DOCUMENT RESUME

ED 069 687 TM 002 138

AUTHOR	Angoff, William H.
TITLE	The Development of Statistical Indices for Detecting Cheaters.
INSTITUTION	Educational Testing Service, Berkeley, Calif.; Educational Testing Service, Princeton, N.J.
REPORT NO	CEEB-RB-72-26; CEED-RDR-72-73-1
PUB DATE	Jul 72
NOTE	25p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Analysis of Variance; *Cheating; College Bound Students; *College Entrance Examinations; Data; *Evaluation Techniques; *Measurement Techniques; Research; Statistical Analysis; Technical Reports; *Testing
IDENTIFIERS	SAT Mathematical; SAT Verbal

ABSTRACT

Comparison data on SAT verbal and mathematical were collected on pairs of examinees in three samples for later use in detecting instances of willful copying. Two of the samples were constructed with the knowledge that no examinee could possibly have copied from the answer sheet of any other examinee in the sample. The third sample was taken entirely from a single center believed to be free of cheating. In each sample the answer sheet of each examinee was compared with the answer sheet of every other examinee. Eight detection indices were developed and distributions were run for possible operational use in making future judgments regarding examinees who were actually suspected of copying. Covariance analyses between samples indicated statistical but not practical significance, and consequently it was judged that any one of the samples could serve the purposes of operational detection as well as either of the other two. Empirical tryout of the indices against known and admitted copiers gave some results which permitted the elimination of three of the indices from further use. Practical considerations removed a fourth, and further statistical study eliminated two others. The remaining two have been in successful operational use at Educational Testing Service for more than two years. (Author)

COLLEGE ENTRANCE EXAMINATION BOARD
RESEARCH AND DEVELOPMENT REPORTS
RDR-72-73, NO. 1

RESEARCH BULLETIN
RB-72-26 JULY 1972

The Development of Statistical Indices for Detecting Cheaters

William H. Angoff

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY
BERKELEY, CALIFORNIA

THE DEVELOPMENT OF STATISTICAL INDICES FOR DETECTING CHEATERS[1]

The problems of cheating during test administrations may be dealt with in one or both of two ways: by discouraging and deterring cheating before it takes place and by detecting it and taking corrective action after it takes place. Most of the deterrent procedures are fairly obvious: They include identity checks, close supervision during the test, the use of two or more forms of the test distributed randomly throughout the testing room, planned seating arrangements to make cheating difficult, threats of punishment for detected cheating, etc. Methods of detecting individual cases of cheating fall into two general categories, depending on whether impersonation or copying was the method of cheating employed.

One solution to the impersonation problem is fairly straightforward, in theory, though often difficult to implement: One compares the handwriting shown on the suspect answer sheet with authentic specimens of handwriting. Here, of course, judgment plays an important role, and it is sometimes necessary to enlist the help of a handwriting expert.

Methods of detecting copying are also difficult, particularly after the test session is over and the answer sheets have been turned in. The obvious method is to compare the responses on the suspect answer sheet with responses on the answer sheets of examinees seated nearby and to look for greater-than-normal similarities. But the question of establishing the range of "normal" similarities itself presents a problem. One solution that suggests itself

---

[1]This research was supported by the College Entrance Examination Board. The author wishes to express his appreciation to the ETS Board of Review for their helpful comments and suggestions in reviewing this manuscript: J. E. Alloway, J. T. Campbell, F. R. Kling, J. S. Kramer, L. R. Lavine, W. B. Schrader, R. E. Smith, E. E. Stewart, and P. W. Williams.

involves the construction of a theoretical distribution of identical responses that would be expected in random pairs of answer sheets for examinees who are known to be honest. However, even a brief consideration of this solution makes it clear that the complexities in making theoretical estimates of such a distribution are far too great to make it practical. For example, the easy assumption that the options of an item are equally attractive, and therefore equally probable, is obviously false and unjustified. Therefore, in the construction of any distribution of similar responses made by random pairs of honest examinees one would have to take into consideration differences in the popularity of the responses. Secondly, although it would certainly make the task of developing those distributions a much easier one if one could assume that the items were uncorrelated, we know that such an assumption is an unreasonable one; the correct responses to a test are not uncorrelated. And even if one had reasonably good estimates of those correlations, the task of using them in generating the distributions appears to be formidable. When it is further recalled that intercorrelations among the patterns of incorrect responses would also have to be considered, it becomes even clearer that the task of developing these distributions theoretically approaches quite unreasonable proportions.

The present paper describes an effort to develop distributions of similar responses made by pairs of "honest" examinees to use in future work in detecting efforts to copy during test administrations. Because of the foregoing considerations, however, it was concluded that the only practical way to develop these distributions was to do so empirically. The remainder of this paper will describe the procedure of developing these distributions and the analyses that followed.

Samples

Three samples of examinees were drawn from actual test administrations, and identical indices were developed for all three samples for comparing the answer sheet of each examinee with the answer sheet of each of the other examinees in the sample. The first of these samples described below is the principal sample in the study and is the basis for the norms used in later actual detection work. The other two samples were used only for verifying the usefulness of the first sample.

1. Sample 1 was constructed by selecting every thousandth examinee taken from every odd-numbered computer tape from the December 1968 administration of the College Board SAT. In the selection of these cases care was taken that each examinee came from a different testing center. If an examinee was chosen who did in fact come from a center represented by a previously selected examinee, he was replaced with the next examinee who came from a unique center. This process of selection yielded a sample of 203 examinees. By comparing the item responses of each of the 203 with the other 202, it was possible to collect data on 20,503 pairs of answer sheets. Since these 203 examinees were sitting for the examination in different geographical locations, it was impossible for them to copy from one another's paper. In that sense, then, and for the purpose of these data, they were "honest" examinees and their responses were therefore usable for developing "norms for honest examinees."

2. Sample 2 was collected in order to check on the hypothesis that the answer sheets for examinees tested in the same geographical location might show greater similarities than the answer sheets of examinees sitting in separate locations, even though they were innocent of improper behavior. This hypothesis might be supported by the possibility that, for example,

examinees in the same center might have studied and learned the same misinformation from the same source. To determine, then, whether such similarities occur more often than similarities in answer sheets coming from different testing rooms, a center with an unblemished security history was chosen and data were developed by comparing the responses on each answer sheet in that center with the responses on each of the other answer sheets in that center. Since there were 122 examinees in that center, it was possible to make 7381 paired comparisons.

3. Sample 3 was also chosen as a check on the first. The purpose of the check was to determine whether data based on a different examinee group, responding to a different form of the SAT, might yield a different set of results. Clearly, if the "norms" to be developed could not be generalized but were unique to the form of the test and unique to the nature of the examinee group, then their usefulness in the course of future operational work in the detection of cheaters would be substantially diminished. Accordingly a set of data was developed by drawing a sample similar to Sample 1, one examinee from each of 209 centers, but taken from the March 1969 administration when a different form of the SAT was given. With 209 examinees in Sample 3, a total of 21,736 paired comparisons were made.

Variables

The observations for each of the variables listed below were derived from the examination of the responses of pairs of examinees, where i = one examinee in a pair and j = the other examinee in that pair. Parallel sets of variables were derived for SAT-verbal and SAT-mathematical.

R_iR_j = the number of items answered correctly by examinee i <u>times</u> the number of items answered correctly by examinee j .

R_{ij} = the number of items answered correctly by both i and j .

W_iW_j = the number of items answered incorrectly by i times the number of items answered incorrectly by j .

W_{ij} = the number of items answered incorrectly by both i and j .

Q_{ij} = the number of items answered incorrectly in the same way (i.e., by making the same incorrect response) by both i and j .

O_iO_j = the number of items omitted by i times the number of items omitted by j . (Note that an "omit" is defined as a nonresponse to an item that appears prior to the last item attempted in the test; hence omits do not include items "not reached.")

O_{ij} = the number of items omitted by both i and j .

W_i (or W_j), whichever is smaller.

O_i (or O_j) for the examinee whose W_i (or W_j) was the smaller.

$S_i = W_i + O_i$.

$S_{ij} = Q_{ij} + O_{ij}$.

K_{ij} = the longest "run" of identically marked incorrect responses and omits. Before defining the "run," it will be useful to define a "succession" of items. This is a consecutive block of items in which all items are marked (or unmarked) in precisely the same way: correct, incorrect, or omit. The "run" is the number of items answered incorrectly in the same way by both i and j (i.e., the Q_{ij}) within the succession plus the number of items omitted by both i and j (i.e., the O_{ij}) within the succession. (Note that although the succession is the length of a consecutive block of items, the run within that succession may not be consecutive. Note also that in any ij comparison there may be more than one run.) K_{ij} is defined as the longest run in an ij comparison.

Analyses

Using the foregoing 12 variables bivariate distributions were prepared for the eight indices shown in Table 1, eight for SAT-verbal and eight for SAT-mathematical. The intent in developing these indices was that in the investigation of an actual case of suspected copying the departure for that case of the value on the dependent variable from the mean of the norms group would be examined, but only after controlling on the independent variable. The value of the dependent variable for that case, or its departure from the mean of the array, is referred to here as the "index of copying."

The first phase of the analysis was conducted in order to evaluate the degree to which the norms tables derived from these bivariate distributions could be generalized to other data. Accordingly, two sets of covariance analyses were conducted: (1) to determine whether the regression systems formed with the data of Sample 1 were significantly different from those of Sample 2; and (2) to determine whether the regression systems resulting from the data of Sample 1 were significantly different from those of Sample 3. The intent of these analyses was to determine whether the data of Sample 1, which presumably would form the basis for developing the norms, were idiosyncratic in the sense that (1) they would behave differently from data collected for noncheaters who were assembled for the test administration in the same room; and (2) they would behave in a way that was somehow characteristic of the particular form of the SAT used at that test administration and/or characteristic of the examinees tested at that time.

The method of analysis of covariance followed the model developed by Gulliksen and Wilks (1950), in which the regression systems are tested successively for differences in errors of estimate, slopes, and intercepts. Tables

Table 1

Description of Copying Indices

Bivariate Distribution (Index)	Independent Variable (x)	Dependent Variable (y)
A	$R_i R_j$	R_{ij}
B	$W_i W_j$	Q_{ij}
C	W_{ij}	Q_{ij}
D	$O_i O_j$	O_{ij}
E	W_i	Q_{ij}
F	O_i	O_{ij}
G	S_i	S_{ij}
H	S_i	K_{ij}

2 and 3 summarize these results and show that for the most part the differences are indeed significant, some of them far beyond the one per cent level. However, in evaluating these results the sizes of the samples on which these analyses were based must be kept in mind. For the purpose of these analyses Sample 1 consisted of 20,503 "cases" (i.e., comparisons, which were not entirely independent in this study); Sample 2 consisted of 7381 "cases," and Sample 3 consisted of 21,736 "cases." (The numbers of actual examinees, it is recalled, were 203, 122, and 209.) With "sample sizes" of these magnitudes even very small differences would have been found to be significant. Indeed, detailed examinations of the array means on the dependent variables for these three samples at each interval on the independent variables revealed only trivial differences. In some very rare instances, as in the data that gave the most highly significant results for the tests of intercepts--e.g., in Table 2, in the test for Index A, Mathematical; also, in Table 3, in the test for Index G, Mathematical--the means of the arrays for the separate samples differed by only two and one-half points at most, and even then only when the data in the arrays were sparse and very likely unstable. In the very large majority of instances the means for Samples 2 and 3 would have rounded to the same whole number as for Sample 1 and would have led to precisely the same conclusion as that based on the data for Sample 1 in the disposition of any actual security case. Accordingly, it was judged that the data of Sample 1 would be sufficiently general to use in developing the "norms."

Validation

The second phase of the analysis involved an attempt to validate the indices and to determine, if possible, which one(s) were most useful in identifying actual cases of copying. From the data already available it was possible to determine the extent to which the independent variable (see Table 1) involved

Table 2

Analyses of Covariance

Sample 1 vs. Sample 2

Bivariate Distribution (Index)	Values of Chi Square		
	Errors of Estimate*	Slopes*	Intercepts*
Verbal			
A: R_iR_j vs. R_{ij}	41.00*	85.59**	12.00**
B: W_iW_j vs. Q_{ij}	10.42**	1.19	1.24
C: W_{ij} vs. Q_{ij}	0.75	9.96**	8.09**
D: O_iO_j vs. O_{ij}	456.28**	32.85**	27.53**
E: W_i vs. Q_{ij}	7.93**	1.04	2.10
F: O_i vs. O_{ij}	94.64**	208.77**	370.76**
G: S_i vs. S_{ij}	2.81	18.63**	107.38**
H: S_i vs. K_{ij}	37.30**	20.31**	36.10**
Mathematical			
A: R_iR_j vs. R_{ij}	0.01	352.91**	1496.30**
B: W_iW_j vs. Q_{ij}	0.07	10.88**	209.49**
C: W_{ij} vs. Q_{ij}	7.48**	3.54	135.25**
D: O_iO_j vs. O_{ij}	1530.57**	375.65**	52.25**
E: W_{ij} vs. Q_{ij}	9.32**	11.50**	156.15**
F: O_i vs. O_{ij}	1312.94**	62.99**	410.23**
G: S_i vs. S_{ij}	9.93**	48.12**	74.08**
H: S_i vs. K_{ij}	20.40**	4.34	65.51**

*One degree of freedom

**Significant beyond 1% level

Table 3

Analyses of Covariance

Sample 1 vs. Sample 3

Bivariate Distribution (Index)	Values of Chi Square		
	Errors of Estimate*	Slopes*	Intercepts*
Verbal			
A: R_iR_j vs. R_{ij}	1.92	7.57**	427.87**
B: W_iW_j vs. Q_{ij}	4.35	22.36**	54.22**
C: W_{ij} vs. Q_{ij}	0.03	1.72	30.89**
D: O_iO_j vs. O_{ij}	264.32**	5.14	43.61**
E: W_i vs. Q_{ij}	0.02	28.24**	16.41**
F: O_i vs. O_{ij}	882.40**	15.15**	5.63
G: S_i vs. S_{ij}	185.06**	0.02	1.02
H: S_i vs. K_{ij}	103.79**	5.53	4.28
Mathematical			
A: R_iR_j vs. R_{ij}	66.30**	0.68	36.97**
B: W_iW_j vs. Q_{ij}	122.96**	26.83**	882.41**
C: W_{ij} vs. Q_{ij}	50.05**	2.68	685.76**
D: O_iO_j vs. O_{ij}	2192.11**	100.84**	842.49**
E: W_i vs. Q_{ij}	27.89**	3.52	693.64**
F: O_i vs. O_{ij}	105.08**	168.28**	806.22**
G: S_i vs. S_{ij}	2.16	45.15**	1071.05**
H: S_i vs. K_{ij}	64.57**	10.87**	397.30**

*One degree of freedom

**Significant beyond 1% level

in each of the indices was useful as a control when referring to the information provided by the dependent variable. Table 4 provides this information in the form of correlation coefficients between the independent and dependent variables involved in each index. Correlations are given for each of Samples 1, 2, and 3, for the verbal and mathematical sections of the test.

The validities of the separate indices cannot be anticipated from these correlations, however, but need to be determined empirically against an independent criterion of known copying. To this end answer sheets for a group of 50 cases of known and admitted copiers from recent administrations were assembled, together with answer sheets for the individuals from whom they copied. For each case 16 t-values, 8 verbal and 8 math, were calculated, based on Sample 1 data, each describing the deviation of the "index of copying" of that case from the mean of the appropriate array. For example, in considering the verbal index, W_iW_j vs. Q_{ij}, the product W_aW_b was calculated, representing the number of items person a was observed to answer incorrectly (W_a) times the number of items answered incorrectly by the person from whose answer sheet he admitted copying (W_b). Then the value Q_{ab}, the number of items that person a and person b were observed to answer incorrectly in the same way--for example, by marking response position d when c was correct--was recorded. Referring to the bivariate distribution of W_iW_j vs. Q_{ij} for Sample 1, the particular array of Q_{ij} was examined for this interval of W_iW_j. The value, $t = (Q_{ab} - \overline{Q}_{ij})/s_{Q_{ij}.W_iW_j}$, was then determined. As already mentioned, 16 t-values of this sort were calculated, 8 for SAT-verbal and 8 for SAT-mathematical, corresponding to the scatterplots described above in Table 1. The rule was adopted in advance that any ab comparison for which any one of the 16 t-values equalled or exceeded 3.0 represented a validation

Table 4

Correlations of Paired Variables Used in

Developing the Detection Indices

Index	SAT-verbal			SAT-mathematical		
	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
A: R_iR_j vs. R_{ij}	.972	.971	.971	.975	.966	.973
B: W_iW_j vs. Q_{ij}	.692	.705	.765	.678	.709	.750
C. W_{ij} vs. Q_{ij}	.759	.756	.808	.774	.784	.811
D: O_iO_j vs. O_{ij}	.892	.910	.906	.848	.848	.910
E: W_i vs. Q_{ij}	.688	.698	.754	.695	.706	.739
F: O_i vs. O_{ij}	.508	.463	.468	.493	.443	.479
G: S_i vs. S_{ij}	.596	.522	.583	.624	.596	.605
H: S_i vs. K_{ij}	.353	.282	.331	.339	.240	.337

of the general procedure. The result of a tabulation of these t-values revealed that every one of the 50 cases was identified as a copying case by at least one of the 16 indices, with most of the t-values ranging from 3.0 to about 23.0 (there was one additional t-value of 27.5 and still another of 45.0!).

The question remained, which of these types of indices were most useful, in terms of their statistical and practical value, for use in operational detection? To answer this question a count was made of the number of times these copying cases were actually detected by each of the eight types of indices. These frequencies of detection are reported in Table 5.

The first and second columns of frequencies in Table 5 report the number of t-values equalling or exceeding 3.0 for each of the eight indices in the Verbal and in the Mathematical sections of the test. The third column merely gives the sums of the frequencies in the first two columns. Finally, since not all of these 50 students necessarily copied on <u>both</u> sections of the test--some appeared to have copied on the verbal section only, others on the mathematical section only--the last column shows the number of cases in the group of 50 that would have been detected on Verbal and/or Math by each of the eight indices.

It appears from an examination of these frequencies that the most successful indices were those involving counts of Rights and those involving counts of Wrongs, especially Index A (R_iR_j <u>vs.</u> R_{ij}), Index B (W_iW_j <u>vs.</u> Q_{ij}), Index E (W_i <u>vs.</u> Q_{ij}), Index G (S_i <u>vs.</u> S_{ij}), and Index H (S_i <u>vs.</u> K_{ij}). The least successful were those involving counts of Omits: Index D (O_iO_j <u>vs.</u> O_{ij}) and Index F (O_i <u>vs.</u> O_{ij}). In order to reduce the number of indices to a manageable size for operational work, Indices D and F were therefore eliminated from

Table 5

Frequencies of Detection of Actual Copying Cases

Index	Frequencies			
	Verbal	Math	Verbal plus Math	Verbal and/or Math
A: R_iR_j vs. R_{ij}	34	37	71	44
B: W_iW_j vs. Q_{ij}	44	37	81	47
C: W_{ij} vs. Q_{ij}	37	22	59	41
D: O_iO_j vs. O_{ij}	8	14	22	18
E: W_i vs. Q_{ij}	44	32	76	47
F: O_i vs. O_{ij}	2	6	8	7
G: S_i vs. S_{ij}	40	36	76	48
H: S_i vs. K_{ij}	41	41	82	49

consideration. Index C (W_{ij} vs. Q_{ij}), which appeared from the frequencies in Table 5 to be somewhat less useful than those initially listed (A, B, E, G, and H), was also eliminated. This left five indices for further consideration. However, five indices were still too many, and there was little question that further reduction was needed.

It seemed likely that quite apart from its statistical validity Index A (R_iR_j vs. R_{ij}) might not be as easily defended and justified to the satisfaction of the typical layman as the other indices. The examinee could argue in his own behalf that a large number of right answers in common with another examinee should be expected since (he could claim) both he and the other examinee were able and knowledgeable students. Therefore, if the ultimate judgment that cheating has occurred is to be made by nonstatisticians, the fact that the R_{ab} -value in his case was significantly higher than the $\bar{R}_{ij}$ for examinees with the same R_iR_j may not be convincing.

This line of reasoning was considered to be sufficiently persuasive to cause the reduction of the number of potential (and presumably face-valid) indices to four: B (W_iW_j vs. Q_{ij}), E (W_i vs. Q_{ij}), G (S_i vs. S_{ij}), and H (S_i vs. K_{ij}). In order to make further selections among these indices, intercorrelations based on Sample 1 data were run among the errors of estimate associated with each index. For example, if Index B is taken as $x_1 - b_{12}x_2$, where Q_{ij} is redefined for simplicity's sake as Variable 1 and W_iW_j is redefined as Variable 2, and if, similarly, Index G is taken as $x_5 - b_{54}x_4$, the correlation between Index B and Index G can be expressed as

$$r_{BG} = \frac{r_{15} - r_{12}r_{25} - r_{14}r_{54} + r_{12}r_{24}r_{54}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{54}^2}} .$$

Table 6 gives the intercorrelations among Indices B, E, G, and H. Correlations among the indices for SAT-verbal appear above the diagonal; correlations among the indices for SAT-mathematical appear below the diagonal.

From the correlations in Table 6 it appears that the overlap between Index B and Index E is sufficiently great ($r = .905$ for verbal; $r = .916$ for math) to warrant dropping one of them. Both of these indices, it is recalled, depended on an examination of Q_{ij}, the number of items answered incorrectly, and in the same way, by both examinees in the comparison. What distinguishes Index B from Index E, it is recalled, is that the former uses W_iW_j, the product of the numbers of wrong responses by i and j, as the control variable and that the latter uses W_i, the number of wrong responses made by examinee i or j, whichever is smaller. Index B appeared on a priori grounds to be the more attractive index because it took into consideration information based on both candidates, rather than just one. It also derives from the logic, as suggested by Saupe (1960), that the expected value of Q_{ij} is the value W_iW_j/K, where K = no. of items in the test. (Saupe actually developed this point in terms of the values, R_{ij} and R_iR_j.) On the other hand it is worth considering that the expected variance of Q_{ij} should depend on the particular values of W_i and W_j separately, since the smaller of the two values imposes an upper limit on Q_{ij}; when W_iW_j is 400, for example, the value of Q_{ij} could be as high as 20 if W_i and W_j are each 20, but only as high as 10 if W_i were 10 and W_j were 40. Ultimately, the decision was made to use Index B (W_iW_j vs. Q_{ij}), in preference to Index E (W_i vs. Q_{ij}) on the basis that it identified the known copiers with more consistency than Index E (see Table 4). With Index E eliminated, the remaining three indices, B, G, and H, were reduced to two, B and H, largely on the basis of the lower correlations of B with H (.382 for SAT-verbal and .516 for SAT-math).

Table 6

Intercorrelations among Indices B, E, G, and H

(Based on Sample 1; N = 20,503)

	B	E	G	H
B		.905	.751	.382
E	.916		.737	.339
G	.802	.803		.582
H	.516	.492	.609	

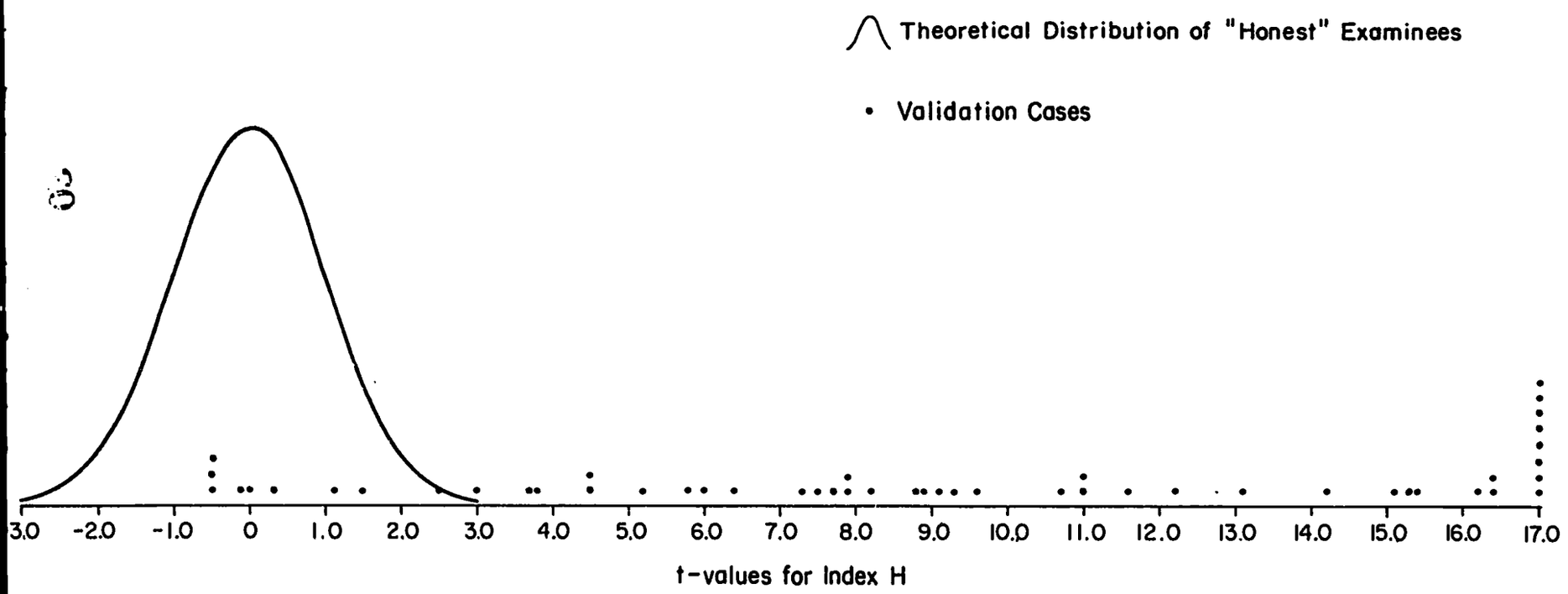
Figure 1 illustrates the sensitivity of Index H (K_{ij} , controlling on S_i) in detecting the copying in the validation group. The distribution shown at the left describes the degree of variation in Index H to be expected in a group of examinees who are not copiers, with t extending from -3σ to $+3\sigma$. The dots shown near the baseline of the graph, most of them to the right of the distribution, represent the frequencies of the t-values for the 50 validation cases on SAT-verbal. (The eight dots plotted at X = 17 represent t-values of 17 or higher for eight examinees in the validation group. Space did not permit plotting the higher t-values, which, as mentioned earlier in this paper, ranged as high as 45.) The appearance of these dots far beyond normally expected values of t makes it dramatically clear that most of these 50 examinees did indeed copy from a neighbor's answer sheet. Now it is also noted that nine of the 50 dots are represented by t-values lower than 3.0, four of them lower than .00. Although Index H fails to show that these nine copied on SAT-verbal, it does show (but not in Figure 1) that eight of the nine copied on SAT-mathematical. Thus, only one of the 50 cases was missed by Index H.

Application

Current operational work in detecting copiers depends most heavily on Index B (Q_{ij} , controlling on W_iW_j). When Index B fails to reveal that a suspected examinee has copied from another paper, data for Index H (K_{ij} , controlling on S_i) are also examined. Data for the other indices may also be used, but only in instances of uncertainty. However, experience with Indices B and H, even when used alone, has been quite satisfactory.

Although the security procedures at Educational Testing Service are under constant review and refinement, they are subject to a philosophy that is

Fig. 1. Values of t for Index H (S_i vs. K_{ij}) as expected in a group of "honest" examinees and as found in the validation sample (SAT-verbal).



intentionally and explicitly permissive. No candidate who is suspected of copying is investigated further in operational work unless one of the indices in use departs from the mean of the appropriate array in the data for Sample 1 by 3.72 standard deviations or more, representing a confidence level of less than 1 in 10,000 (assuming normal distributions in the arrays). Thus, only if an examinee's paper shows such a strong similarity to another examinee's paper that such an occurrence would be observed less than once in 10,000 in comparisons made of the papers of honest examinees would the investigation of the examinee's case be continued. (Lists of smoothed values, used to implement these procedures, are shown for illustration as Tables 7 and 8, below.) In the course of this investigation the examinee may be asked to take a retest to confirm his questioned score. If he agrees, arrangements are made for retesting under standard conditions and he is given the same form of the test on which he received the questioned score. If, on this retest, he earns a score more than 100 points lower than the questioned score, then the questioned score is cancelled. Otherwise the questioned score is confirmed. All communications and arrangements for retesting are made privately between the examinee and ETS. Information regarding the events is withheld from the examinee's high school and colleges of application, except on the initiative of the examinee himself.

Summary

Comparison data on SAT-verbal and mathematical were collected on pairs of examinees in three samples for later use in detecting instances of willful copying. Two of the samples were constructed with the knowledge that no examinee could possibly have copied from the answer sheet of any other examinee in the sample. The third sample was taken entirely from a single center believed to be free of cheating. In each sample the answer sheet of each examinee was

Table 7

Decision Points for Index B*

SAT-Verbal		SAT-Mathematical	
W_iW_j	Q_{ij}	W_iW_j	Q_{ij}
0- 99	4	50- 99	6
100- 199	5	100- 149	7
200- 299	6	150- 199	8
300- 399	7	200- 299	9
400- 499	8	300- 349	10
500- 599	9	350- 449	11
600- 699	10	450- 549	12
700- 799	11	550- 649	13
800- 899	12	650- 799	14
900-1099	13	800- 899	15
1100-1299	14	900-1049	16
1300-1499	15	1050-1199	17
1500-1699	16	1200-1349	18
1700-1999	17	1350-1499	19
2000-2199	18	1500-1649	20
2200-2499	19	1650-1799	21
2500-2799	20	1800-1949	22
2800-3099	21	1950-2099	23
3100-3499	22		
3500-4099	23		

*Defined as occurring in "honest" comparisons no more frequently than once in 10,000 times.

Table 8

Decision Points for Index H*

SAT-Verbal		SAT-Mathematical	
S_i	K_{ij}	S_i	K_{ij}
1- 8	2	1- 8	3
9-21	3	9-18	4
22-34	4	19-40	5
35-47	5		
48-60	6		

*Defined as occurring in "honest" comparisons no more frequently than once in 10,000 times.

compared with the answer sheet of every other examinee. Eight detection indices were developed and distributions were run for possible operational use in making future judgments regarding examinees who were actually suspected of copying. Covariance analyses between samples indicated statistical but not practical significance, and consequently it was judged that any one of the samples could serve the purposes of operational detection as well as either of the other two.

Empirical tryout of the indices against known and admitted copiers gave some results which permitted the elimination of three of the indices from further use. Practical considerations removed a fourth, and further statistical study eliminated two others. The remaining two have been in successful operational use at Educational Testing Service for more than two years.

References

Gulliksen, H., & Wilks, S. S. Regression tests for several samples. Psychometrika, 1950, 15, 91-114.

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