Army Project Number
29563000

Contract DA-49-083 OSA-638
C-9-252-30

PRB Technical Research Note 33

THE ASSESSMENT OF PARTIAL KNOWLEDGE IN OBJECTIVE TESTING

Clyde H. Coombs
John E. Milholland
Frank B. Womer

Engineering Research Institute
University of Michigan
Ann Arbor

January 1955

PRB Technical Research Notes are intended primarily for research agencies in the Armed Forces as a means of guiding further research in the area of human resources. While conclusions affecting military policy or operations may appear in this report, they are not intended as a basis for official action.

engn

UMR0934

ACKNOWLEDGEMENTS

The writers are indebted to a number of individuals who generously gave their time and interest. Through the willing cooperation of Dr. Kent Leach of the University of Michigan Bureau of School Services, contact was made with Mr. W. Earl Holman, Principal of Jackson High School, Jackson, Michigan. It was because of Mr. Holman's interest in educational research that his entire junior and senior classes were made available for this project. A program of this sort is very disruptive of the normal flow of activity in a high school, and we greatly appreciate his willingness to meet our research needs. Dr. Philip DuBois of Washington University, St. Louis, Missouri, generously allowed the use of items from an experimental form of the DuBois-Gleser Object Aperture Test and also rendered invaluable service as a consultant on IBM procedure along with Mr. Kurt Benjamin, then Director of the University of Michigan Tabulating Service. Through the assistance of Mr. Earl Allgaier of the American Automobile Association, Washington, D.C., several of their tests on driving information were made available. Finally, our recognition of assistance is gratefully extended to Dr. David Birch and to several graduate students, David Beardslee, William Hays, and J. E. Keith Smith, who participated in a running seminar on the project throughout its duration and made many valuable criticisms and suggestions.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF TABLES (continued)

LIST OF FIGURES

BRIEF

Conventional procedures for giving and scoring multiple-choice tests may not fully use the information that could be obtained from this test form. The examinee simply indicates the alternative he thinks is correct; he is scored only on his one choice; and the other alternatives function merely as distracters. The objective of the study described in this research note was to find out whether measuring partial, as well as complete, knowledge (as indicated by responses to multiple-choice items) would result in better tests and more meaningful spread of scores.

A new method of administering and scoring multiple-choice items was used. In administering the four-alternative, multiple-choice items, the examinees were instructed to cross out those alternatives which they knew to be wrong. Each of the three wrong alternatives (of an item) correctly crossed out was scored one point; but a right alternative incorrectly crossed out was scored a penalty of minus three points. Thus, a range of seven score points was possible for each item (from minus three to plus three). An examinee who crossed out only the right alternative of a given item received a score of minus three; an examinee who crossed out the right alternative and one wrong alternative received a score of minus two, and so on, up to the examinee who crossed out the three wrong alternatives and received a score of three.

Three tests of vocabulary, driver information, and spatial visualization were administered to 855 junior and senior high school students. These examinees were divided into three matched groups, one group for each of three testing procedures—the conventional method, the experimental method, and a procedure combining both methods.

The findings may be summarized as follows:

1. There was objective evidence that partial information influenced the responses to the multiple-choice items.
2. On the average, the experimental method increased the reliability of the tests to a degree equivalent to a 20-percent increase in the tests' effective length. What the effectiveness of the method would be after repeated experience was not investigated.
3. The experimental method appeared to measure the same complex of abilities as the conventional method.
4. The experimental method was better able to discriminate among the examinees.
5. The examinees expressed a preference for the experimental method over the conventional procedure.

These advantages may be outweighed in the operational situation by increased testing time and the complexity of scoring; however, comparative cost information is not yet available.

CHAPTER I. INTRODUCTION

The general acceptance of the multiple-choice-type test item as the best one for objective measurement of aptitude or achievement does not imply that its merits are optimal. Any variation upon an already widely accepted and useful technique which indicates promise of improved measurement is deserving of further investigation. A response method* for multiple-choice items which has certain theoretical advantages over the conventional response method is considered here, and this study is an empirical investigation of some of its relative merits and demerits.

The conventional response method (C method) for multiple-choice items requires selecting and marking the answer from among the choices offered. In this study one of four was to be picked. The conventional item score in a power test is one point when the answer is chosen and zero when a distracter is chosen. Complete information leads to an item score of one and misinformation to a score of zero. Partial information may lead to a score of either one or zero. The inability of the conventional method to discriminate between partial information and complete information or misinformation is a disadvantage. A second disadvantage is the encouragement of guessing. The conventional correction formula, used with speeded tests, only attempts to compensate for guessing, not penalize it.

The experimental response method (E method) under study here[1] attempts to differentiate between various degrees of partial information and partial misinformation. The task presented to the examinees is that of selecting and marking the distracters rather than the answer.** Differential choice of distracters

*Originality for the method is not being claimed here. It has apparently been part of the lore among psychometricians for some years although no mention of it has been found in the literature. One of the referees of this paper has been teaching the method for 15 to 20 years and it is not original with him.

**A proposed method[3] complementary to this is one called the free-choice method in which an individual marks as many choices as he desires, to be sure of having selected the right answer. With appropriate scoring procedures this method is formally isomorphic to the one studied here, but whether the task is psychologically complementary is an experimental question.

allows an examinee to exhibit varying degrees of partial information or partial misinformation which is not possible in the conventional method. For a four-choice item with one answer and three distracters, one point credit is gained for each distracter correctly identified (marked) and three points credit is lost if the answer is incorrectly identified as a distracter (marked). Thus, item scores for a four-choice item may range from minus three to plus three, and each score represents a different amount of information. This seven-point item score scale should produce greater item and test variance than the conventional two-point item score scale. In addition, the experimental method has the advantage of penalizing random guessing associated with partial information.

The line of argument that gives rise to this method begins with the notion that while an individual may not know the answer to an item, he may know some of the things which are wrong. This is called partial knowledge. If he knows the answer then he knows all of the things which are wrong and has complete information. If he thinks the correct alternative is wrong he has misinformation, and if in addition he also recognizes some of the distracters as wrong it becomes partial misinformation. These notions would seem to be applicable only to those items which have one and only one answer among the choices and even here there may be some domains to which the method is not applicable. One thinks of an arithmetic item in which an individual arrives at the answer and then seeks it among the several choices.

A serious disadvantage of the method may be the difficulty of altering established sets with respect to answering multiple-choice items. If examinees who actually have only partial information proceed by first selecting an answer and then marking the remaining alternatives as distracters, the purposes of the method are defeated since it is in the differential recognition of distracters that it differs from the conventional one.

On the practical side, there are the disadvantages of increased time for scoring and, very possibly, increased time for administration. Present scoring procedures require that each paper be scored twice—once for distracters crossed out (a positive score), and once for answers crossed out (a negative score)—and then these two scores added. It should be possible to devise scoring methods which would cut the time to less than twice that for conventional scoring.

No attempt was made in this study to compare the experimental and conventional methods with respect to administration time. However, experience with the experimental method in regular classroom testing, especially after students have become familiar with the method, indicates that little additional testing time is required.

The problem of the correction for guessing does not arise, but in its place there is a question of the standard of assurance. Suppose, for example, an individual has crossed out two distracters he knows to be wrong on an item and has a sure two points to his credit. There remain two choices, one the answer and the other a distracter. If he has no knowledge and guesses, he is gambling an additional point credit vs a three-point loss on a fifty-fifty chance. This is not a profitable game to play. A number of individuals with identical knowledge would presumably not behave alike in this situation. Some would take a chance and others would need to feel more assured as to which choice was the distracter before marking it. Over a number of items these individuals would get different scores because of their different standards of assurance. This variable is one of temperament, like utility for risk, and it would be of interest and value if a measure of it could also be secured.

The primary purposes of this study were a comparison of test reliabilities produced by using the experimental response method and the conventional response method, and a comparison of certain item selection techniques appropriate to the two methods. In addition, comparisons of test validities and of a coefficient of discrimination were made; two standard-of-assurance indices were evaluated; and some implications of partial information for the conventional correction for guessing formula were examined.

CHAPTER II. DESIGN OF THE STUDY

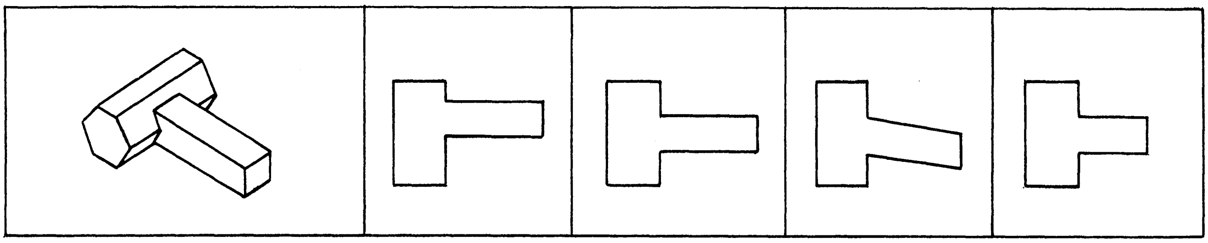
A. THE TESTS

Three tests were developed for use in this study to provide greater generality of results than a single test would. A vocabulary test, a test of driver information, and a test of spatial visualization were developed. Each test contained forty items.

Vocabulary items were constructed for the purpose and were of the form of a lead word followed by four choices, one of which had the same meaning as the lead, e.g., "FALL: autumn, spring, summer, winter."

Driver information items were secured from several tests of sportsmanlike driving of the American Automobile Association, e.g., "Just before making a left turn you should always: blow your horn, increase your speed, give a turning signal, drive in the extreme right hand lane."

Spatial visualization items were obtained from an experimental form of the DuBois-Gleser Object Aperture Test. These are of the form of a lead element which is a drawing in perspective of a three-dimensional object followed by four drawings of irregularly shaped holes, only one of which the object could slide over and drop through, e.g.,



An attempt was made to select items in which one or two distracters were easily identifiable, in order that partial information would be available for most examinees.

B. THE TESTING DESIGN

Three different methods of testing were used: (1) the conventional method [C], (2) the experimental method [E], and (3) the "both" method [B]. The first two are described in the introduction. The B method was designed so that a single test could be scored by both the C method and the E method for the same subjects. In the B method the examinees were asked to rank three distracters according to ease of recognition. The unmarked choice is the one that would have been marked as the answer in the C method. They were then asked to circle those of the ranked distracters which they were confident were actually distracters and which they wanted to be scored as such by the E method.

The subjects were divided into three groups and the testing programmed so that each subject would use each method once and take each test once, and so that each test would be taken by each method once. Table 1 illustrates the design for administering tests. The three experimental groups are designated by Roman numerals.

TABLE 1

TESTING.DESIGN FOR THE THREE GROUPS OF SUBJECTS

		Response Method	
Test	Conventional (C)	Experimental (E)	Both (B)
Vocabulary	I	II	III
Driver Information	III	I	II
Object Aperture	II	III	I

C. TESTING PROCEDURE

In each test group the C method test was administered first, the E method test next, and the B method test last. There were two testing sessions, two weeks apart, but at each session the subjects participating in that session were given all three tests. Three hours and fifteen minutes were available for the first of the two administrations. Time limits were estimated for each test and each set of instructions. These estimated times proved to be too generous, and they were reduced to a total of two hours and fifteen minutes for the second administration.

Three steps were taken in an effort to secure maximally honest responses. First, the importance of the testing program was stressed in a "press release" to the high school. The interest of the United States Army was mentioned. Second, provision was made to report examinees' scores back to them through their counselors. Third, an attempt was made to select test items with high face validity.

Administration was handled by experienced test administrators. Proctoring was done by advanced psychology students who were specially trained for these testing sessions.

More time than perhaps was necessary was taken in giving the instructions for the tests, particularly in the E and B sessions but it was imperative that the students fully understand what they were to do. The tests were administered with generous time limits in order that the same N could be used on all items which were retained for analysis in a given test. The duration of instructions and testing is indicated in Table 2.

TABLE 2

INSTRUCTION AND TESTING TIME

Test	Instruction			Testing		
	C	E	B	C	E	B
Vocabulary	10	9	10	15	25	35
Driver Information	13	13	10	25	35	*63
Object Aperture	17	14	22	25	30	38

*This time was artificially extended to use up the time allotted by the school for the testing program.

A question naturally arises as to the relative instruction time and testing time for a given test administered by the conventional response method compared with administration by the experimental response method. This was not a subject of investigation in this study and Table 2 offers no evidence on this. The student took his three tests in the order C, E, B, and in the first test, by the C method, time was taken to familiarize him with the IBM answer sheet so that little additional time was required on this aspect for the subsequent tests.

As to testing time, because the tests were to be used as power tests, the time to stop was left to the discretion of the administrator and a later adjustment was made in the length of the test and size of the group to achieve a constant N over all the items in a test. The procedures employed in this editing of the data are described in the next section.

D. EDITING OF DATA

Decisions as to how to treat mistakes in recording responses were made prior to the test administration. All mistakes which actually happened were foreseen, as well as some which did not occur. Mechanical recording mistakes were corrected, e.g., marking the fifth response position on an answer sheet when the items had only four responses. Omitted items were scored zero since the subjects were asked to answer all questions.

For purposes of item analysis it was desirable to have equal N's for all of the test items used in the computations. Since not all examinees finished every test, it was decided not to use some of the items at the end of each test. Twelve items were eliminated from one test and eight each from the other two. There were 21 examinees who did not complete the reduced-length tests. The test scores of these examinees were removed from the data. The loss of 21 examinees was less than 2-1/2 percent of the total. Whatever bias might be introduced by eliminating these examinees was deemed unavoidable in order to keep from losing any more data than was already lost with the elimination of twenty-eight items out of one hundred and twenty. Table 3 shows how many examinees were lost in each group and which tests they did not finish.

TABLE 3

NUMBER OF EXAMINEES ELIMINATED
BY TEST AND BY GROUP

Test		Group	
	I	II	III
Vocabulary (Eliminate 37 through 44*)	0	1	2
Driver Information (Eliminate 33 through 44*)	0	10	0
Object Aperture (Eliminate 37 through 44*)	5	1	2
Total N**	292	269	294

*Items numbered 1 to 4 were used in the instructions.

**After elimination of items and subjects.

E. SUBJECTS

The subjects were 855 juniors and seniors of Jackson High School, Jackson, Michigan, tested in two sessions, the majority on April 8 and the remainder on April 21, 1953.

The distributions of the subjects by sex and class in each experimental group are given in Table 4. It may be noted that Group I had some preponderance of seniors over juniors, whereas in the other two groups the reverse was true. The age distribution in each experimental group is indicated in Table 5.

TABLE 4

SEX AND CLASS DISTRIBUTIONS IN THE EXPERIMENTAL GROUPS

Experimental	Class		
Group	Junior	Senior	Total
I			
Male	50	93	143
Female	57	92	149
Total	107	185	292
II			
Male	83	62	145
Female	77	47	124
Total	160	109	269
III			
Male	88	58	146
Female	98	50	148
Total	186	108	294
Grand Total			
Male	221	213	434
Female	232	189	421
Total	453	402	855

Within each experimental group a random split was made into two subgroups, A and B, for item analysis. It is pertinent, therefore, to examine the test performance of the groups to see if any important differences exist.

TABLE 5

AGE DISTRIBUTIONS IN THE EXPERIMENTAL GROUPS

Statistic	Group I	 II	 III	 Total
N	292	269	294	855
Range	15-10 to 22-5	15-11 to 20-5	15-5 to 20-10	15-5 to 22-5
Q_3	18-1	17-11	17-9	17-11
Median	17-7	17-4	17-4	17-5
Q_1	17-1	16-11	16-9	16-10

The results on the tests are shown in Table 6, which gives the ranges and various point measures in the score distributions. The differences between the item analysis subgroups all seem to be trivial: for the C scores there is one difference of two points, the rest are 0 or 1; and for the E scores all differences but three are less than four points.

A within- and between-groups analysis of variance was carried out, and the results are given in Table 7. Two of the variance ratios were statistically significant at the 1-percent level of confidence and these two were attributable to the better performance of Group I on the Object Aperture Test than the other two groups.

F. REFERENCE VARIABLES

In order to determine whether the three experimental groups were matched on aptitude, a series of reference variable scores were secured from the high school's files. These scores were not complete for all examinees since they had been secured at different educational levels. The one which was least complete had scores from 38 percent of the examinees; the one which was most complete had scores from 74 percent.

Sixteen reference variable scores were available from five different tests. The Differential Aptitude Test yielded eight scores: verbal reasoning, numerical ability, abstract reasoning, space relations, mechanical reasoning, clerical speed and accuracy, language usage—spelling, and language usage—sentences. This test had been administered during the tenth grade.

The Stanford-Binet IQ was available. It had been administered during

TABLE 6

PERFORMANCES OF THE GROUPS ON THE TESTS

Statistic	Subgroup	Vocabulary				Driver Information				Object Aperture			
				B				B				B	
		C	E	C Score	E Score	C	E	C Score	E Score	C	E	C Score	E Score
		I	II	III	III	III	I	II	II	II	III	I	I
Range	A	7-30	108-176	1-27	69-180	7-26	93-166	7-27	77-162	5-31	80-192	12-32	112-192
	B	9-28	92-178	10-28	106-183	11-27	97-166	6-27	92-164	5-31	84-192	7-32	92-192
Mean	A	18.83	141.41	18.02	140.14	19.05	132.15	19.59	132.46	23.54	160.86	26.34	168.24
	B	18.44	139.52	18.08	140.50	19.10	134.58	19.24	130.88	22.94	164.22	26.36	167.96
Q_3	A	22	152	21	150	22	143	22	142	28	180	29	182
	B	21	151	21	152	22	145	21	140	28	180	30	184
Median	A	18	140	18	142	19	134	19	134	25	169	27	172
	B	18	140	18	141	19	135	19	132	24	170	27	172
Q_1	A	16	131	15	130	17	120	17	124	21	147	24	157
	B	16	128	15	128	17	125	16	121	19	153	24	161

TABLE 7

RESULTS OF WITHIN- AND BETWEEN-GROUPS ANALYSIS OF VARIANCE FOR THE SIX EXPERIMENTAL GROUPS ON THE THREE TESTS

Test	Kind of Score	Group and Method of Administration			Mean Square		F	d.f.	Significance
		C	E	B	Within	Between			
Vocabulary	C	IA IB		IIIA IIIB	17.76	20.33	1.14	3/582	NS
	E		IIA IIB	IIIA IIIB	266.55	83.67	3.19	559/3	NS
Driver Information	C	IIIA IIIB		IIA IIB	12.78	4.67	2.74	559/3	NS
	E		IA IB	IIA IIB	198.47	332.33	1.67	3/557	NS
Object Aperture	C	IIA IIB		IA IB	30.63	327.33	10.69	3/557	S(<1%)
	E	IIIA IIIB		IA IB	466.69	1786.00	3.83	3/582	S(<1%)

kindergarten. The language, nonlanguage, and total IQ's were available from the California Intelligence Test. It had been administered during the ninth grade. The overall, mechanical, and clerical IQ's were available from the Detroit General Aptitude Test. It had also been administered during the ninth grade. The last reference variable score was from the MacQuarrie Mechanical Aptitude Test. It had been administered during the sixth grade.

An analysis of variance of the scores on the sixteen reference variables was carried out and the results are presented in Table 8. Only one F was significant at the 5-percent level, so there is no reason to reject the hypothesis of equivalence of groups on the reference variables.

The levels of performance for all students on whom reference variable scores were available are shown in Table 9. These results indicate that the students tested in this study were fairly representative of the norm groups for the reference tests.

TABLE 8

RESULTS OF WITHIN- AND BETWEEN-GROUPS ANALYSES OF VARIANCES FOR THE SIX EXPERIMENTAL GROUPS ON SIXTEEN REFERENCE VARIABLES

Variable	Mean Square		F	d.f.	Significance
	Within	Between			
Differential Aptitude Test					
Verbal reasoning	84.70	90.80	1.07	5/439	NS
Numerical	70.36	61.60	1.14	442/5	NS
Abstract reasoning	100.75	181.60	1.80	5/445	NS
Space relations	565.03	252.80	2.24	440/5	NS
Mechanical reasoning	202.91	99.00	2.05	441/5	NS
Clerical speed and accuracy	110.38	17.40	6.34	447/5	S (5%)
Language usage:					
Spelling	696.94	869.00	1.25	5/424	NS
Sentences	233.33	287.00	1.23	5/442	NS
Stanford-Binet IQ	167.03	295.00	1.77	5/484	NS
California Intelligence Test					
Total IQ	164.59	230.00	1.40	5/644	NS
Language IQ	211.06	335.60	1.59	5/644	NS
Nonlanguage IQ	221.42	387.60	1.75	5/644	NS
MacQuarrie Mechanical Aptitude Test	92.12	25.20	3.66	521/5	NS
Detroit General Aptitude Test					
IQ	185.41	187.60	1.01	5/328	NS
Mechanical IQ	151.68	122.00	1.24	328/5	NS
Clerical IQ	153.11	110.60	1.38	328/5	NS

G. DISTRIBUTIONS OF SCORES

The C and the E score distributions of the experimental tests are presented in the form of frequency polygons in Figs. 1, 2, 3, and 4. The distributions indicate that the Object Aperture Test was easier than the Vocabulary or Driver Information Tests and, in fact, somewhat too easy for optimum usability in a study of this kind.

TABLE 9

REPRESENTATIVENESS OF JACKSON HIGH SCHOOL JUNIORS AND SENIORS ON CERTAIN SELECTED REFERENCE VARIABLES, 1953

Reference Variables	Jackson H.S. Median	National Norms Median		N[9]
		Male	Female	
Differential Aptitude Test[1]				
Verbal reasoning	22	22[5]	22	445
Numerical ability	20	18	17	447
Abstract reasoning	31	29	28	451
Space relations	51	49	38	446
Mechanical reasoning	34	41	25	447
Clerical speed and accuracy	54	51	59	453
Language usage:				
Spelling	45	38	56	430
Sentences	34	30	38	448
MacQuarrie Mechanical Aptitude Test[2]	45	44[6]		527
Stanford-Binet IQ[3]	111	Mean = 108[7]		490
California Intelligence Test[4]		Otis Mean = 106[8]		
Total IQ	107			650
Language IQ	108			650
Nonlanguage IQ	104			650
Detroit General Aptitude[4]				
IQ	98			334
Mechanical IQ	99			334
Clerical IQ	95			334

1. Given in tenth grade.
2. Given in sixth grade.
3. Given in kindergarten.
4. Given in ninth grade.
5. Tenth grade norms, Form A.
6. Median for twelve year olds.
7. Mean Stanford-Binet IQ of high school seniors from three studies.[5]
8. Mean Otis IQ of high school seniors from six studies.[5]
9. Total N of juniors and seniors used was 855.

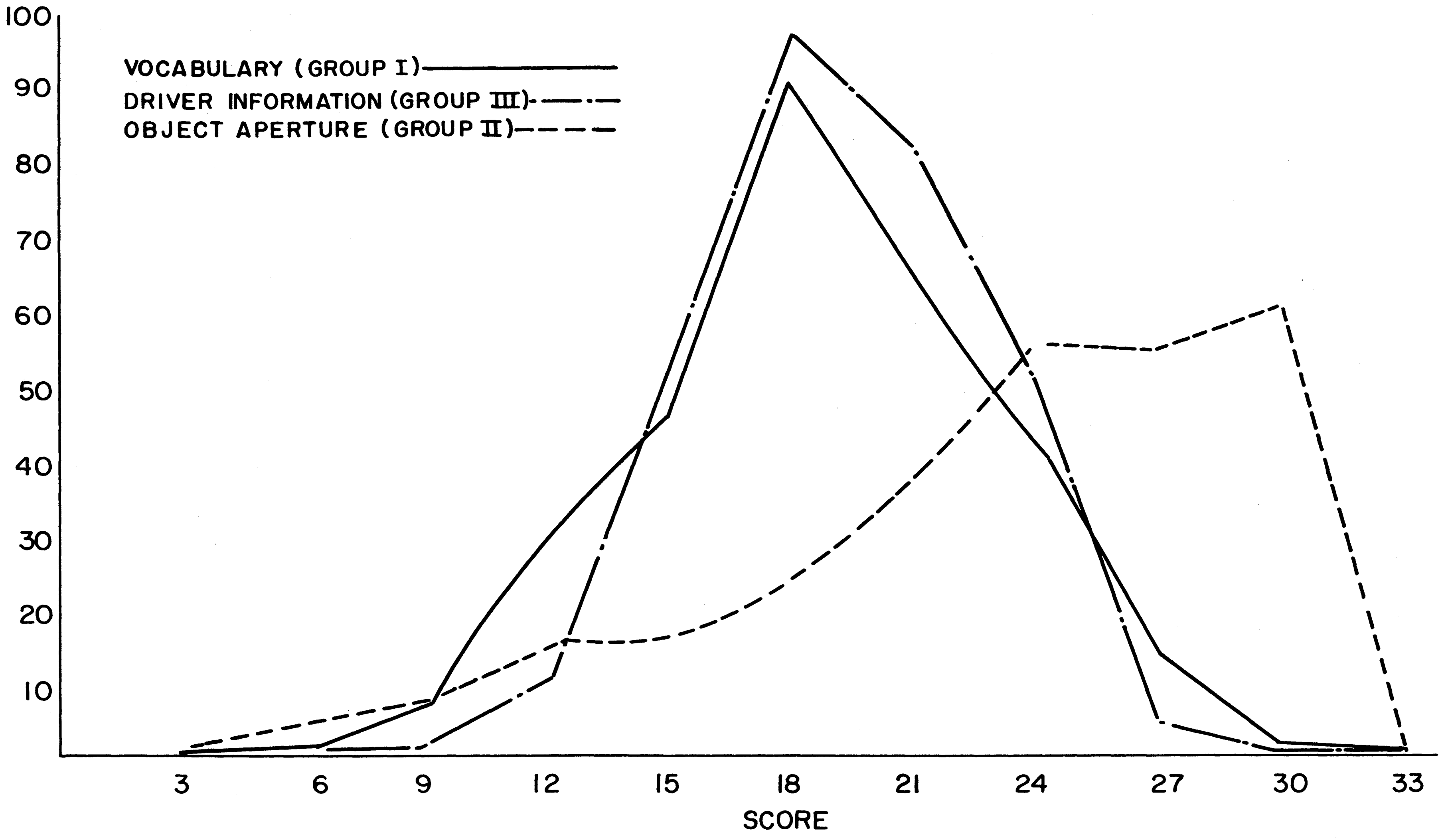


Fig. 1. Distributions of Conventional Method Scores, C Administration.

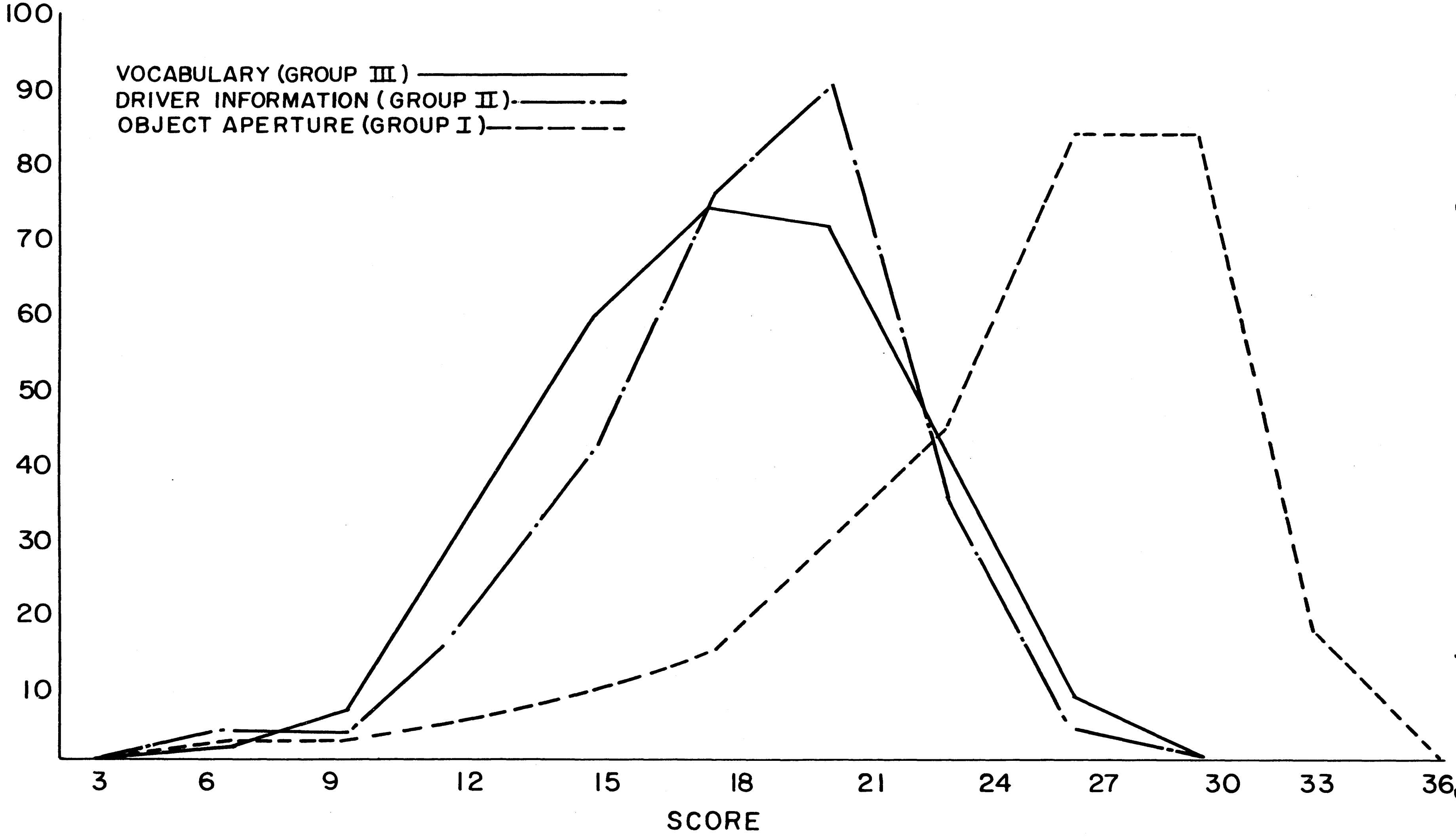


Fig. 2. Distributions of Conventional Method Scores, B Administration.

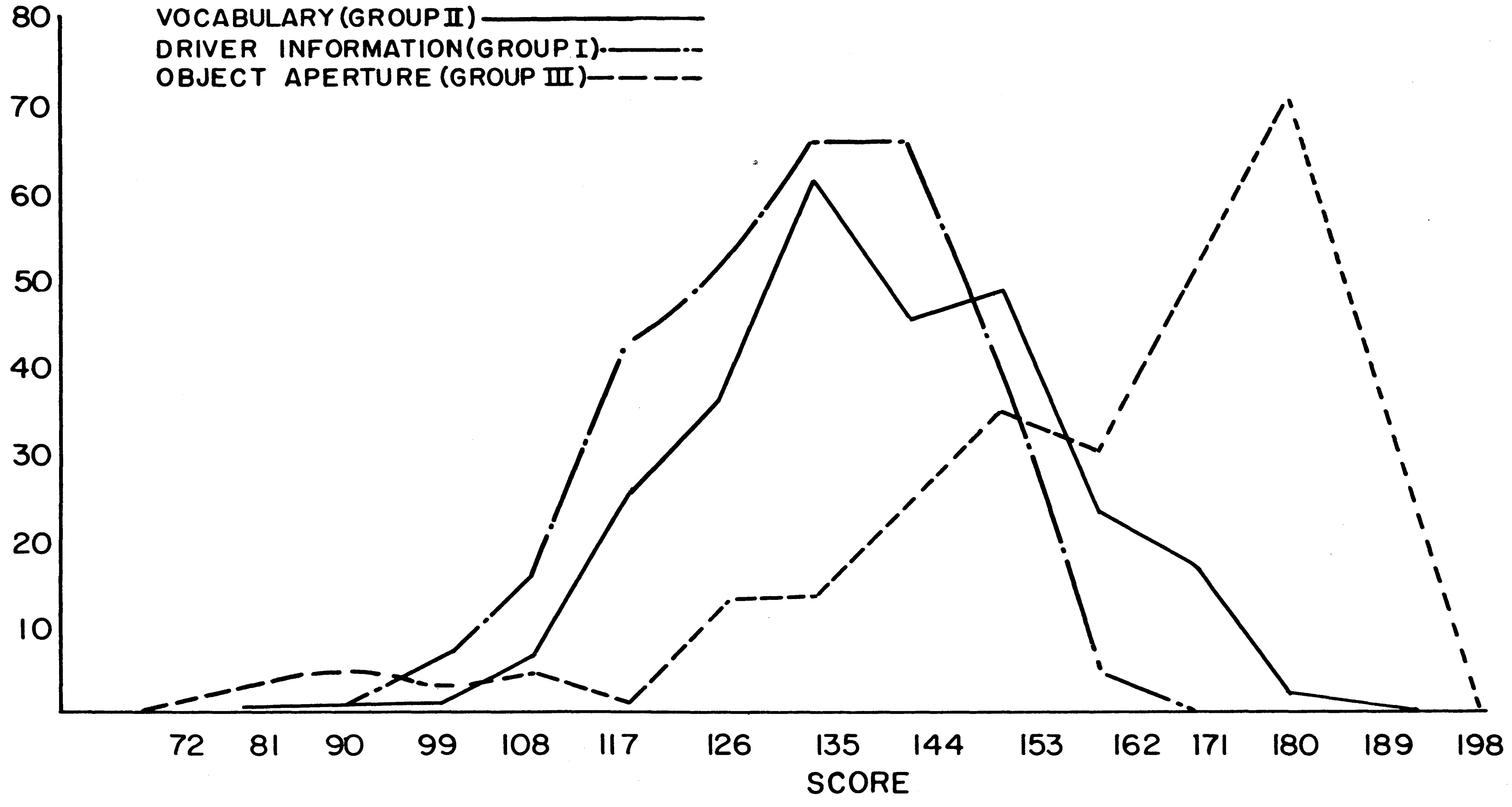


Fig. 3. Distributions of Experimental Method Scores, E Administration.

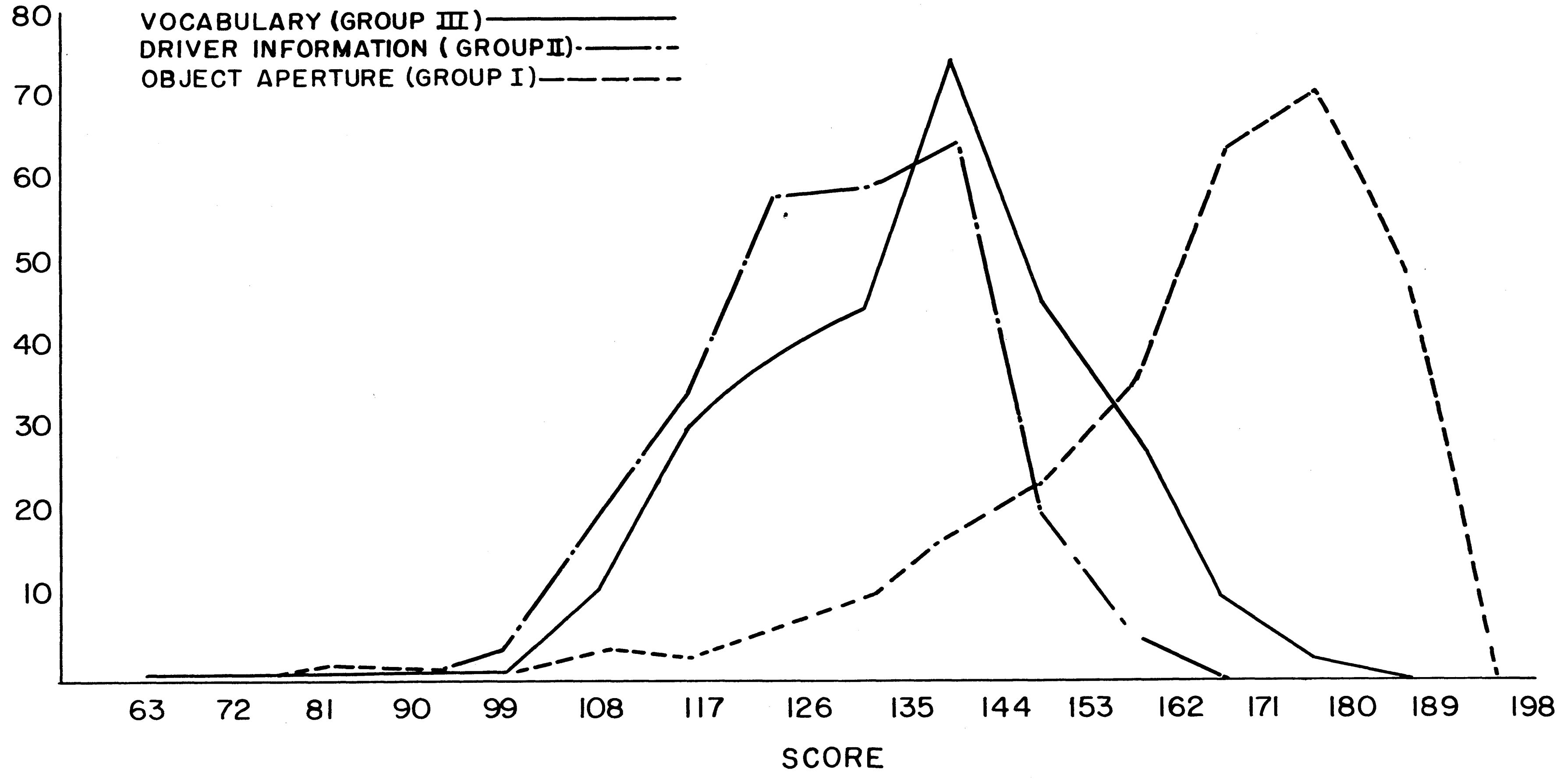


Fig. 4. Distributions of Experimental Method Scores, B Administration

CHAPTER III. EVIDENCE FOR PARTIAL INFORMATION AND EFFECT OF CORRECTING FOR CHANCE

A. EVIDENCE FOR PARTIAL INFORMATION

An assumption underlying this investigation was that partial information exists and enters into answering multiple-choice items. The presence of partial information has been assumed for many years. This study offers an opportunity to test that assumption.

Consider the conventional method of correcting for guessing in speeded power tests. The formula usually used is

$$S_i = R_i - \frac{W_i}{k-1} ,$$

where:

S_i = an individual's score corrected for guessing,
R_i = number right,
W_i = number wrong, and
k = number of alternatives.

This formulation assumes that an individual either knows the answer or guesses, that there is neither partial information nor misinformation. If there were neither partial information nor misinformation, and there were some way of telling, on those items an individual missed, what his second choice for the right answer would be, he would be expected to get a chance proportion, 1/k-1, of them correct.

If partial information exists and is operative, there would be a disproportionate number of individuals getting more than 1/k-1 of these items correct on their second choice. If misinformation exists and is operative, there would be, independently of the preceding hypothesis, a disproportionate number of individuals getting less than 1/k-1 correct.

Essentially the chance hypothesis says that there is a binomial distribution $(p+q)^n$ where $p = 1/k-1$, the probability of getting an item right on second choice when it was missed on the first choice and n is the number of items missed on the first choice. The hypothesized existence of partial information and misinformation says that the obtained distribution should exceed the chance distribution in both tails. More individuals would get more than the mean number of items right expected by chance by virtue of partial information and more individuals would get less than the mean number of items right expected by chance by virtue of misinformation. Data collected by the both (B) method provide a possible basis for testing these hypotheses. In this method the subject was instructed to rank three alternatives as being incorrect in order from one, the one he was most certain was incorrect, to three, the one he was least certain was incorrect. The fourth alternative he left unmarked as his selection for the right alternative. It was assumed then that the alternative ranked third by an individual on an item would have been his second choice for the right answer. Then looking only at those items an individual missed on his first choice, the proportion of these items he got right on his second choice could be obtained.

An evaluation of this was made from the B method data for each of the three tests separately. Only those individuals who missed at least ten items on a particular test were used in order to provide a minimum stability for the estimate of the proportion correct on second choice. For each individual who missed ten or more items, a count was made over those items of the number of times he gave the answer a rank of three.

Unfortunately, these data could not be used to test for the existence of both partial information and misinformation because of the varying number of items missed by different individuals. The hypothesis that could be tested was whether partial information was operative as against the alternative hypothesis that either chance or misinformation was operative.

Another pair of alternative hypotheses is the hypothesis that misinformation was operative as against the hypothesis that chance or partial information was operative. These two pairs of alternative hypotheses are not independent as they both depend on the relative influence of partial information vs misinformation. The pair chosen to be tested was that of partial information vs either chance or misinformation. This pair was chosen on the expectation that the influence of partial information would exceed that of misinformation.

The hypothesis of the existence of partial information is sustained vs its alternative if the number of individuals having more than one third of their second choices correct significantly exceeds the number of individuals having exactly one third or less of their second choices correct. This is a very conservative test. The sign test was used to determine whether a significant number of examinees, over each test separately, supported the assumption of partial information.

There were 248 examinees who missed at least ten items on the Vocabulary Test. Of these 248, 202 gave the answer a rank of three more often than one would predict by chance. This is significant at less than the 1-percent level, indicating that the assumption of partial information is sustained for the Vocabulary Test.

There were 114 examinees who missed at least ten items on the Driver Information Test. Of these 114, 86 gave the answer a rank of three more often than one would predict by chance. This, also, is significant at less than the 1-percent level, which supports the assumption of partial information for the Driver Information Test.

There were 61 examinees who missed at least ten items on the Object Aperture Test. Of these 61, 36 gave the answer a rank of three more often than one would predict by chance. This is not a significant difference. The assumption of partial information is not supported for the Object Aperture Test.

These results demonstrate that partial information does operate, in certain test situations, in the selection of responses to multiple-choice test items. Since it can operate, its measurement would contribute to differentiation between individuals.

Having sustained the hypothesis that partial information does exist, it was further hypothesized that it would be related to complete information. Examinees who know the most answers probably have more partial information about the items they miss than do those examinees who know the fewest answers. To test this hypothesis, product moment correlations were obtained between the C score (total number right) on the B method data and the percent of times the answers to the missed items were ranked three rather than one or two. The same subsamples were used as were used to determine whether partial information exists. These subsamples represented a severely restricted range of ability on each particular test, since all examinees missing nine or fewer items were excluded. Therefore, the variances of the test scores for these subsamples were computed for comparison with the total test variances of the entire groups of examinees in order to estimate the correlations for these groups. Table 10 presents these results.

For both the Vocabulary and Driver Information Tests the correlation coefficients are positive and significantly different from zero. In both cases the test score variances show considerable curtailment. Estimates of the correlation for the total group are contained in the column headed "R". The result for the Object Aperture Test did not produce a significant correlation. This is not surprising, however, since the presence of partial information was not established for this test.

TABLE 10

PRODUCT MOMENT CORRELATIONS BETWEEN TEST SCORES AND PERCENT OF ANSWERS TO MISSED ITEMS GIVEN A RANK OF THREE

Test	N	r	σ_s^2 *	σ_t^2 **	R†
Vocabulary	248	.425	12.34	17.66	.490
Driver Information	112	.354	6.82	13.71	.473
Object Aperture	61	.035	15.24	24.94	.045

*Variance of subsample's test scores.

**Variance of total group's test scores.

†r corrected for curtailment. (Gulliksen,[6] p. 137, Equation 18)

The results indicate that examinees with less than complete information on a given subject may have considerable partial information and that this may be used as a valid basis for discrimination among them.

B. EFFECT OF CORRECTING FOR CHANCE

This evidence for the existence of partial information raises the question of what meaning the conventional correction for chance has.[8] Some indication is given by the following development.

Let:

x = a continuous variable defined from 0 to ∞ representing ability;

$f(x)$ = the probability density function of items over $0 \leqslant x \leqslant \infty$;

$p_i(x)dx$ = the probability of an individual i getting an item right of difficulty between x and x+dx, on the basis of ability alone;

n = number of items;

k = number of alternatives in each item;

R_i = number of items right for individual i; and

W_i = number of items wrong for individual i.

The items in the test are multiple-choice items, ordered in difficulty, and the individual takes them in succession without skipping any.

The function $p_i(x)$ attempts to capture the idea of partial information as distinct from the two-valued function in which an individual either knows the answer $[p_i(x) = 1]$ or guesses $[p_i(x) = 0]$.

Let $k = 1/c$. Then the probability of an individual getting an item right in the interval between x and x+dx is

$$p_i'(x)dx = p_i(x)dx + c[1-p_i(x)dx]. \tag{1}$$

The actual number of items an individual would get right in a given test with item distribution f(x) is

$$R_i = n\int_0^\infty \left\{p_i(x) + c[1-p_i(x)]\right\} f(x)dx, \tag{2}$$

which upon expansion becomes

$$R_i = n\int_0^\infty p_i(x)f(x)dx + cn\int_0^\infty f(x)dx - cn\int_0^\infty p_i(x)f(x)dx . \tag{3}$$

Let

$$T_i = n\int_0^\infty p_i(x)f(x)dx , \tag{4}$$

where T_i is interpreted as the individual's true number of items right or true score on the test. Also,

$$\int_0^\infty f(x)dx = 1 . \tag{5}$$

Substituting (4) and (5) in (3),

$$R_i = (1-c)\ T_i + cn . \tag{6}$$

Solving for T_i,

$$T_i = \frac{R_i - cn}{1-c} \quad . \tag{7}$$

For a power test,

$$R_i + W_i = n \quad . \tag{8}$$

Substituting (8) in (7),

$$T_i = R_i - \frac{c}{1-c} W_i \ , \tag{9}$$

which is the conventional formula for correcting guessing. It follows then that in a power test the corrected score is an estimate of the individual's true score on a test.

Of special interest is the speeded power test, which is the same test as before but administered with a time limit so that not everyone finishes, i.e., $R_i + W_i \leqslant n$. In this case the number of items the individual will get right is given by a modification of equation (3) as follows:

$$R_i^{(1)} = n\int_0^{X_i} p_i(x)f(x)dx + cn\int_0^{X_i} f(x)dx - cn\int_0^{X_i} p_i(x)f(x)dx \ , \tag{10}$$

where X_i is the level of difficulty in the test reached by the individual. Equation (4) becomes

$$T_i = n\int_0^{X_i} p_i(x)f(x)dx + n\int_{X_i}^{\infty} p_i(x)f(x)dx \ . \tag{11}$$

Let

$$T_i^{(1)} = n\int_0^{X_i} p_i(x)f(x)dx \ , \tag{12}$$

$$T_i^{(2)} = n\int_{X_i}^{\infty} p_i(x)f(x)dx \ , \tag{13}$$

where $T_i^{(1)}$ is the individual's true score on that part of the test he attempted.

The number of items the individual attempted is given by

$$n_i = n \int_0^{X_i} f(x)dx \ . \qquad (14)$$

Substituting (12) and (14) in (10),

$$R_i^{(1)} = (1-c)T_i^{(1)} + cn_i \quad , \qquad (15)$$

where n_i represents the number of items the individual attempted in the test and $T_i^{(1)}$ his true score on that segment of the test.

Solving equation (15) for $T_i^{(1)}$ where $R_i^{(1)} + W_i^{(1)} = n_i$ and rearranging terms gives

$$T_i^{(1)} = R_i^{(1)} - \frac{c}{1-c} W_i^{(1)} \quad . \qquad (16)$$

Thus, correcting scores on a speeded power test for chance yields an estimate of the true score of the individual on that segment of the test he finished. But, as $T_i^{(1)}$ may be based on a different number of items for different individuals depending on the speed at which they work, this score does not represent an individual's power ability in the sense that T_i does, unless speed and power of performance are functionally related. This latter is an experimental question that still remains to be solved.

It should be noted that the above results are independent of the specific functions assumed for $p_i(x)$ and for $f(x)$ provided that certain analytic conditions are satisfied, e.g., the existence and absolute convergence of the integrals considered.

CHAPTER IV. COMPARATIVE RELIABILITIES AND VALIDITIES

A. COMPARATIVE RELIABILITIES

There are two grounds for an expectation that administering a set of items by the experimental method would result in improved reliability. First, scores obtained by the conventional method contain a chance, or error, component which, it was hoped, would not be contained in the scores obtained by the E method. Second, since each subject is making more responses by the E method than by the conventional, the E method may have the effect of lengthening the test somewhat.

These grounds are based on the formal concept that reliability varies inversely with the relative proportion of error variance in the total variance. When, on the other hand, one considers that such measures of reliability as the Kuder-Richardson formulas are to a large extent a reflection of the homogeneity of a test, then in that sense the experimental method should not differ in reliability from the conventional.

As a means of comparing the reliabilities obtained by the two response methods on the same test, the view of the E method as a "lengthening" of the conventional method suggested setting up an index which we called the "coefficient of effective length" (CEL).* In this coefficient the E method is regarded as the "lengthened" form of the conventional method and the CEL is the "k" in the Spearman-Brown prophecy formula for the reliability of a test lengthened k times. Thus,

$$\text{CEL} = k = \frac{r_{kk}(1-r_{11})}{r_{11}(1-r_{kk})} ,$$

where r_{kk} is the reliability of the test administered by the experimental method and r_{11}, the reliability of the test as conventionally administered.

The CEL should be interpreted as the length of the test under the experimental method in units of the length of the test under the conventional method. The CEL may take on any non-negative value. A CEL = 1 signifies that

*Philip DuBois suggested this name for the index.

the test under the experimental method had effectively the same length as when administered under the conventional method in so far as reliability is concerned, i.e., a CEL = 1 indicates that the test had the same reliability under the two methods of administration. Another way of looking at the index is that the CEL is a measure of how much a test administered in the conventional manner would have to be lengthened in order to produce the same reliability as that same test (not lengthened) administered by the experimental method.

The reliability estimate used throughout was the well-known Kuder-Richardson Formula 20.[7] Reliability comparisons for the conventional and experimental methods appear in Table 11. Since, in every case, the groups taking a test by the C method consisted of different individuals from those taking it by the E method there was no reason for pairing one item analysis group, A or B, with any particular one of the others. Accordingly, the CEL shown in the table for these groups is the mean of the four possible coefficients in each case.

TABLE 11

RELIABILITY ESTIMATES (K-R #20) AND COEFFICIENTS OF EFFECTIVE LENGTH FOR CONVENTIONAL AND EXPERIMENTAL METHODS

	Vocabulary			Driver Information			Object Aperture		
	A	B	Total	A	B	Total	A	B	Total
Reliability									
C method	.72	.72	.72	.64	.63	.64	.89	.88	.89
E method	.71	.75	.73	.73	.66	.70	.92	.89	.91
Coefficient of effective length	1.06*		1.05	1.32*		1.31	1.25*		1.25

*Mean of the four intergroup coefficients.

The CEL's are all larger than unity, but the magnitude of the increment in reliability is not spectacular. The CEL's ranged from 1.05 to 1.31 with an average value of 1.20. This means that on the average the use of the experimental method had the effect of increasing the reliability of the test equivalent to a 20-percent increase in the length of the test. This should not be interpreted, on the basis of these data, as a characteristic parameter of the method. The several tests over which the mean index was taken were not equivalent in number of items nor in administration time. They were all power tests. A proper estimate of an expected CEL should be based on administration time and might be different for different content areas and levels of difficulty.

Three additional questions of reliability were investigated. The first had to do with the possible influence of test difficulty upon the relation between reliability and method of responding. The procedure was to construct two subtests from each test, one consisting of ten easy items and one consisting of ten difficult items. The CEL was then computed for each comparison.

In the mechanics of constituting the tests, three controls were employed. The subtests were matched as closely as possible on the discrimination indices of the component items; a separate pair of tests was made from the data from each item analysis group; and separate sets of tests were constituted on the basis of difficulty values obtained from data on the C and E methods. There were thus twelve ten-item subtests: three tests x two item analysis groups x two methods of administration. The mean C and E method scores made on these tests are shown in Table 12.

TABLE 12

MEANS OF TEN-ITEM CONSTITUTED TESTS

		Vocabulary		Driver Information		Object Aperture	
		A	B	A	B	A	B
			C Method*				
Picked by	Easy	7.93	7.62	8.14	8.48	8.79	8.69
C Method	Difficult	3.56	3.27	4.32	4.25	6.23	5.81
Picked by	Easy	7.64	7.25	7.92	8.39	8.75	8.48
E Method	Difficult	3.14	3.01	4.35	4.22	5.54	5.69
			E Method**				
Picked by	Easy	52.36	50.48	52.69	54.34	55.04	56.22
C Method	Difficult	35.88	33.99	37.53	38.68	46.36	46.56
Picked by	Easy	51.55	48.99	52.30	53.67	55.05	55.80
E Method	Difficult	33.62	32.79	37.00	38.78	43.73	45.51

*Maximum score is 10.
**Maximum score is 60.

The relations between difficulty, response method, and reliability are shown in Table 13. In this table, the coefficients for item analysis groups A are based on ten-item tests whose items were chosen on the basis of their performances with B groups, and vice versa. In all but two cases (C method data, Driver Information, A group; and E method data, Vocabulary, B group) the CEL's were larger for the difficult tests. By the sign test, a 10-2 split is significant at the 5-percent level. Making the comparison another way, eleven of the twelve CEL's for difficult tests are greater than unity (significant at the 1-percent level by the sign test), whereas only seven of those for the easy tests are greater than unity. There seems to be some support, then, for the hypothesis that the experimental method is more likely to result in improved reliability with difficult rather than with easy tests.

TABLE 13

K-R #20'S AND CEL'S FOR CONSTITUTED EASY AND DIFFICULT TEN-ITEM TESTS

Test	Group	C Score (K-R)		E Score (K-R)		CEL	
		Easy	Dif.	Easy	Dif.	Easy	Dif.
		Items Selected on C Method Data					
V	A	.532	.434	.365	.507	0.506	1.341
	B	.500	.485	.589	.589	1.433	1.522
DI	A	.415	.386	.578	.528	1.931	1.779
	B	.388	.428	.249	.445	0.523	1.072
OA	A	.830	.770	.828	.852	0.986	1.720
	B	.744	.717	.745	.751	1.005	1.190
		Items Selected on E Method Data					
V	A	.506	.619	.415	.669	0.693	1.244
	B	.435	.538	.592	.559	1.885	1.089
DI	A	.431	.388	.552	.547	1.627	1.905
	B	.369	.459	.312	.421	0.775	0.857
OA	A	.769	.719	.819	.793	1.359	1.497
	B	.684	.740	.723	.781	1.206	1.253

The second additional reliability investigation also dealt with difficulty, but in this case the samples were drawn from pools of individuals rather than from items. Each item analysis group was divided at the median score on each test: for the group above the median, a test was considered an easy test; for the group below the median, that same test was considered a difficult test. The results for the groups of high and low scorers appear in Table 14. Again there is a tendency for high CEL to be associated with greater difficulty. The presence of negative reliability coefficients, however, detracts from the clarity of this distinction. Such coefficients may be explained by item heterogeneity, but it seems more likely that they represent sampling fluctuations.[2]

TABLE 14

K-R #20'S AND CEL'S FOR HIGH-SCORING AND LOW-SCORING EXAMINEES

Test	Group	High Scorers			Low Scorers		
		K-R		CEL	K-R		CEL
		C Score	E Score		C Score	E Score	
V	A	.378	.337	0.836	.182	-.155	---
	B	.112	.232	2.395	.245	.407	2.115
DI	A	-.201	.179	---	.258	.304	1.256
	B	.005	-.265	---	-.407	-.092	---
OA	A	.402	.381	0.916	.802	.866	1.596
	B	.340	.251	0.651	.762	.806	1.298

The third attack upon the question of reliability was made by computing reliability coefficients based on only those individuals who had used the experimental method at least once, i.e., all persons who when using the E method had crossed out three alternatives on every item were excluded. The effect of this procedure on the mean scores is shown in Table 15.

TABLE 15

MEANS OF THE ORIGINAL GROUPS AND THE GROUPS REDUCED BY ELIMINATING EXAMINEES NOT USING THE EXPERIMENTAL METHOD

Test	Group	Original		Reduced	
		N	M	N	M
V	A	135	141.41	106	141.64
	B	134	139.52	104	142.23
DI	A	146	132.15	115	133.93
	B	146	134.58	119	135.50
OA	A	147	160.86	97	156.48
	B	147	164.22	85	158.82

The results of this aspect of the study are presented in Table 16. Four of the six CEL's are greater than unity, and likewise four are greater than the corresponding CEL's for the groups before the exclusion of individuals not making use of the experimental method. On the basis of these comparisons, it can hardly be said that the restriction to persons making use of the method had any significant effect on the CEL.

TABLE 16

K-R #20'S AND CEL'S COMPUTED FOR THOSE EXAMINEES WHO USED THE EXPERIMENTAL METHOD AT LEAST ONCE—WHO MARKED TWO OR FEWER RESPONSES ON AT LEAST ONE ITEM ON THE EXPERIMENTAL METHOD

Test	Group	N	K-R Computed	K-R* Adjusted	K-R C Method	CEL
V	A	106	.702	.720	.723	0.985
	B	104	.679	.761	.720	1.238
DI	A	115	.725	.751	.642	1.682
	B	119	.687	.675	.634	1.199
OA	A	97	.899	.913	.891	1.284
	B	85	.888	.877	.884	0.936

*Estimated K-R for group with variance equal to original group.

$$r_{22} = 1 - \frac{1-r_{11}}{k}, \quad \text{where } k = \frac{\sigma_2^2}{\sigma_1^2}$$

B. COMPARATIVE VALIDITIES

Tables 17, 18, and 19 present the validity coefficients for the three tests used in this project when the sixteen reference variable scores are used as criteria. They were developed for A and B data separately and together. Table 20 presents the number of times the corresponding coefficients are higher for one method or the other.

It had been hypothesized that validity would not be affected by the use of the E method, except as it might be related to changed reliability. When considering all three tests for A and B groups together, twenty-four of the forty-eight comparisons are larger for C method scores and twenty-four are larger for E method scores. When considering the tests separately the E method scores seem to do a better job for the Driver Information Test and the C method scores for the Object Aperture Test. These differences are not significant in terms of the sign test.

TABLE 17

VALIDITY COEFFICIENTS: C AND E SCORES VS SIXTEEN REFERENCE VARIABLES (A GROUPS)

Variable	Vocabulary				Driver Information				Object Aperture			
	C		E		C		E		C		E	
Differential Aptitude Tests												
Verbal	.706	(57)*	.661	(73)	.387	(89)	.492	(57)	.517	(73)	.470	(89)
Numerical	.399	(58)	.365	(79)	.183	(87)	.263	(58)	.355	(79)	.438	(87)
Abstract	.320	(59)	.521	(79)	.448	(87)	.324	(59)	.535	(79)	.657	(87)
Space	.231	(58)	.417	(79)	.465	(90)	.354	(58)	.666	(79)	.698	(90)
Mechanical	.377	(61)	.312	(77)	.545	(87)	.625	(61)	.647	(77)	.742	(87)
Clerical	.140	(58)	.174	(80)	-.267	(88)	.035	(58)	.055	(80)	-.084	(88)
Spelling	.621	(58)	.434	(68)	.000	(87)	.482	(58)	.184	(68)	.028	(87)
Sentences	.768	(61)	.603	(74)	.175	(88)	.497	(61)	.320	(74)	.205	(88)
Stanford-Binet IQ	.371	(82)	.507	(84)	.406	(84)	.343	(82)	.494	(84)	.440	(84)
California Intelligence Test												
Total IQ	.607	(120)	.547	(101)	.381	(116)	.533	(120)	.428	(101)	.404	(116)
Language IQ	.651	(120)	.514	(101)	.261	(116)	.480	(120)	.330	(101)	.324	(116)
Nonlanguage IQ	.284	(120)	.354	(101)	.324	(116)	.408	(120)	.443	(101)	.395	(116)
MacQuarrie Mechanical Aptitude Test	.287	(92)	.307	(87)	.295	(86)	.416	(92)	.413	(87)	.412	(86)
Detroit General Aptitude Test												
General IQ	.606	(76)	.635	(42)	.423	(47)	.460	(76)	.444	(42)	.407	(47)
Mechanical IQ	.538	(76)	.486	(42)	.467	(47)	.446	(76)	.373	(42)	.345	(47)
Clerical IQ	.491	(76)	.442	(42)	.202	(47)	.307	(76)	.079	(42)	.337	(47)

*N's are in parentheses.

TABLE 18

VALIDITY COEFFICIENTS: C AND E SCORES VS SIXTEEN REFERENCE VARIABLES (B GROUPS)

Variable	Vocabulary				Driver Information				Object Aperture			
	C		E		C		E		C		E	
Differential Aptitude Tests												
Verbal	.661	(63)*	.717	(76)	.586	(87)	.527	(63)	.614	(76)	.520	(87)
Numerical	.481	(62)	.374	(76)	.357	(86)	.635	(62)	.441	(76)	.434	(86)
Abstract	.452	(63)	.437	(76)	.425	(87)	.548	(63)	.613	(76)	.603	(87)
Space	.395	(58)	.510	(75)	.562	(86)	.512	(58)	.677	(75)	.729	(86)
Mechanical	.241	(61)	.432	(75)	.657	(86)	.557	(61)	.673	(75)	.734	(86)
Clerical	.233	(63)	.167	(78)	.153	(86)	.219	(63)	.046	(78)	-.095	(86)
Spelling	.545	(62)	.465	(69)	.077	(86)	.172	(62)	.126	(69)	.254	(86)
Sentences	.607	(64)	.593	(79)	.407	(82)	.382	(64)	.311	(79)	.341	(82)
Stanford-Binet IQ	.435	(87)	.524	(76)	.235	(77)	.379	(87)	.504	(76)	.320	(77)
California Intelligence Test												
Total IQ	.700	(110)	.597	(98)	.430	(105)	.480	(110)	.663	(98)	.455	(105)
Language IQ	.738	(110)	.619	(98)	.347	(105)	.423	(110)	.527	(98)	.306	(105)
Nonlanguage IQ	.379	(110)	.369	(98)	.416	(105)	.351	(110)	.664	(98)	.536	(105)
MacQuarrie Mechanical Aptitude Test	.275	(100)	.257	(86)	.266	(76)	.351	(100)	.523	(86)	.461	(76)
Detroit General Aptitude Test												
General IQ	.629	(85)	.581	(39)	.420	(45)	.383	(85)	.519	(39)	.325	(45)
Mechanical IQ	.482	(85)	.225	(39)	.459	(45)	.372	(85)	.510	(39)	.368	(45)
Clerical IQ	.487	(85)	.472	(39)	.309	(45)	.232	(85)	.325	(39)	.342	(45)

*N's are in parentheses.

TABLE 19

VALIDITY COEFFICIENTS: C AND E SCORES VS SIXTEEN REFERENCE VARIABLES (TOTAL GROUPS)

Variable	Vocabulary				Driver Information				Object Aperture			
	C		E		C		E		C		E	
Differential Aptitude Tests												
Verbal	.679	(120)*	.690	(149)	.496	(176)	.508	(120)	.565	(149)	.485	(176)
Numerical	.438	(120)	.364	(155)	.265	(173)	.420	(120)	.395	(155)	.432	(173)
Abstract	.388	(122)	.481	(155)	.431	(174)	.420	(122)	.575	(155)	.638	(174)
Space	.313	(116)	.470	(154)	.514	(176)	.422	(116)	.672	(154)	.710	(176)
Mechanical	.310	(122)	.379	(152)	.602	(173)	.596	(122)	.660	(152)	.729	(173)
Clerical	.187	(121)	.174	(158)	-.210	(174)	.127	(121)	.051	(158)	-.086	(174)
Spelling	.585	(120	.456	(137)	.041	(173)	.335	(120)	.156	(137)	.127	(173)
Sentences	.690	(125)	.601	(153)	.299	(170)	.432	(125)	.318	(153)	.271	(170)
Stanford-Binet IQ	.404	(169)	.504	(160)	.313	(161)	.355	(169)	.496	(160)	.372	(161)
California Intelligence Test												
Total IQ	.657	(230)	.579	(199)	.371	(221)	.500	(230)	.556	(199)	.426	(221)
Language IQ	.697	(230)	.573	(199)	.301	(221)	.468	(230)	.433	(199)	.313	(221)
Nonlanguage IQ	.335	(230)	.373	(199)	.365	(221)	.376	(230)	.566	(199)	.459	(221)
MacQuarrie Mechanical Aptitude Test	.280	(192)	.281	(173)	.285	(162)	.382	(192)	.471	(173)	.431	(162)
Detroit General Aptitude Test												
General IQ	.616	(161)	.604	(81)	.421	(92)	.422	(161)	.474	(81)	.369	(92)
Mechanical IQ	.508	(161)	.440	(81)	.463	(92)	.407	(161)	.437	(81)	.351	(92)
Clerical IQ	.488	(161)	.457	(81)	.245	(92)	.269	(161)	.196	(81)	.335	(92)

*N's are in parentheses.

TABLE 20

NUMBER OF TIMES THE SIXTEEN VALIDITY COEFFICIENTS
ARE HIGHER FOR EACH TYPE OF SCORE

Group	Method	No. of Times Larger
	Vocabulary	
A	C	9
	E	7
B	C	12
	E	4
	Driver Information	
A	C	4
	E	12
B	C	8
	E	8
	Object Aperture	
A	C	11
	E	5
B	C	11
	E	5
Test	**Method**	**No. of Times Larger**
	Total	
V	C	9
	E	7
DI	C	4
	E	12
OA	C	11
	E	5

The hypothesis is thus borne out that the experimental method does not appear to differ significantly in what it measures from the conventional method.

Incidentally, the tests used in this study relate most highly to those reference variables which one would expect. The Vocabulary Test is most highly related to the "Verbal" and "Sentences" scores of the Differential Aptitude Tests and to the "Language IQ" of the California Intelligence Test. The Driver Information Test is most highly related to the "Mechanical" score of the Differential Aptitude Tests. The Object Aperture Test is most highly related to the "Space" and "Mechanical" scores of the Differential Aptitude Tests.

CHAPTER V. COMPARATIVE METHODS OF ITEM ANALYSIS

The experimental response method has a seven-point scale on each item as contrasted with the two-point scale of the conventional method. Also, the E method provides information not previously available and which may be of interest, e.g., the difficulty of a distracter measured in terms of its being recognized as wrong instead of being measured in terms of its being selected as the answer. Furthermore, the difficulty of the answer can be measured separately in the E method from the difficulty of the item as a whole. For reasons such as these, the E method was evaluated in terms of its possible contribution to item analysis techniques.

A. INDICES USED IN THIS STUDY

1. Conventional Response Method.—The index of item difficulty chosen was p, the percent passing. Both point biserial r and biserial r were computed as indices of item discrimination. These are the most common indices used in conventional item analysis. A short-cut discrimination index, $p_u - p_l$, where p_u is the proportion of a high-scoring group which passes an item, p_l the proportion of a low-scoring group, was also computed.

2. Experimental Response Method.—Mean score on the item was used as the index of item difficulty. This index parallels p for the conventional method. The p's for the distracters were also computed, as well as an index of answer difficulty defined by:

$$d = N_3 + (1/2)N_2 + (1/3)N_1 + (1/4)N_0 ,$$

where

N_3 = the number of examinees marking the three distracters,
N_2 = the number marking two distracters only,
N_1 = the number marking one distracter only, and
N_0 = the number omitting the item.

This index of answer difficulty is relative to the difficulties of the distracters associated with it in the item.

Item-test product moment r, with Sheppard's correction, was selected as the index of item discrimination. This is comparable to point biserial for the conventional method. A serial correlation for seven categories could also have been used to parallel the use of biserial. Ferguson's delta,[4] proposed as an index of test discrimination (in a different sense from item discrimination), was computed as an item index. A short-cut discrimination index, $M_u - M_l$, where M_u is the mean score on an item from an upper group, and M_l from a lower group, was also computed.

B. GENERAL EFFECTIVENESS OF CONVENTIONAL VS MODIFIED TECHNIQUES

General effectiveness here means doing the job of choosing items for tests. The comparisons which were made stem from the design for comparing the test reliabilities produced by the two response methods. Easy and difficult tests of ten items each, matched on item discrimination, were constituted using conventional and experimental method data independently for each of the six-item selection groups. Each of these twenty-four tests was scored on a cross-validation group (e.g., item analysis groups B served as cross-validation groups for items selected on A group data, and vice versa) for both C and E data. For each of these forty-eight tests a Kuder-Richardson#20 reliability estimate was computed. Table 21 presents these estimates.

TABLE 21

K-R #20 RELIABILITY ESTIMATES FOR EASY AND DIFFICULT TEN-ITEM TESTS

Group		Test	C Score Easy		C Score Difficult		E Score Easy		E Score Difficult	
C	E		C*	E**	C*	E**	C*	E**	C*	E**
I-A	II-A	V	.532	.506	.434	.619	.365	.415	.507	.669
I-B	II-B	V	.500	.435	.485	.538	.589	.592	.589	.559
III-A	I-A	DI	.415	.431	.386	.388	.578	.552	.528	.547
III-B	I-B	DI	.388	.369	.428	.459	.249	.312	.445	.421
II-A	III-A	OA	.830	.769	.770	.719	.828	.819	.852	.793
II-B	III-B	OA	.744	.684	.717	.740	.745	.723	.751	.781

*Items selected on C method data.

**Items selected on E method data.

Of the twenty-four comparisons, twelve are higher for items selected on C method data and twelve for items selected on E method data. This overall comparison contains no evidence favoring selection indices of either method. In the case of C score, easy items, five of the six comparisons favor C method selection. In the case of C score, difficult items, five of the six comparisons favor E method selection. This possibly indicates that C method selection indices do a better job on easy items and E method selection indices do a better job on difficult items, but this conclusion is not substantiated when comparing E scores.

It should be kept in mind that efforts were made to equate discrimination indices. The question was whether any differences in reliability would appear on cross-validation data when no effort had been made to build it in.

C. RELIABILITY OF CONVENTIONAL AND OF MODIFIED INDICES

To measure the reliability of item difficulty indices for the conventional method, Kendall's tau between the p's obtained in the two groups was used; for the experimental method, taus between item means and between answer difficulties were used. Product moment r was used as a measure of reliability for all difficulty indices. Table 22 presents the reliability estimates for the item difficulty indices computed for each of the response methods.

All of the product moment r's between indices computed on comparable groups were at least 0.97. There were no differences between the reliability estimates of difficulty indices for the conventional response method and those for the experimental method.

All of the taus between indices with small N's were equal to or greater than 0.80. Here again there were no differences between C method indices and E method indices. The lower taus for indices from the Object Aperture Test may be accounted for by a greater homogeneity of item difficulty indices for that test.

Tau was used as a measure of the reliability of item discrimination indices, computed separately from A and B group data. Table 23 presents the reliability estimates of the item discrimination indices computed for the two response methods. The taus are small, the largest being 0.66. This confirms the findings of others that larger N's are necessary for securing adequate reliability for these discrimination indices. There is one atypical tau, for biserial on Driver Information. Aside from this one instance, there is no evidence that any of the three indices tested is more reliable than the others.

TABLE 22

RELIABILITY OF ITEM DIFFICULTY INDICES

Index Used	Test	Groups A vs B	N*	r	Tau
		C Method			
Answer	V	I	32	0.99	0.93
difficulty	DI	III	28	0.99	0.90
p	OA	II	32	0.98	0.85
Distracter	V	I	96	0.99	
difficulty	DI	III	84	0.98	
p	OA	II	96	0.97	
		E Method			
Item	V	II	32	0.99	0.93
difficulty	DI	I	28	0.98	0.90
M	OA	III	32	0.98	0.80
Answer	V	II	32	0.99	0.91
difficulty	DI	I	28	0.98	0.91
d**	OA	III	32	0.98	0.82
Distracter	V	II	96	0.99	
difficulty	DI	I	84	0.97	
p	OA	III	96	0.97	

*Number of items.

**$d = N_3 + (1/2)N_2 + (1/3)N_1 + (1/4)N_0$.

TABLE 23

RELIABILITY OF ITEM DISCRIMINATION INDICES

Index Used	Test	Groups A vs B	N*	Tau**
Point	V	I	32	0.66
biserial	DI	III	28	0.33
(C score)	OA	II	32	0.46
Biserial	V	I	32	0.64
(C score)	DI	III	28	0.07
	OA	II	32	0.50
Product	V	II	32	0.54
moment†	DI	I	28	0.38
(E score)	OA	III	32	0.45

*Number of items.

**Tau between rank order of discrimination indices on two-item analysis groups.

†Using Sheppard's correction for broad categories.

D. RELATION OF CONVENTIONAL TO MODIFIED INDICES

The indices from the A data and the B data for one method were related to the indices from the A data and the B data for the other. Taus were used to relate the discrimination indices of the two methods, and product moment r's to relate the difficulty indices. Table 24 presents the relationships of the conventional item and distracter difficulty indices to the modified indices.

TABLE 24

THE RELATIONS OF CONVENTIONAL DIFFICULTY INDICES TO MODIFIED DIFFICULTY INDICES

Group		Test	Item Difficulty*		Distracter Difficulty**	
C	E		N/	r	N//	r
I-A	II-A	V	32	0.98	96	-0.96
I-A	II-B	V	32	0.99	96	-0.97
I-B	II-A	V	32	0.98	96	-0.97
I-B	II-B	V	32	0.98	96	-0.98
III-A	I-A	DI	28	0.98	84	-0.96
III-A	I-B	DI	28	0.98	84	-0.95
III-B	I-A	DI	28	0.98	84	-0.96
III-B	I-B	DI	28	0.98	84	-0.94
II-A	III-A	OA	32	0.98	96	-0.93
II-A	III-B	OA	32	0.96	96	-0.93
II-B	III-A	OA	32	0.97	96	-0.94
II-B	III-B	OA	32	0.97	96	-0.93

*p vs M.

**p vs p.

/Number of items.

//Number of distracters.

The relationships are all very high, another indication that difficulty is stable and easy to measure. The correlations between distracter difficulties are negative because examinees were asked to mark answers in the conventional method and to mark distracters in the experimental method.

Table 25 presents the relationships of the conventional discrimination indices to the modified indices. The relationships of item biserials to item product moments are consistently low, although eight of them are significantly different from zero and all of them are positive. These correlations are not

TABLE 25

THE RELATIONS OF CONVENTIONAL DISCRIMINATION INDICES
TO MODIFIED DISCRIMINATION INDICES

Group		Test	N*	r_{bis} vs r		$r_{p\ bis}$ vs r	
C	E			Tau	Normal Deviate	Tau	Normal Deviate
I-A	II-A	V	32	.298	2.40 (5%)	.552	4.45 (1%)
I-A	II-B	V	32	.270	2.18 (5%)	.512	4.13 (1%)
I-B	II-A	V	32	.379	3.06 (1%)	.593	4.78 (1%)
I-B	II-B	V	32	.286	2.31 (5%)	.460	3.71 (1%)
III-A	I-A	DI	28	.236	1.76 (NS)	.384	2.87 (1%)
III-A	I-B	DI	28	.098	0.73 (NS)	.294	2.19 (5%)
III-B	I-A	DI	28	.206	1.54 (NS)	.413	3.08 (1%)
III-B	I-B	DI	28	.185	1.38 (NS)	.503	3.75 (1%)
II-A	III-A	OA	32	.298	2.40 (5%)	.302	2.44 (5%)
II-A	III-B	OA	32	.343	2.77 (1%)	.532	4.29 (1%)
II-B	III-A	OA	32	.327	2.64 (1%)	.460	3.71 (1%)
II-B	III-B	OA	32	.306	2.47 (5%)	.593	4.78 (1%)

*Number of items.

high even if the unreliabilities of the indices themselves are taken into consideration by correcting for attenuation. As could be anticipated, the relationships of item point biserials to item product moments are consistently larger than for biserials. These correlations are roughly of the same magnitude as the reliability estimates of both point biserial and product moment.

The sizes of these relationships vary considerably between item selection groups. In one instance, Object Aperture, one tau is twice another, a further indication of the unreliability of these discrimination indices. In summary, apparently what constitutes a good discriminating item is the same for both methods.

E. RELATION OF TWO CONVENTIONAL INDICES

Table 26 represents the relationship of point biserial r to biserial r. Two comparisons were made, one involving the indices for all items of a test, the other for those items whose difficulty indices lay between p values of 0.10 and 0.90. This was done because biserial r is particularly unreliable for extreme p values.

TABLE 26

THE RELATION OF POINT BISERIAL r TO BISERIAL r

Group	Test	N*	Tau	N**	Tau**
I-A	V	32	.625	22	.913
I-B	V	32	.702	26	.852
III-A	DI	28	.675	23	.842
III-B	DI	28	.534	22	.740
II-A	OA	32	.605	27	.761
II-B	OA	32	.621	29	.732

*Number of items.

**After eliminating items whose difficulty indices were outside the range p = 0.10 to 0.90.

The taus for the items with restricted range of difficulty are consistently larger than those computed for all items. They are large enough to indicate that the two indices would do approximately the same jobs of item selection if that selection were limited to items with intermediate difficulties.

F. SIZE OF DISCRIMINATION INDICES

The sign test was used to compare the sizes of biserial and point biserial with product moment item discrimination indices. Table 27 presents these relationships.

In each item, selection group biserials were significantly larger than product moments. This is not surprising as biserial estimates of correlation are inflated by the assumption of normality. In only two of the item selection groups are there significant differences between item product moments and point biserials, but if one compares the one hundred and eighty-four pairs of indices over all six groups, in one hundred and two cases the product moment r (E method) is larger than the point biserial r (C method). This difference is significant at the 1-percent level and may be interpreted as an indication that items discriminate better when administered by the experimental method.

TABLE 27

NUMBER OF TIMES r_{bis} AND $r_{p\ bis}$ WERE LARGER THAN PRODUCT MOMENT r, WHEN USED FOR ITEM-TEST CORRELATIONS

Group C	Group E	Test	N*	r_{bis} Larger than r	Significance of Difference**	$r_{p\ bis}$ Larger than r	Significance of Difference**
I-A	II-A	V	32	25	S (1%)	16	NS
I-B	II-B	V	32	24	S (1%)	13	NS
III-A	I-A	DI	28	20	S (5%)	7	S (5%)
III-B	I-B	DI	28	24	S (1%)	14	NS
II-A	III-A	OA	32	28	S (1%)	9	S (5%)
II-B	III-B	OA	32	29	S (1%)	13	NS

*Number of items.

**Using the sign test.

G. A SHORT-CUT ESTIMATE OF ITEM-TEST SCORE RELATIONSHIP

The examinees were split at the median scores for both response methods, and the differences in item means between high and low groups were computed for all items. Table 28 presents the reliability estimates (correlations between item analysis groups) of these item mean differences (D) and also the relationships of D to item point biserials and product moments.

The reliability estimates are fairly constant. The one atypical value may be a chance variation, but, on the other hand, it is consistent with the reliability estimates of point biserial and product moment for the same test. The reliability estimates of this short-cut index are numerically larger than those for either point biserial or product moment on an overall basis and are more consistent over tests.

For two of the tests, Vocabulary and Driver Information, the relationships of item mean differences to point biserial and product moment are large enough to indicate that the short-cut index would be a reasonable substitute for the more precise indices. For Object Aperture the correlations are quite low.

In order to examine these relationships more carefully, scatter plots were made with item mean differences on the abscissa and item-test score correlations on the ordinate. These plots resemble the upper half of a tilted ellipse.

This means that a large item mean difference, $p_u - p_l$ or $M_u - M_l$, for upper vs lower groups implies a high item-test score relationship with a very high degree of accuracy. The converse, however, is not true. A small item mean difference does not imply a low item-test score relationship with much accuracy. Hence, if the pool of items is large relative to the number to be selected, the short-cut method is recommended.

TABLE 28

THE RELIABILITY AND RELATION TO ITEM DISCRIMINATION OF ITEM MEAN DIFFERENCES (D) BETWEEN UPPER AND LOWER GROUPS*

Group		Test	N**	Reliability		D vs r_p bis	D vs r
C	E			C	E	C Score	E Score
I-A	II-A	V	32	.682	.608	.733	.823
I-B	II-B	V	32			.797	.742
III-A	I-A	DI	28	.343	.601	.618	.615
III-B	I-B	DI	28			.832	.674
II-A	III-A	OA	32	.642	.603	.382	.239
II-B	III-B	OA	32			.439	.358

*All correlations are taus.

**Number of items.

H. RELATION OF FERGUSON'S DELTA TO OTHER INDICES

Item deltas were computed for all items administered by the E method. To investigate delta's relation to item-test discrimination estimates, taus were computed between delta and item-test product moment r's for E method data. For C method data items were given delta ranks, in terms of distance from a p of 0.50,* and taus were computed for their correlation with point biserial. Table 29 presents the reliability of delta and its relationship to item point biserials and product moment correlations.

Delta is more reliable than either point biserial or product moment, and its reliability is quite consistent over tests. Delta has very low relationships to both point biserial and product moment r. To investigate delta's relation to difficulty in the E method, item means were plotted against delta values. Delta shows a close curvilinear relationship with item difficulty, the

*Delta for an item scored 1 or 0 is 4 pq which is maximal when p = q = 0.50.

peak of the curve falling above an item score of zero, relatively close to plus one. Apparently delta is maximal when partial information is operating to some extent.

TABLE 29

THE RELIABILITY AND RELATION TO ITEM DISCRIMINATION OF FERGUSON'S DELTA*

Group		Test	N**	Reliability		Delta vs $r_{p\ bis}$	Delta vs r
C	E			C	E	C	E
I-A	II-A	V	32	.861	.794	.393	.383
I-B	II-B	V	32			.395	.464
III-A	I-A	DI	28	.858	.810	.402	.302
III-B	I-B	DI	28			.496	.376
II-A	III-A	OA	32	.811	.815	.071	-.117
II-B	III-B	OA	32			.069	-.004

*All correlations are taus.

**Number of items.

Because of the close relation between delta and item difficulty, there is little to be gained from using delta as an item index for the experimental method.

CHAPTER VI. THE STANDARD OF ASSURANCE

The responses of an individual using the experimental method are probably to some extent a function of his willingness to take a chance by going beyond his sure knowledge. It is conceivable that each individual, independently of his knowledge, sets up a criterion level of "degree of certainty of being right" which serves as a threshold for responding. We have chosen to call this threshold the individual's "standard of assurance." If individuals with the same amount of knowledge may differ in their standards of assurance, then this will contribute to the variance of the test score distribution independently of individual differences in ability per se.

While the experimental response method does not have the "guessing" component contributing to total variance and hence no "correction for guessing" is called for, it may have variance contributed by individual differences in standard of assurance independent of ability. If this is so, it would seem desirable to try to obtain a measure of this standard of assurance at the same time as the test score.

It is easy enough to dream up some possible indices, each plausible and rationalizable, but they may well be unrelated to each other and one is then left with the problem of which index, if any, is a measure of standard of assurance. Obviously, what is called for is a criterion index against which indices from the experimental method could be validated. This was one of the principal reasons for collecting data on one test in each group by the "both" response method. The data from the two response methods on the same test for each individual provided a basis for constructing a criterion index of standard of assurance for each individual. If reasonably reliable, then any index based on the experimental response method alone could be tested against this criterion.

The criterion index of standard of assurance used was the difference between the individual's conventional score on the test and a theoretical score obtained as follows. Wrong alternatives which an individual thought were wrong but did not cross out under the experimental response method represent partial information possessed but not used. The more of this, the higher must an individual's standard of assurance be. Using the two response methods on the same test, it was possible to construct such an index. The individual obtained an E method score on the test and from this it was possible to construct a theoretical conventional score by assuming he had responded by chance to the remaining alternatives in each item. His actual conventional score was based on his actual responses to the remaining alternatives when he was forced to choose, and they may have been responses he did not want included in his E method score. Hence, to the degree that these further responses are correct beyond chance, they represent partial information which the individual did not feel secure enough about to want it to affect his score.

An estimate, then, of the individual's standard of assurance is contained in the disparity between his conventional score on the test and his theoretical conventional score. To summarize, the individual's conventional score uses all the information he has plus a chance component. The theoretical conventional score uses only such information as the individual is assured of plus a chance component. The difference represents information the individual has but which is below some threshold. The greater the disparity of the two scores, the higher the individual's standard of assurance.

Having established a criterion index of standard of assurance for each individual on the test he took by both response methods, the next problem is to construct estimates of this index from data secured by the experimental response method alone. Unfortunately, due to external limitations, it was not possible to study the complete variety of such indices that one might construct. One possible index to be compared with the criterion index is the number of right alternatives crossed out.

The rationale behind this index is based on the notion that the number of alternatives the individual correctly crossed out as compared with the total number he crossed out is a direct reflection of the standard of assurance. Thus, if an individual had a standard of assurance of 80 percent, this would mean that out of an infinite population of alternatives of all levels of difficulty he would be correct 80 percent of the time in the alternatives he chose to cross out. The difference between the numerator and the denominator of such a ratio is the number of correct alternatives crossed out as being incorrect. This difference, the number of correct alternatives crossed out, would serve as a crude but simple index of the individual's standard of assurance. The more correct alternatives crossed out, the lower the standard of assurance. This index assumes that misinformation is not mediating responses to items. The problem of constructing an index of standard of assurance on data obtained by the experimental response method is the difficulty of controlling for partial information and for misinformation.

The index of standard of assurance based on data obtained by the experimental response method can be obtained on data collected by both response methods. These data were used to estimate the reliabilities of this index, labeled SA-E, and also the reliabilities of the criterion index, labeled SA-Crit.

The split-half reliabilities, stepped up by Spearman Brown, are presented in Table 30. It will be observed that the reliabilities are moderate except for the criterion index from the Vocabulary Test, which has too low a reliability to be of use.

TABLE 30

RELIABILITIES OF TWO STANDARDS OF ASSURANCE; SPLIT-HALF INCREASED BY SPEARMAN-BROWN; COMPUTED ON B DATA

Group	Test	SA-Crit	SA-E
I	Object Aperture	0.605	0.783
II	Driver Information	0.639	0.665
III	Vocabulary	0.250	0.765

Another measure of possible interest is the number of distracters not crossed out. These represent cautiousness on the part of the individual and might also serve as an index from E method data of the standard of assurance.

The correlations of the criterion and the SA-E index to various other scores are presented in Table 31. It is evident that the criterion index of standard of assurance is unrelated to ability and that the SA-E index is significantly related to ability. The SA-E index was computed on B method data so its higher correlations with C and E scores from B method data can be attributed to experimental dependence.

The SA-E score obtained for individuals on the test taken by the E method has insignificant correlation with the criterion index for the same individuals on the test taken by the B method. This may mean that the SA-E index is not a measure of standard of assurance as defined by the criterion or it is possible that an individual may have different standards of assurance for different content areas. The data presented here do not permit a choice between these two possible interpretations.

TABLE 31

RELATIONSHIPS OF THE TWO STANDARD-OF-ASSURANCE INDICES TO VARIOUS OTHER SCORES

Score	Method	Group	Test	SA-Crit*	SA-E*
C	C	I	V	-0.079	-0.342
		II	OA	0.145	-0.560
		III	DI	-0.062	-0.345
E	E	I	DI	-0.223	-0.460
		II	V	0.062	-0.469
		III	OA	-0.060	-0.413
C	B	I	OA	-0.100	-0.909
		II	DI	0.237	-0.786
		III	V	0.007	-0.690
E	B	I	OA	-0.255	-0.929
		II	DI	-0.040	-0.800
		III	V	-0.189	-0.799
Distracters not crossed out, E method		I	DI	0.252	0.157
		II	V	-0.016	0.174
		III	OA	0.102	0.227
SA-E, E method data		I	DI	0.121	0.491
		II	V	-0.067	0.476
		III	OA	0.027	0.479

*From B method data.

CHAPTER VII. COMPARATIVE COEFFICIENTS OF DISCRIMINATION

The extension of the range of item scores from one and zero to minus three and plus three makes it possible to distribute individuals into more categories. The degree to which this added capacity for discrimination was realized is shown in Table 32. The Ferguson index, δ, is the ratio of the number of between-persons discriminations actually made to the number for a rectangular distribution, which is maximum. It is thus a relative index, and perhaps there should be no expectation of advantage for the experimental method. The figures in the table, however, show slight, but consistent, differences in favor of the experimental method. These differences may be a bit more substantial than at first apparent, since δ tends to have rather high values.*

TABLE 32

NUMBERS OF DISCRIMINATIONS MADE (D), FERGUSON'S[4] INDEX OF DISCRIMINATION (δ), AND COMPARATIVE DISCRIMINATING ABILITY (δ') OF THE TESTS WHEN GIVEN BY THE C METHOD AND BY THE E METHOD

Test	N		D		δ		δ'*
	C	E	C	E	C	E	
Vocabulary (32 items)	292	269	39657	35410	0.959	0.984	1.009
Driver Information (28 items)	294	292	39572	41573	0.948	0.981	1.010
Object Aperture (32 items)	269	294	33987	42076	0.969	0.979	1.004

*δ' = ratio of the number of discriminations made by the test when given by the E method to the maximum possible number for that same test given the C method, keeping the number of individuals constant at the E method sample size.

*Ferguson, in his paper,[4] gives some examples of discrimination indices for various kinds of distributions. Whereas the δ for a rectangular distribution was 1.000; that for a binomial was 0.9035.

Because of the differing N's the numbers of discriminations made (D) are not comparable between methods. Some data on this point are provided, however, by the δ' column of the table. Here, fcr each test, the number of discriminations actually made when it was administered by the E method is compared with the maximum number possible with the same number of individuals, using the C method.

It is seen that the E method made at least as many discriminations as the C method could possibly make on the same test. Of general relevance to these results is the fact that 229 individuals (26.8%) when responding by the E method crossed out three alternatives on every item in the test and only 9 of these people had perfect scores. This means that a considerable number of the individuals, when presumably using the E method, did the same thing they do under the C method. With proper instruction and feedback from successive tests more individuals might use the E method and discrimination would increase further.

CHAPTER VIII. ATTITUDE OF SUBJECTS

In the interval between the administration of the test which was given by the E method and the one given by the B method, a questionnaire was distributed to the students, asking for their opinions about certain aspects of the test method used. The questions asked were:

1. Which of the two ways of taking a test do you prefer?
2. Which of the two ways of taking a test do you think is more fair?
3. Which of the two ways of taking a test do you think is harder?

The responses to the questions are shown in Table 33.

TABLE 33

PERCENTAGE DISTRIBUTIONS OF QUESTIONNAIRE RESPONSES

Questions	Group I		Group II		Group III		Total	
	Method and Test							
	C V	E DI	C OA	E V	C DI	E OA	C All	E All
1. Which method do you prefer?	30	70	40	60	20	80	30	70
2. Which method is fairer?	15	85	25	75	12	88	17	83
3. Which method is harder?	46	54	52	48	69	31	56	44
N	292		269		294		855	

The general tendency was for the students to say that the E method was preferred, was fairer, and was easier than the C method. How much of this favorableness was engendered by a desire to please the examiners is, of course, not known, but at least it may be said that the experimental method does not arouse antagonism among most subjects who use it for the first time.

Judgments of C vs E method were confounded with tests as the individuals took one test by the C method and a different one by the E method. When the data are examined for comparisons between the tests there is some indication that the Object Aperture Test was preferred and was easiest and that the Driver Information Test was hardest. This variability in the percentages over tests suggests that the questionnaire responses may be dependent in part upon characteristics of the test such as content and difficulty. Detailed comparisons are presented in Table 34.

TABLE 34

RELATION OF TEST CONTENT AND ATTITUDES TOWARD E METHOD

Test	Percent of Times Judged:		
	Preferred	Fairer	Harder
Vocabulary	30	30	31
Driver Information	30	32	41
Object Aperture	40	38	28
	100	100	100

CHAPTER IX. SUMMARY AND CONCLUSIONS

A. SUMMARY

This paper reports an intensive study of an experimental response method for multiple-choice test items. The method is based on the theory that if an individual does not know the answer to an item, he may still know that some of the distracters are wrong. This is partial information. The experimental method requires that the examinee cross out wrong alternatives with the understanding that he gets one point credit for each such distracter crossed out and 1-k (where k is the number of alternatives to an item) credit if he crosses out the answer. For a four-alternative item, the score scale may range from minus three to plus three. All positive scores represent some degree of partial information and all negative scores represent some degree of misinformation. A score of zero on an item is obtained by crossing out all the alternatives or by skipping the item and represents complete ignorance.

With a possible seven-point scale for each item, instead of the two-point scale obtained by the conventional method, the experimental method offers promise of increased test score variance and increased discrimination between individuals. Of interest is the effect of the experimental method on the reliability of a test, and on what the test measures (validity). Some of the implications of the experimental method for item analysis were evaluated, and also an attempt was made to construct a score representing an individual's standard of assurance, which is a variable replacing the guessing component in the conventional method.

Three multiple-choice tests were used, all of which had four-alternative items including an answer and three distracters. The tests were a Vocabulary Test (V), a Driver Information Test (DI), and an Object Aperture Test (OA), the latter being presumably a test of spatial relations. Each test was administered under three different testing procedures to three groups of subjects in a Latin square design. The three testing procedures were the conventional method (C), the experimental response method (E), and both methods jointly (B). The subjects were 855 Jackson High School (Jackson, Michigan) juniors and seniors.

B. CONCLUSIONS

1. Clear evidence for the existence of partial information mediating responses to multiple-choice items was obtained, and some of the implications of the conventional correction for chance formula investigated.

2. On the average, the experimental response method increased the reliability of the tests to a degree equivalent to a 20-percent increase in the test's effective length. The effect on reliability, however, is clearly dependent on the difficulty of the test, the reliability being increased more for more difficult tests.

3. A test administered by the experimental method appears to measure the same complex of abilities as it does when administered by the conventional method.

4. A number of relations between item indices for the two methods and implications for item analysis techniques are pointed out. Basically, what constitutes a good discriminating item is the same for the two methods.

5. A criterion index of the standard-of-assurance variable was constructed, but the two experimental measures of it which were investigated were not valid.

6. The experimental response method makes at least as many discriminations between individuals in the test score distribution as the conventional method could have possibly made under maximally optimal conditions.

7. In response to a three-item questionnaire, the subjects predominantly reported that the experimental method was preferred (70% to 30%) and was fairer (83% to 17%). There was some slight indication that they regarded the conventional method as harder (56% to 44%). The highly cooperative attitude of the subjects, however, may have induced them to give answers they thought the experimenter wanted.

REFERENCES

1. Coombs, C. H., "On the Use of Objective Examinations", _Educational and Psychological Measurement_, _13_, 308-310 (1953).

2. Cronbach, L. J. and Hartmann, W., "A Note on Negative Reliabilities", _Educational and Psychological Measurement_, _14_, 342-346 (1954).

3. Dressel, P. L. and Schmid, J., "Some Modifications of the Multiple-Choice Item", _Educational and Psychological Measurement_, _13_, 574-595 (1953).

4. Ferguson, G. A., "On the Theory of Test Discrimination", _Psychometrika_, _14_, 61-68 (1949).

5. Finch, F. H., "Enrollment Increases and Changes in the Mental Level of the High-School Population", _Applied Psychology Monographs_, No. 10 (1946).

6. Gulliksen, H., _Theory of Mental Tests_, John Wiley and Sons, New York, 1950.

7. Kuder, G. F. and Richardson, M. W., "The Theory of the Estimation of Test Reliability", _Psychometrika_, _2_, 151-160 (1937).

8. Lyerly, S. B., "A Note on Correcting for Chance Success in Objective Tests", _Psychometrika_, _16_, 21-30 (1951).