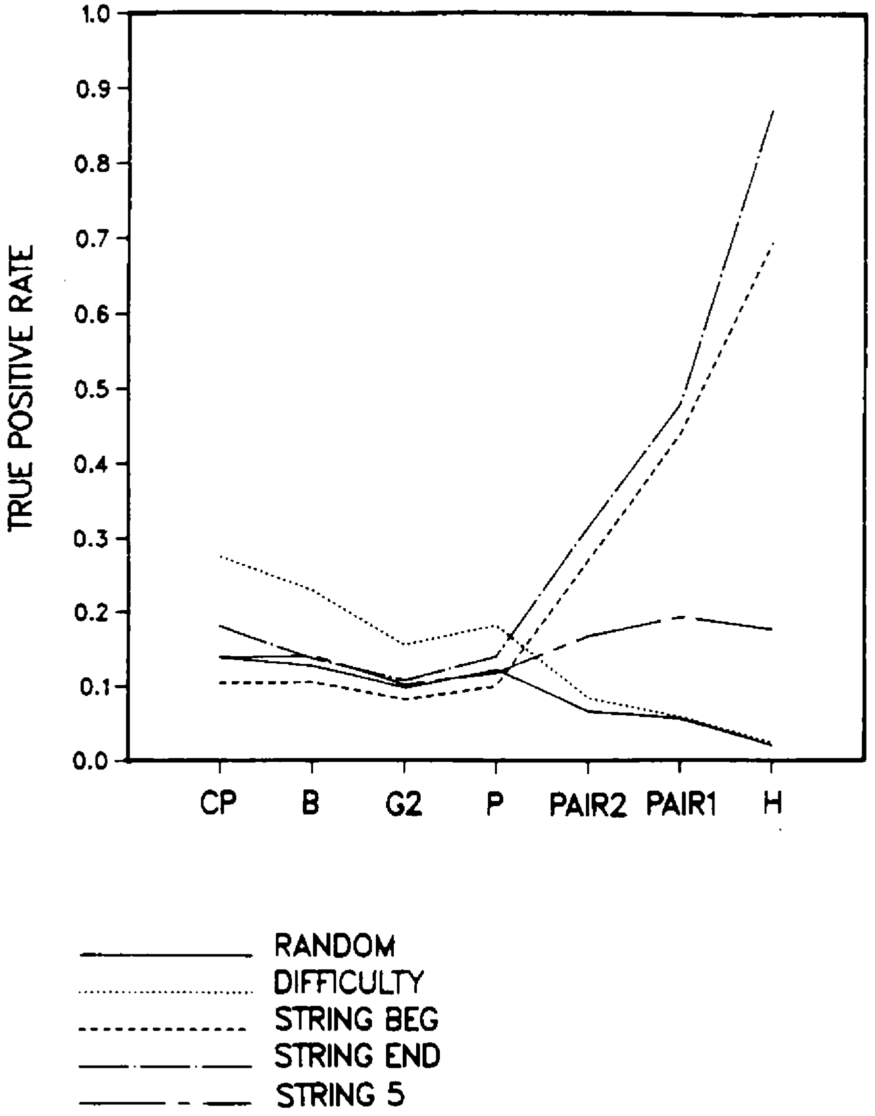

Figure 2. True positive rates corresponding to a false positive rate of .001 for 20 items copied.

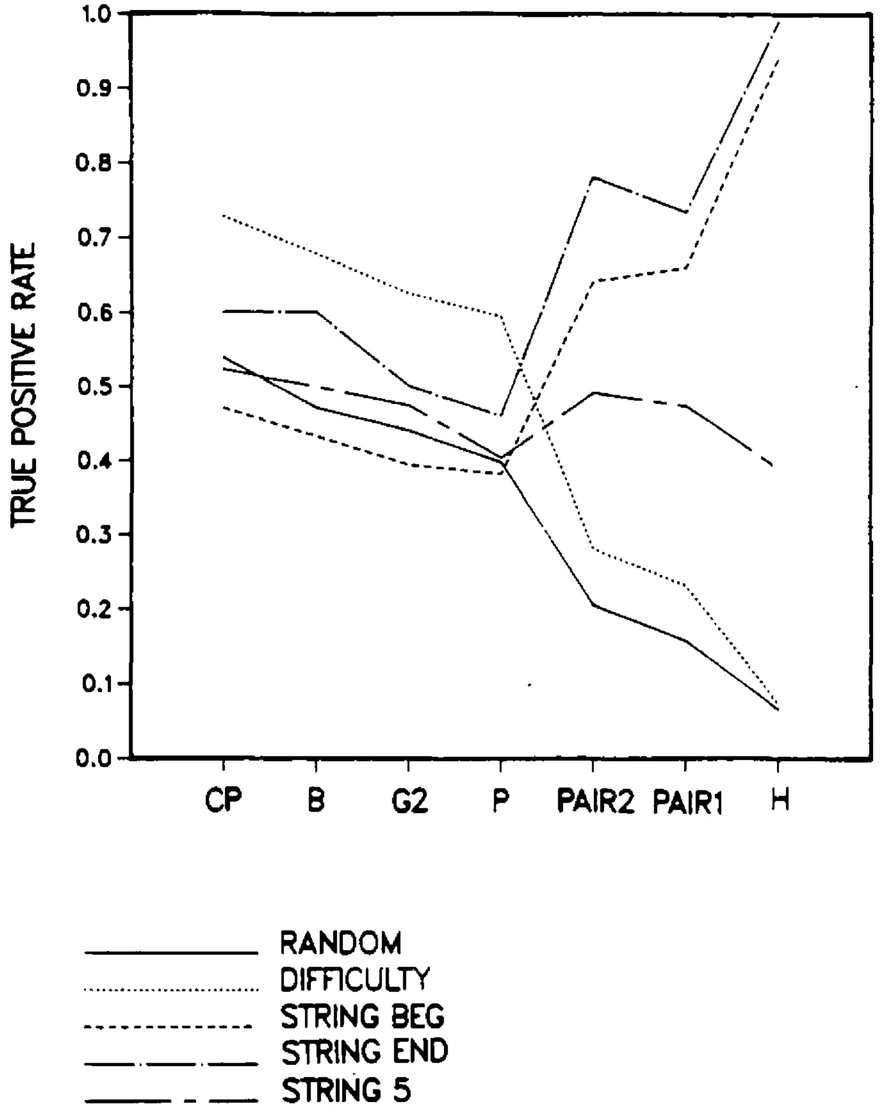

Figure 3. True positive rates corresponding to a false positive rate of .001 for 30 items copied.

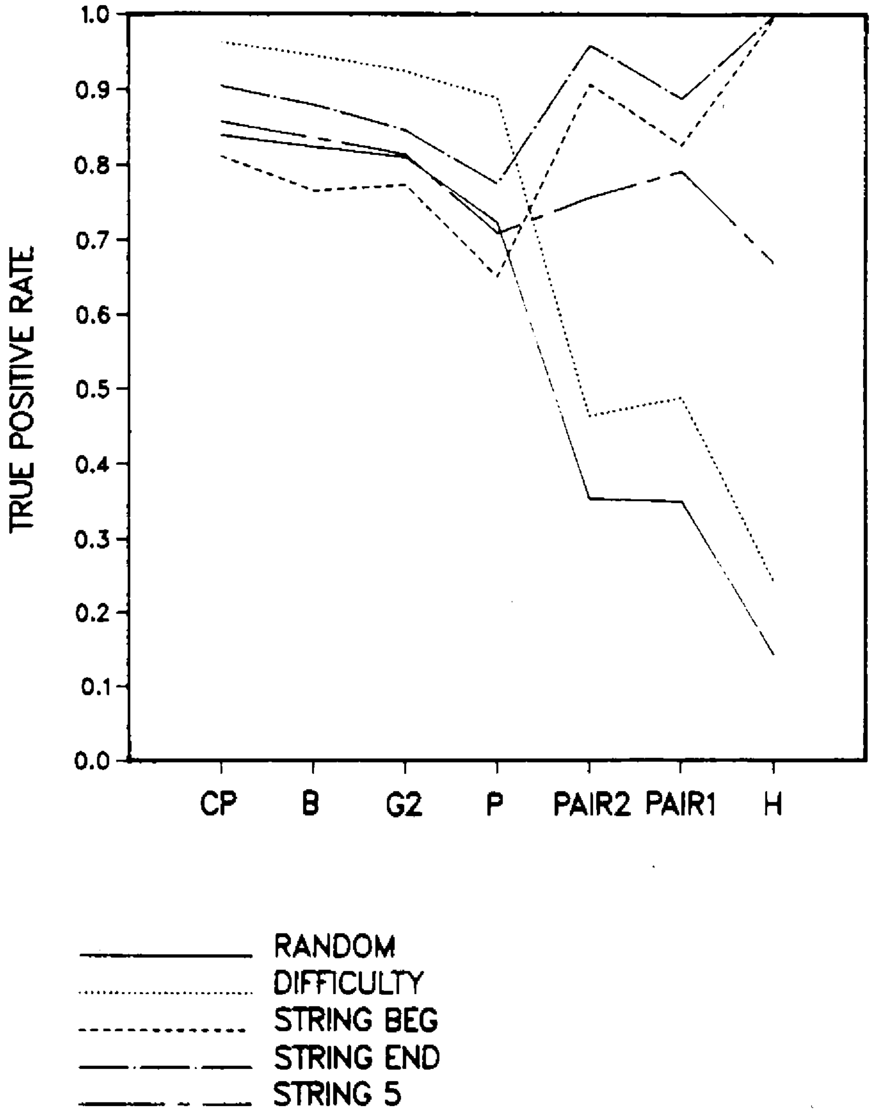

Figure 4. True positive rates corresponding to a false positive rate of.001 for 40 items copied.

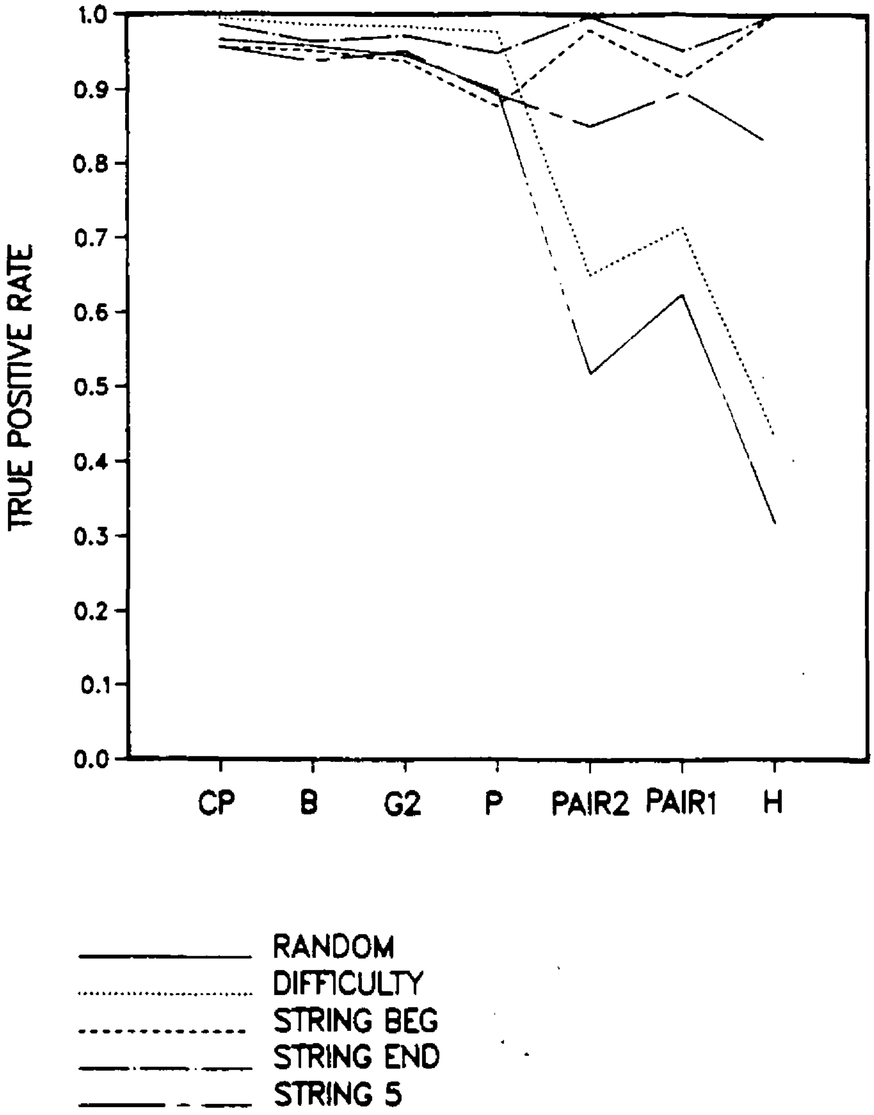

Figure 5. True positive rates corresponding to a false positive rate of .001 for 50 items copied.