THE UNIVERSITY OF MICHIGAN
INDUSTRY PROGRAM OF THE COLLEGE OF ENGINEERING

THE ASSESSMENT OF PARTIAL KNOWLEDGE

C. H. Coombs
J. E. Milholland
F. B. Womer

This report is an adaptation of a final report describing contract work done by the Engineering Research Institute, the University of Michigan, under Project 2119, for the Department of the Army, on Contract DA-49-083. The views expressed in this adaptation are those of the authors, and such views are not to be construed as indorsed by the Department of the Army.

IP-123

August, 1955

ENSM

UMR0935

ACKNOWLEDGEMENT

The Industry Program of the College of Engineering wishes to express its appreciation to Dr. Clyde H. Coombs, Professor of Psychology, the University of Michigan, for making it possible to distribute this report under the Industry Program cover.

THE ASSESSMENT OF PARTIAL KNOWLEDGE[1]

By

C. H. Coombs,[2] J. E. Milholland, and F. B. Womer

University of Michigan

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[3] 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.

The conventional response method (C method) for multiple-choice items requires selecting and marking the answer from among the choices offered. In this study it was pick one of four. 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 to 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.[4] Differential choice of distracters 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 plus three to minus 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 information. 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 mark 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 answer crossed out (a negative score)--and then these two scores added. It should be possible to devise scoring methods which would out 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 50-50 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 standards of assurance indices were evaluated, and some implications of partial information for the conventional correction for guessing formula were examined.

II. Design of the Study

A. Test Development and Administration

To provide greater generality of results than a single test would, three 40-item tests: a vocabulary test, a test of driver information, and a test of spatial visualization[5] were developed. 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.

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 have been described above. 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 required to rank three of the four alternatives according to ease of recognition as distracters. The unmarked choice was the one that would have been marked as the answer in the C method. They were then asked to circle those of the ranked alternatives which they were confident enough were actually distracters to want them to be scored as such by the E method.

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 subjects were divided into three groups and the testing programmed so that each group would use each method once and take each test once, and so that each test would be taken by each method once. 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. 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.

B. Editing of Data

In spite of the generous time limits not all examinees finished every test, and it was decided to drop 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 still twenty-one examinees who did not complete the reduced length tests and the scores of these examinees were removed from the data. The loss

of twenty-one cases was less than two and one-half per cent of the total. Whatever bias might be introduced by eliminating them 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-twenty.

C. Equivalence of Groups

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 thirty-eight per cent of the examinees; the one which was most complete had scores from seventy-four per cent. These results indicate that the students tested in this study were fairly representative of the norm groups for the reference tests (6).

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 kindergarten. The language, non-language, 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 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 only one F was significant at the 5% level, so there is no reason to reject the hypothesis of equivalence of groups on the reference variables.

Within each experimental group a random split was made into two subgroups, A and B, for item analysis cross-validation, and these subgroups were examined to see if any important differences existed in their test performances. The differences between the item analysis subgroups all seemed to be trivial: for the C scores there was one difference of two points, the rest were 0 or 1; for the E scores all differences but three were less than four points.

A within- and between-groups analysis of variance was carried out and two of the variance ratios were statistically significant at the 1% level of confidence. These two were attributable to the better performance of Group I on the Object Aperture Test than the other two groups.

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,

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 then 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 $p = 1/k\text{-}1$, 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 hypothesis of partial information and misinformation says that the obtained distribution should exceed the chance distribution in <u>both</u> 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 provides a possible basis for testing these hypotheses. In this method the subject was instructed to rank three alternatives as being incorrect in order from <u>one</u>, the one he was most certain was incorrect, to <u>three</u>,

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 at only 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 influences of partial information and misinformation. The pair chosen to be tested was that of partial information vs. either chance or misinformation, the expectation being that the influence of partial information would exceed that of misinformation.

The hypothesis of the existence of partial information would be sustained 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 expected by chance. This is significant at less than the one per cent 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 expected. This, also, is significant at less than the one per cent 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 expected. 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.

If partial information does exist, it should, perhaps, 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 per cent of times the answers to the missed items were ranked three rather than one or two. The same sub-samples were used as were used to determine whether partial information exists. These sub-samples 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 sub-samples 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 1 presents these results.

Table 1

Product Moment Correlations between Test Scores and Per Cent of Answers to Missed Items Given a Rank of Three

Test	N	r	$\sigma_s^{2\,a}$	$\sigma_t^{2\,b}$	R[c]
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

[a]Variance of sub-sample's test scores.
[b]Variance of total group's test scores.
[c]Corrected for curtailment. (Gulliksen (7), p. 137, equation 18).

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.

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 (9). 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 $0 \leq x \leq \infty$;

$p_i(x)dx$ = 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;

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

$$(1) \qquad p_i'(x)dx = p_i(x)dx + c\,[1 - p_i(x)dx]$$

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

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

which upon expansion becomes:

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

Let:

$$(4) \qquad T_i = n \int_0^\infty p_i(x)f(x)dx$$

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

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

Substituting (4) and (5) in (3), we have:

$$(6) \qquad R_i = (1 - c)\,T_i + cn$$

Solving for T_i:

$$(7) \qquad T_i = \frac{R_i - cn}{1 - c}$$

For a power test:

$$(8) \qquad R_i + W_i = n$$

Substituting (8) in (7):

$$(9) \qquad T_i = R_i - \frac{c}{1 - c} W_i$$

which is the conventional formula for correcting for 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., $R_i + W_i \leq n$. In this case the number of items the individual will get right is given by a modification of equation (3) as follows:

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

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

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

Let:

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

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

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:

$$(14) \qquad n_i = n \int_{0}^{X_i} f(x) dx$$

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

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

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:

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

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 upon 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 covergence of the integrals considered.

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, scores obtained by the E method would not have. 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 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.

The view of the E method as a "lengthening" of the conventional method suggested setting up an index we called[6] 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:

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

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. A CEL = 1 signifies that 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., 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 (8) Formula 20. Reliability comparisons for the conventional and experimental methods appear in Table 2. 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.

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% 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.

Table 2

Reliability Estimates (K - R #20) and Coefficients of Effective Length for Conventional and Experimental Methods

	Vocabulary			Driver Information			Object Aperture		
	A	B	Tot.	A	B	Tot.	A	B	Tot.
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 inter-group coefficients.

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 C method and from E method data. There were thus twelve ten-item subtests: 3 tests x 2 item analysis groups x 2 methods of administration. The mean C method and E method scores made on these tests are shown in Table 3.

Table 3

Means of Ten-Item Constituted Tests, C Method*

		V		DI		OA	
		A	B	A	B	A	B
Picked by C Method	Easy	7.93	7.62	8.14	8.48	8.79	8.69
	Difficult	3.56	3.27	4.32	4.25	6.23	5.81
Picked by E Method	Easy	7.64	7.25	7.92	8.39	8.75	8.48
	Difficult	3.14	8.01	4.35	4.22	5.54	5.69

*Maximum score is 10.

Table 3 (cont.)

Means of Ten-Item Constituted Tests, E Method*

		V		DI		OA	
		A	B	A	B	A	B
Picked by C Method	Easy	52.36	50.48	52.69	54.34	55.04	56.22
	Difficult	35.88	33.99	37.53	38.68	46.36	46.56
Picked by E Method	Easy	51.55	48.99	52.30	53.67	55.05	55.80
	Difficult	33.62	32.79	37.00	38.78	43.73	45.51

*Maximum score is 60.

The relations between difficulty, response method, and reliability are shown in Table 4. In this table, the coefficients for item analysis groups A

Table 4

K-R #20's and CEL's for Constituted Easy and Difficult Ten-Item Tests

Items Selected on C Method Data

		C Score K-R		E Score K-R		CEL	
		Easy	Dif.	Easy	Dif.	Easy	Dif.
V	A	.532	.434	.365	.507	.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	.523	1.072
OA	A	.830	.770	.828	.852	.986	1.720
	B	.744	.717	.745	.751	1.005	1.190

Items Selected on E Method Data

		C Score K-R		E Score K-R		CEL	
		Easy	Dif.	Easy	Dif.	Easy	Dif.
V	A	.506	.619	.415	.669	.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	.775	..857
OA	A	.769	.719	.819	.793	1.359	1.497
	B	.684	.740	.723	.781	1.206	1.253

are based on 10-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% level. Making the comparison another way, eleven of the twelve CEL's for difficult tests are greater than unity (significant at the 1% level by the sign test), whereas only seven of those for the easy tests are gerater 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 than with easy tests.

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 5. 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 heterogenity, but it seems more likely that they represent sampling fluctuations (3).

Table 5

K-R #20 and CEL's
for High Scoring Examinees and Low Scoring Examinees

		High Scorers			Low Scorers		
		K-R		CEL	K-R		CEL
		C Score	E Score		C Score	E Score	
V	A	.378	.337	.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	.916	.802	.866	1.596
	B	.340	.251	.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 upon the mean scores is shown in Table 6. The results of this aspect of the study are presented in Table 7. 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 upon the CEL.

Table 6

Means of the Original Groups and the Groups Reduced by Eliminating Examinees Not Using the Experimental Method

		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

Table 7

K-R #20's and CEL's Computed for those Examinees Who Used the E Method at Least Once--Who Marked Two or Fewer Responses on at Least One Item on the E Method

		N	K-R Computed	K-R* Adjusted	K-R C Method	CEL
V	A	106	.702	.720	.723	.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	.936

*Estimated K-R for group with variance equal to original group; cf. equation (5), p. 111 (7).

B. Comparative Validities

Validity coefficients for the three tests used in this project were based on the sixteen reference variable scores as criteria. They were developed for A and B data separately and together. Table 8 presents the number of times the corresponding coefficients are higher for one method or the other.

Table 8

Number of Times the Sixteen Validity Coefficients Are Higher for Each Type of Score

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

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 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.

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

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. Details of this study are reported elsewhere (2). The general result of the item analysis phase of this investigation is that the characteristics of a good test item are the same for both response methods.

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.

While it is easy enough to dream up some possible indices, each plausible and rationalizable, 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 the individual did not feel secure enough about it 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 only. 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, say, 80%, this would mean that out of an infinite population of alternatives of all levels of difficulty he would be correct 80% 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, labelled SA-E, and also the reliabilities of the criterion index, labelled SA-Crit. The split-half reliabilities, stepped up by the Spearman-Brown formula, are presented in Table 9. 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 9

Reliabilities of Two Standards of Assurance; Split-Half Increased by the Spearman-Brown Formula; Computed on B Data

Group	Test	SA-Crit.*	SA-E*
I	Object Aperture	.605	.783
II	Driver Information	.639	.665
III	Vocabulary	.250	.765

*From B method data.

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 10.

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

Table 10

Relationships of the Two Standards of Assurance Indices to Various Other Scores

	Group and Test	SA-Crit*	SA-E*
C Score - C Method	I (V)	-.079	-.342
	II (OA)	.145	-.560
	III (DI)	-.062	-.345
E Score - E Method	I (DI)	-.223	-.460
	II (V)	.062	-.469
	III (OA)	-.060	-.413
C Score - B Method	I (OA)	-.100	-.909
	II (DI)	.237	-.786
	III (V)	.007	-.690
E Score - B Method	I (OA)	-.255	-.929
	II (DI)	-.040	-.800
	III (V)	-.189	-.799
Distracters Not Crossed Out E Method	I (DI)	.252	.157
	II (V)	-.016	.174
	III (OA)	.102	.227
SA-E - E Method Data	I (DI)	.121	.491
	II (V)	-.067	.476
	III (OA)	.027	.479

*From B Method Data.

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.

VII. COMPARATIVE COEFFICIENTS OF DISCRIMINATION

The extension of the range of item scores from 1 and 0 to -3 and +3 makes it possible to distribute individuals into more categories. The degree to which this added capacity for discrimination was realized is shown in Table 11. 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.[7]

Table 11

Numbers of Discriminations Made (D), Ferguson's (5) 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	.959	.984	1.009
Driver Information (28 items)	294	292	39572	41573	.948	.981	1.010
Object Aperture (32 items)	269	294	33987	42076	.969	.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.

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, for 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 nine 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 feed-back from successive tests more individuals might use the E method and discrimination would increase further.

VII. 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 12.

Table 12

Percentage Distributions of Questionnaire Responses

Questions	Group I Method and Test C V	Group I Method and Test E DI	Group II Method and Test C OA	Group II Method and Test E V	Group III Method and Test C DI	Group III Method and Test E OA	Total C all	Total 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.

IX. SUMMARY AND CONCLUSIONS

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 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 -3 to +3. 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 method on the reliability of a test, and on what the test measures (validity). Some of the implications of the 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 (Michigan) juniors and seniors.

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% 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 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 new method was preferred and was fairer. There was some slight indication that they regarded the conventional method as harder. The highly co-operative attitude of the subjects, however, may have induced them to give answers they thought the experimenter wanted.

FOOTNOTES

1. This is an adaptation of a final report to The Personnel Research Branch, The Adjutant General's Office, Department of the Army on Contract DA-49-083 OSA-638. The views expressed are those of the authors and not to be construed as indorsed by the Department of the Army.

2. The writers are indebted to a number of individuals who gave generously of their time and interest. The list includes Dr. Kent Leach, Director of the University of Michigan Bureau of School Services; Mr. W. Earl Holman, Principal of Jackson High School, Jackson, Michigan; Mr. Earl Allgaier of the American Automobile Association, Washington, D.C.; and many professional colleagues: Philip DuBois, David Birch, David Beardslee, William Hays, and J. E. Keith Smith.

3. Originality for the method is not being claimed here. It has apparently been part of the lore amongst psychometricians for some years although we were not able to find any mention of it in the literature. One of the referees of this paper informed us that he has been teaching the method for 15 to 20 years and it is not original with him.

4. A method complementary to this is one proposed in (4) 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.

5. Spatial visualization items were obtained from an experimental form of the DuBois and 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 be slid over and dropped through.

6. Philip DuBois suggested this name for the index.

7. Ferguson, in his paper (5), 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 .9035.

REFERENCES

1. Coombs, C. H., On the Use of Objective Examinations. Educational and Psychological Measurement. 1953, 13, 308-310.

2. Coombs, C. H., Milholland, J. E., and Womer, F. B. The Assessment of Partial Knowledge. PRB Tech. Res. Note 33, Dept. of the Army, Wash. D.C.

3. Cronbach, L. J. and Hartmann, W. A Note on Negative Reliabilities. Educational and Psychological Measurement, 1954, 14, 342-346.

4. Dressel, P. L. and Schmid, J. Some Modifications of the Multiple-Choice Item. Educational and Psychological Measurement, 1953, 13, 574-595.

5. Ferguson, G. A. On the Theory of Test Discrimination. Psychometrika, 1949, 14, 61-68.

6. Finch, F. H. Enrollment Increases and Changes in the Mental Level of the High-School Population. Applied Psychology Monographs, 1946, No. 10.

7. Gulliksen, H. Theory of Mental Tests. New York: John Wiley and Sons, 1950.

8. Kuder, G. F. and Richardson, M. W. The Theory of the Estimation of Test Reliability. Psychometrika, 1937, 2, 151-160.

9. Lyerly, S. B. A Note on Correcting for Chance Success in Objective Tests. Psychometrika, 1951, 16, 21-30.