DOCUMENT RESUME

ED 128 362 TM 005 484

AUTHOR Frary, Robert B.; Tideman, T. Nicolaus
TITLE Evaluation of Statistics for Detection of Cheating on Multiple-Choice Tests.
PUB DATE [Apr 76]
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Bayesian Statistics; *Cheating; Electronic Data Processing; *Multiple Choice Tests; *Probability; Response Style (Tests); *Statistical Analysis; Statistical Bias

ABSTRACT

The development of an index reflecting the probability that the observed correspondence between multiple choice test responses of two examinees was due to chance in the absence of copying was previously reported. The present paper reports the implementation of a statistic requiring less restrictive underlying assumptions but more computation time and a related Bayesian procedure designed to adjust the standard error estimates to counteract the effect of the presence of a substantial proportion of cheaters in a sample. The Bayesian adjustment did reduce the bias; however, the original index appears to be the most accurate and least expensive in terms of processing cost. With either method, results suggest that cheaters may be conclusively identified when they copy more than 50 percent of their answers from anyone answering less than 90 percent of test items correctly. (BW)

EVALUATION OF STATISTICS FOR DETECTION OF CHEATING ON MULTIPLE-CHOICE TESTS

Robert B. Frary and T. Nicolaus Tideman
Virginia Polytechnic Institute
and State University

Paper presented at the Annual Meeting of
the American Educational Research Association
San Francisco, California, April, 1976
(Session 11.11: Test Item Analysis, Division D)

Evaluation of Statistics for Detection of Cheating on Multiple-Choice Tests

Robert B. Frary and T. Nicolaus Tideman
Virginia Polytechnic Institute and State University

In an earlier paper, the authors reported development of indices reflecting the probability that the observed correspondence between multiple-choice test responses of two examinees was due to chance in the absence of copying (Frary and Tideman, 1975). Each of these statistics resembles the t statistic in that the numerator contains the difference between the observed and expected degree of correspondence and the denominator a sample-based estimate of the standard error of this difference. The statistics were designated the *fast t*, *intermediate t* and *slow t* according to the computer time required to produce them. *fast t*'s were actually computed and evaluated in a variety of situations, which suggested that this statistic is exceedingly effective in identifying cheaters.

The present paper reports the successful implementation of the *intermediate t* and a related "Bayesian" procedure designed to adjust the standard error estimates to counteract the effect of the presence of a substantial proportion of cheaters in a sample. Results are compared with those from the *fast t* in a novel manner which reveals the relative effectiveness of each statistic at varying levels of resources that an institution might apply for investigating possible cheating and prosecuting alleged cheaters.

Development of the *fast t*

The following paragraphs review the *fast t* to lend perspective to the newer outcomes reported thereafter. Computation of the *fast t* and other statistics described in this paper require as input the item responses of each examinee, readily available in any setting involving optical scanning of response sheets followed by the usual computer processing to produce item and total score analyses.

Expected Correspondence. For each examinee in the group under consideration, an individual probability of using each response including omissions on the test is computed. Initially, for each examinee, these probabilities are equivalent to the proportion of examinees choosing each choice. If the examinee under consideration has an above-average score, the probabilities for wrong answers are reduced by the ratio of the average score to his score, and these reductions are added to the probability corresponding to the right answer. For below-average examinees, the probability for the right answer is reduced by the ratio of the examinee's score to the average score, and this reduction is distributed among the wrong choices in proportion to their popularity. This procedure assures that the probabilities assigned to all answers are in the range of 0 to 1, that the sum of probabilities assigned to all answers to a given question is equal to 1, and that the sum of the probabilities attached to the right answers is equal to the student's score.

For a given examinee, i, the probabilities just described are used to determine the expected correspondence of each other examinee to i under the assumption of no cheating. The expected number of choices of examinee, i,

that would be the same as those of another examinee, j, is the sum over all responses of j of the assigned probabilities that i would use them.

Standard Error for Observed Correspondence. The variance of the observed number of j's choices that i uses would be the sum of all elements in a variance-covariance matrix for i's use of j's choices. If for an item of the test, the assigned probability of examinee i choosing j's choice is p, the corresponding variance estimate under the assumption of no cheating is $p(1 - p)$. In order to facilitate evaluation of the difference between the expected and observed number of j's choices that i used, the off-diagonal elements may be assumed to sum to zero. This assumption is not unreasonable because between right and wrong choices covariances would tend to be negative, while between two right or two wrong choices they would tend to be positive.

fast t. Under the assumption just stated, a statistic (the *fast t*) similar to the t statistic is the observed number of j's choices that i used minus the expected number divided by the square root of the sum of the variances for i of the choices that j used. A very high value of the statistic would suggest that i copied from j. The reverse hypothesis can be tested separately. Thus for every pair of examinees two *fast t*'s should be calculated, one for i's copying from j and one for j's copying from i.

Application of *fast t*. The authors have applied the *fast t* in a number of situations involving suspected cheating and in others where cheating was believed absent. In all cases the *fast t* has performed consistently, yielding extreme outlying values for most previously suspected cases of cheating. The *fast t* has also been applied to monitor the prevalence of cheating in connection with efforts to prevent it.

Modifications of the *fast t*. Three modifications to the *fast t* were proposed but not implemented in earlier work:

1) To take into account the intercorrelations among items. The resulting statistic was designated the *intermediate t*.
2) To use a "Bayesian" procedure to improve estimation of inter-item correlations possibly distorted by the presence of cheaters in the sample.
3) To use discriminant analysis to enhance distinctions between cheaters and noncheaters. This procedure would require inventing a matrix of order equal to the number of items on the test for each statistic (on a pair of examinees) generated. Hence its result was designated the *slow t*.

Application of the *intermediate t*

If X_k and X_m are the number of students who used responses k and m, X_{km} is the number of students who used both response k and response m, and N is the total number of examinees, then the correlation between responses k and m can be estimated as:

$$R_{km} = \frac{NX_{km} - X_k X_m}{\sqrt{X_k(N - X_k)X_m(N - X_m)}} .$$

This correlation across all examinees may then be taken as the estimate of interchoice correlation across many hypothetical, independent administrations of the test to the single examinee i being compared with j. R_{km} may then be converted to a covariance, V_{km}, by multiplying it by the product of the standard deviations of items k and m for examinee i. The denominator of the *fast t* is then changed from ΣV_{kk} to $\Sigma\Sigma V_{km}$, and the resulting statistic is designated the *intermediate t*.

The recommended "Bayesian" procedure is accomplished by separating interchoice correlations into categories: both choices wrong, both right, and one wrong and the other right. For each category, the mean and standard deviation of the elements is calculated for a distribution standardized to the interval (0,1). Then the standardized beta distribution with the same first and second moments is adopted as the prior distribution for each R_{km} in that category (see Raiffa and Schlaiffer, pp 218-20, 1968). The information in the sample for the R_{km} is then added, and the mean of the resulting posterior beta distribution, transformed back to the interval (-1,1), is then taken as the estimate of R_{km}.

Data for the study came from a test with 60 four-choice items administered simultaneously to 356 examinees in two rooms. There were two forms of the test, the second containing the same items as the first but in a substantially different order. Responses from the second form were reordered to correspond with the first and the test was split into two 30-item tests containing items 1-30 and 31-60.

Using the procedures described above, *fast t*'s, *intermediate t*'s and *intermediate t*'s with adjusted interitem correlation estimates were computed for both 30-item tests and for each pair of examinees in two mutually exclusive groups:

Group 1: Pairs of examinees who were in the same room *and* took the same form of the test. There were 19,210 such pairs, which for each statistic produced 38,420 values, since the computation changes when the potential copier and person copied from are interchanged.

Group 2: Pairs of examinees who were in different rooms *and* took different forms of the test. A total of 12,296 pairs were available, which yielded 24,592 values of each statistic.

Figure 1 shows the two *fast t* distributions for the first 30-item test for the groups described above. The distribution of the 12,296 *fast t*'s from Group 2 above has a slightly positive mean, .11, and a slightly positive skew, .21. Its standard deviation is .97. If it can be assumed that there was no crossform-crossroom cheating, this distribution may be taken as a norm with which to compare distributions arising from others subsamples of pairs of examinees on the same test. The slightly positive mean and skew are typical of numerous other *fast t* distributions studied by the authors, where cheating was believed absent. Also characteristic of such distributions is the absence of outliers, that is, the distribution for Group 2 appears to be monotonically decreasing in the high *fast t* range. The highest value observed in group 2 was 3.91.

Figure 2 shows the *intermediate t* distribution for the first 30-item test for Groups 1 and 2. For Group 2, the mean was .09, close to that for the corresponding Group 2 *fast t*'s. Also, the Group 2 skew was again slightly positive .20. However, the Group 2 standard deviation, .83, was substantially lower than for the *fast t*. The highest Group 2 *intermediate t* was 3.5.

For the second 30-item test, the *fast t* and *intermediate t* distributions were nearly identical with those shown in Figures 1 and 2. This result suggests that cheating was about equally prevalent over both halves of the test.

All distributions of *intermediate t*'s with adjusted interitem correlations were quite similar to the corresponding *intermediate t* distributions except for slightly larger standard deviations for both Groups 1 and 2.

Evaluation of Statistics

Inspection of the various Group 1 distributions reveals that the more extreme cases of correspondence are identified as probable cheaters by all three statistics. However, consider a *fast t* value of 3.2. Inspection of Figure 1 reveals that relatively about twice as many Group 1 examinee pairs correspond to this value as in Group 2. Therefore it is reasonable to assume that about half the Group 1 pairs with *fast t*'s of 3.2 cheated and half attained this degree of correspondence by chance in the absence of cheating. If some authority wished to investigate suspicious cases based on *fast t* values, it would be necessary to determine some *fast t* value below which investigations would not be made. This decision might be made on the basis of the percentage of innocent cases expected or on the basis of the resources available for investigation. In either case it is possible to estimate which statistic identifies the greater proportion or number of actual cheaters for a given number of investigations.

Figure 3 shows the relationship between number of cases investigated and, on the basis of each statistic, the number of guilty persons which subsequent investigation might be expected to confirm. This relationship was produced from the relative differences between the Group 1 and Group 2 distributions for the first test, under the assumption of two approximately equal statistics for each pair of examinees. A refined calculation based on the higher statistic for each pair did not seem justified since the *fast t* proved markedly superior and inspection of higher values of all three statistics showed the assumption to be largely correct. Each statistic reaches a plateau beyond which further investigations would produce very few additional confirmable cases of cheating.

Figure 4 shows the results for the second test, which yielded results similar to those for the first. If an authority were willing based on the second test to undertake investigating 75 individuals, the *fast t* might yield as many as 50 confirmations of guilt, while the *intermediate t* in either of its forms would yield no more than about 40. If investigative resources are more restricted or the authority believed the innocent/guilty ratio of 25/50 too high, other comparisons could be made. For example if 50 investigations is the maximum possible, confirmation of up to 43 guilty cases may be expected using the *fast t* for an innocent/guilty ratio of only 7/43. Investigation of 50 cases using either version of the *intermediate t* could yield up to about 40 confirmation of guilt with an innocent/guilty ratio of 10/40, about 50 percent higher than for the *fast t*. Of course, if 30 or fewer investigations are made use of any of the three statistics should identify almost 100 percent guilty cases. Results for the first test (items 1-30) were similar to those for the second. The *fast t* appeared superior in identifying cheaters as shown in Figure 3.

Discussion

One explanation of the somewhat surprising result that the simple statistic, the *fast t*, works best is that the assumption of zero correlations among pairs of responses is very close to the truth. The distributions of observed interitem correlation coefficients in the data from the first test are shown in Table 1. With average values so close to 0, and with the distributions centered so tightly around 0, there is little accuracy lost in employing the assumption that all covariances are 0 or that they sum to 0. Furthermore, the presence of cheaters in the sample will give an upward bias to the estimates of the correlation coefficients for the matrix of

response pairs that the cheaters use, which will increase the estimate of the denominator of the t-statistic for pairs of cheaters and therefore lower the calculated value. This ability of cheaters to increase the estimates of the covariances for their responses is reduced when the "Bayesian" adjustment is used. That is why the performance of the adjusted *intermediate* statistic is better than the unadjusted *intermediate*. But the effect of the bias that cheaters impart to the estimates of the correlation coefficients for their responses even with the "Bayesian" adjustment is probably greater than the effect of the simplifying assumption that all correlations are zero. Hence the *fast t* works better than the adjusted *intermediate t*.

Future Work

The earlier proposal for computation of the *slow t* with its associated matrix inversion requirement seems unlikely to yield improved identification of cheaters in the light of present results. In addition computer time even for the *intermediate t* becomes excessive when large numbers of items or examinees are involved. (CPU time was about 110 minutes on an IBM 370 for each of the tests reported above.) Accordingly, future work will be done under the assumption of zero covariances in the variance-covariance matrix for one examinee's use of another's choices.

This accommodation permits the application of discriminant analysis to enhance cheater/noncheater distinctions without a lengthy matrix inversion operation. However, preliminary tests of this procedure suggest the need for adjustment of the probabilities assigned to each examinee for using the various choices. This adjustment might be accomplished by another "Bayesian" procedure similar to that used to adjust interitem correlations.

REFERENCES

Frary, R.B. and Tideman, T.N. Detecting remarkable similarity between multiple-choice test responses. Paper presented at Annual Meeting of the American Psychological Association, Chicago, September, 1975. (Submitted for publication.)

Raiffa H. and Schlaiffer, R. *Applied Statistical Decision Theory.* Boston: MIT Press, 1968.

Table 1

Means and Standard Deviations of Distributions of Correlation Coefficients for Pairs of Responses Test 1 (items 1-30)

	Wrong-Wrong	Right-Wrong	Right-Right
Mean	.024	-.018	.056
Standard dev.	.064	.064	.070

Figure 1 - Relative Frequency Distributions for Two Sets of *fast t*'s from Test 1 (items 1-30)

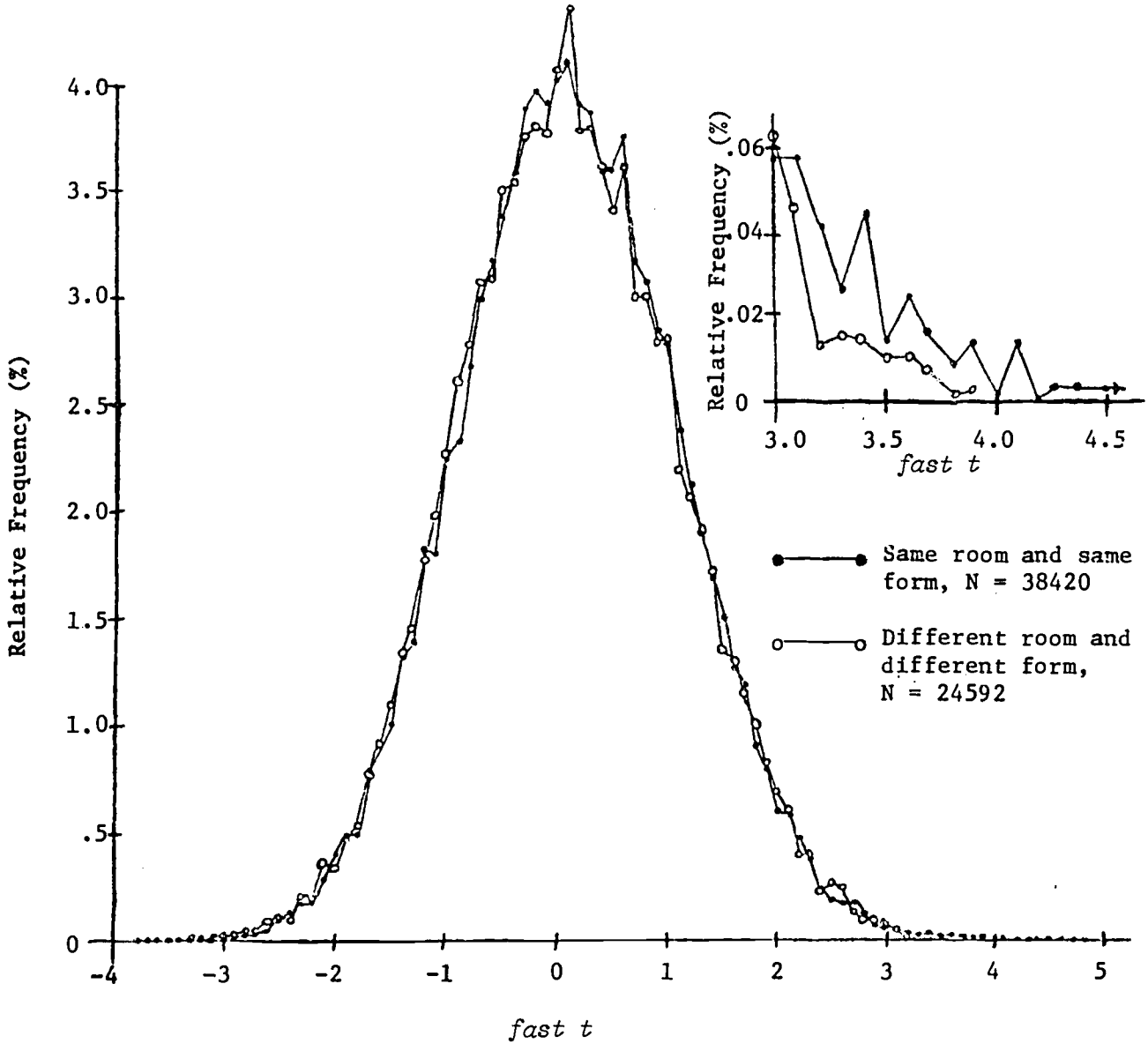

Figure 2 - Relative Frequency Distributions for Two Sets of *intermediate t*'s from Test 1 (items 1-30)

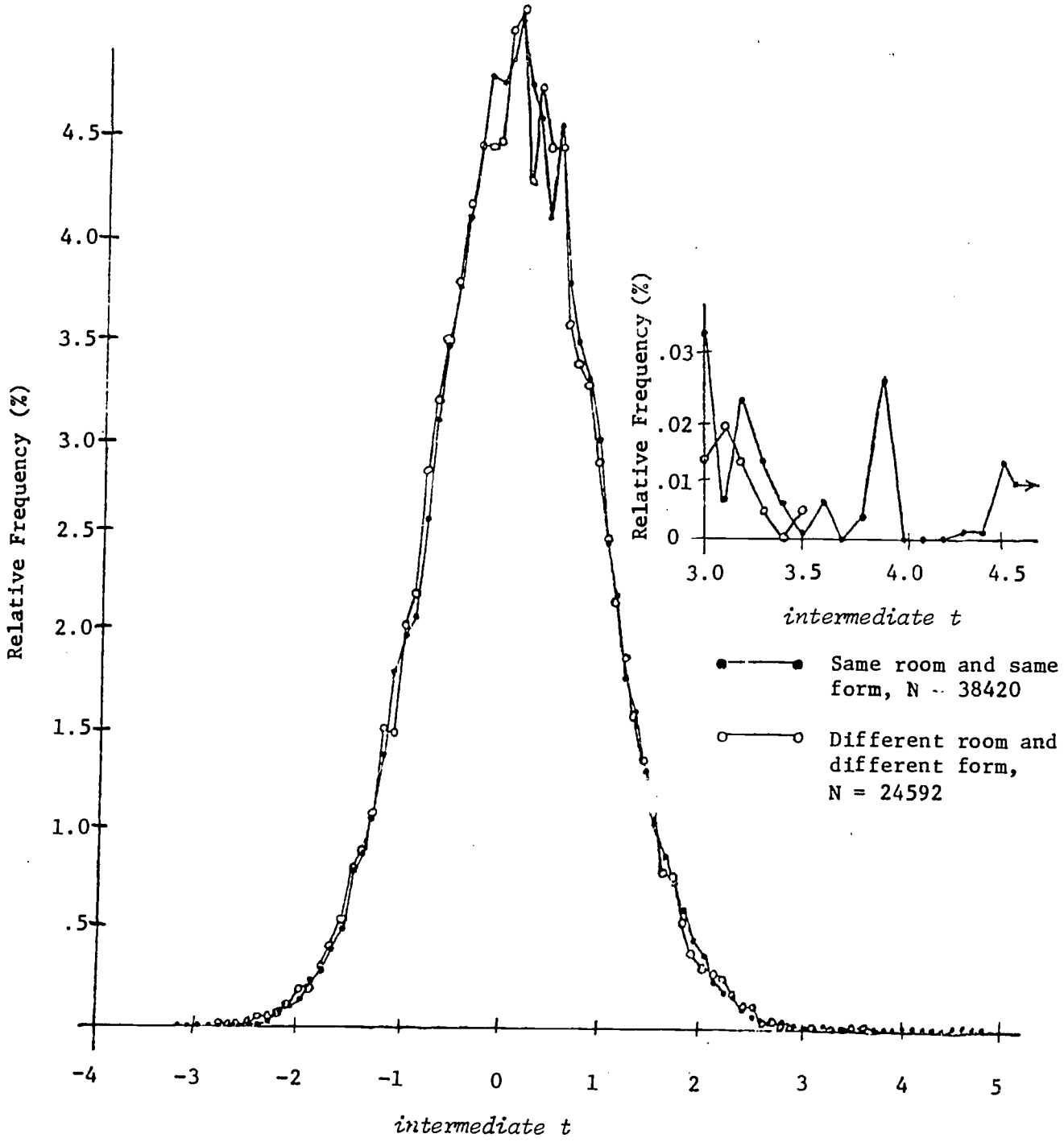

Figure 3 - Number of Confirmable Cases of Cheating as a Function of Number of Cases Investigated: Test 1 (items 1-30)

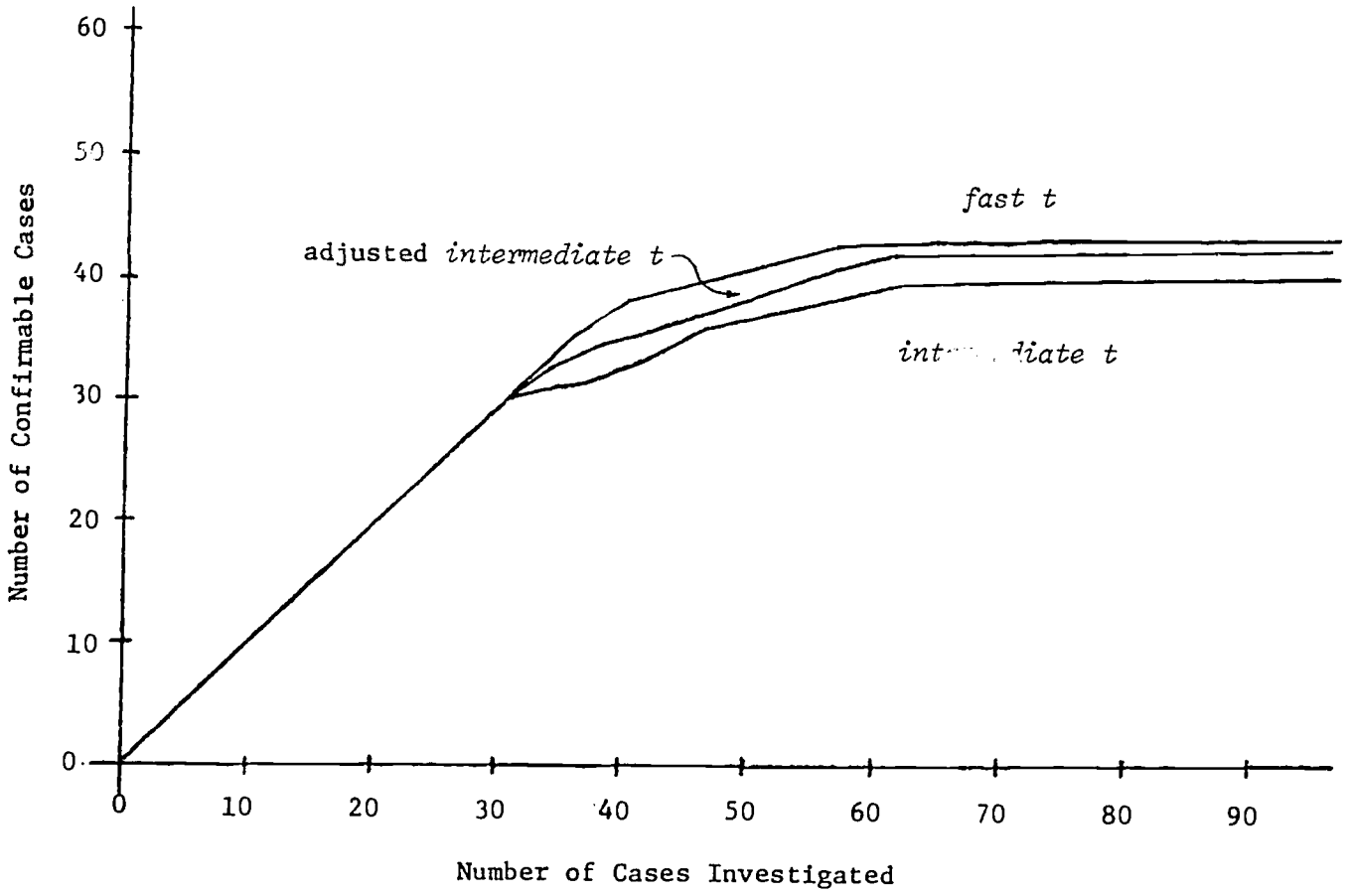

Figure 4 - Number of Confirmable Cases of Cheating as a Function of Number of Cases Investigated: Test 2 (items 31-60)