DOCUMENT RESUME

ED 041 037 TM 000 001

AUTHOR Collet, LeVerne S.
TITLE Elimination vs. Best Answer Response Modes for M-C Tests.
PUB DATE Mar 70
NOTE 21p.; Paper presented at joint annual meetings of American Educational Research Association and National Council on Measurement in Education, Minnespolis, Minnesota, March 2-6, 1970

EDRS PRICE EDRS Price MF-$0.25 HC-$1.15
DESCRIPTORS *Multiple Choice Tests, *Scoring, *Test Reliability, *Test Validity

ABSTRACT

A critical review of systems of scoring multiple choice tests is presented and the superiority of a system based upon elimination method over one based upon the best answer mode is hypothesized. This is discussed in terms of the capacity of the mode to reveal the relationships among decoy options and the effects of partial information, misinformation and guessing. Tests were administered to two groups of subjects and scored according to each of the three treatment modes, classical, weighted choice, and elimination. In addition, subjects were asked to indicate their confidence in the correctness of each answer. Thus, treatment, confidence, and knowledge scores were computed for each subject and a Gulliksen-Wilks regression test was performed on the data to compare the validity and reliability of the three scoring modes. The results generally support the hypothesized superiority of the elimination scores. Elimination produced higher validities and reliabilities and less guessing than either of the other two treatments. Although the design did not permit a definitive comparison of elimination and confidence scores, there was some evidence that elimination scores were at least as valid as confidence scores. (Author/PR)

ELIMINATION VS. BEST ANSWER RESPONSE MODES FOR M-C TESTS

by

LeVerne S. Collet

The University of Michigan

Most multiple-choice scoring systems currently in vogue use the best-answer response mode, which requires testees to choose a single answer from a set of k alternative options. The original k-category scale is then transformed to a binomial by classifying the response as either right or wrong. Relationships among decoy options are thus lost, and the effects of partial information, misinformation and guessing inextricably confounded. It is the contention of this paper that the full information potential of a multiple-choice item can be retained only if the response mode utilizes all the degrees of freedom contained in the original item. The proposed elimination response mode has that capacity.

In order to place the arguments for elimination scoring in context, a brief review of the various models of the multiple-choice item is desirable. During the first three decades of this century the classical model was developed. In the classical view, a k-option multiple-choice item consisted of one correct answer and (k-1) equally-incorrect decoys. It was explicitly assumed that subjects who did not know the correct answer would choose randomly among all k options in the set. From this assumption was derived the classical correction for guessing formula:

$$T = R - W/(k-1)$$

where T is the estimated true score, R the number right, W the number wrong, and k the number of options per item.

A major criticism of the classical model was advanced by Horst (1933), who pointed out that variations in plausibility among decoy options can drastically affect the role of chance. For example, if an examinee can eliminate one decoy option his chance of a correct guess is increased from

1/k to 1/(k-1). In the presence of partial information, the classical formula will always undercorrect.

Horst (1933, 1966 Ch. 14) proposed an item model which, rather than equally-plausible decoys, posited only that the best answer and decoys could be arranged in order of plausibility on a unidimensional scale. The model explicitly assumes that subjects who do not know the correct answer will choose randomly among only the uneliminated options.

Thus, (k-1) levels of "guessers" are identified: those with zero knowledge (G_0), those with enough knowledge to eliminate the least plausible option (G_1), and so on, up to those with knowledge enough to eliminate all options but the best answer (G_{k-1}). Assuming random distribution of guesses, the number of responding to each option would be:

Least plausible option:	$\frac{1}{k} G_0$
Second least plausible option:	$\frac{1}{k} G_0 + \frac{1}{k-1} G_1$
. . . .	
Most plausible option:	$\frac{1}{k} G_0 + \frac{1}{k-1} G_1 + \frac{1}{k-2} G_2 + \ldots + \frac{1}{2} G_{k-2}$
Best answer option:	$\frac{1}{k} G_0 + \frac{1}{k-1} G_1 + \frac{1}{k-2} G_2 + \ldots + \frac{1}{2} G_{k-2} + G_{k-1}$

The difference between the last two options is obviously the term G_{k-1}, that is, the number of people who knew the right answer or who could eliminate all the incorrect options. Although the foregoing rationale permits the computation of group item statistics such as the "true" difficulty and "true" variance, the application of Horst's model to the prediction of the "true" scores of individuals presents some mathematical problems which are as yet unsolved. However, the Horst argument has stimulated the development of a number of "weighted-choice" scales, such as those reported by Davis and Fifer (1959) and Merwin (1959). Both investigators reported that the procedure yielded somewhat higher reliability and validity coefficients than the classical procedure.

Nevertheless, both the classical and Horst models assume that all wrong answers are the result of guessing. This assumption is both logically invalid, and contrary to the empirical evidence (Davis 1959; Davis and Fifer 1959; Sax and Collet 1968). It appears that the majority of incorrect responses are due to the presence of misinformation, rather than the absence of positive information.

The confidence scoring system represents an attempt to account for the effects of both guessing and misinformation. In general, confidence scoring requires a subject to attach to each optional response a number (usually a percent) which represents his confidence that it is the correct answer. Note that the confidences attached to various options within an item must sum to 100 percent. The score is determined by the amount of confidence attached to the correct answer; 100% confidence would yield a maximum score +1, (100/k)% would represent a guessed answer and yield a score of zero. Misinformation would depress the confidence in the correct answer below the chance level, resulting in a negative score. Complete confidence in an incorrect answer would result in zero confidence in the correct option and yield a score of -1. In the more sophisticated systems, the assigned score is a non-linear function of the confidence attached to the correct answer. Schuford et al (1966) suggested a scoring formula which they demonstrated mathematically to yield scores such that, regardless of his level of skill, a subject can maximize his score if and only if he honestly reflects his degree-of-belief probabilities for each option. The confidence model assumes that:

1. Subjects are aware of the state of their own knowledge.
2. Subjects have or can easily acquire the skills necessary to transform their internal confidence into a (numerical) response.
3. When shown it is to their advantage, subjects will not guess, but honestly record their confidence in each answer.
4. Misinformation is of negative value.

Although the confidence model deals nicely with guessing and misinformation, the skills required to use such a response system restrict its utility. Even with the aid of a tool such as the scorule (designed by the Schuford Massengill Corporation), it seems likely that scores will be influenced by variations in manipulative dexterity and spatial ability which may be independent of the variable being assessed. At the very least, confidence scoring is both slow and, if a scorule is required, expensive.

Elimination Model

The three models above share the implicit assumption that a subject's knowledge must be assessed in terms of a single response -- which option was chosen, or how much confidence was attached to the keyed answer. However, both observation and personal experience suggest that the response to all but the most elementary factual items requires a set of sequential decisions rather than a single act of recognition. For example, if a student is asked to justify his answer, he is likely to do so by comparing it to each decoy option in turn. A process analogous to the paired-comparisons technique seems to be an efficient, if not required, approach to the multiple-choice test -- at least for items requiring application, analysis, or evaluation. The basic premise of the elimination model is that partial knowledge can be assessed by breaking the total item response into paired-comparison components. This is accomplished by the simple expedient of requiring subjects to respond to items by eliminating any (or all) of the options they know to be incorrect. The elimination scoring procedure assumes that:

1. When responding to multiple-choice items by marking out the eliminated options, subjects will conduct a series of paired-comparisons, at each step eliminating one option and carrying the other to a subsequent comparison.

2. Regardless of the combinations compared, the answer to any item with k options will be determined by k-1 paired-comparison decisions. Each comparison is assumed to involve 1/(k-1) of the total knowledge content of the item.

3. Misinformation has negative value of a weight equivalent to the knowledge displaced. Since eliminating the correct option will always result in a wrong answer it is assigned a value equal to the negative of the weight assigned to the whole item.

4. Subjects will not guess if they are shown that it is not to their advantage to do so.

The elimination scoring formula is derived directly from the model. If an item with k options is assigned a weight W, then each elimination of an incorrect answer is worth W/(k-1), and elimination of the correct answer is worth -W. Thus the score on a test of N items having k options each may be calculated as:

$$\text{Score} = \sum_{i=1}^{N} \frac{E_i - C_i\,(k_i - 1)}{k_i - 1}\,W_i$$

where the subscript i identifies the item, E_i is the number of incorrect options eliminated, C_i is the number of correct answers eliminated (always one or zero), k_i is the number of options in item i and W_i is the weight assigned to the whole item. If a test is composed of items having the same number of options and equal weights, the subscripts and the denominator may be ignored, and the formula rewritten as:

$$\text{Score} = E - C(k-1)$$

where E is the total number of incorrect decoys eliminated on the whole test, C is the total number of correct answers eliminated, and k is the number of options in each item.

The formula outlined above shares with confidence scores the property that it is never to a subject's advantage to guess. The proof for this proposition is outlined below. If k is the number of options in an item and j is the number of eliminations a subject has made, then the probability that the correct answer was eliminated (the probability that C=1) is given by the expression:

$$P(C = 1) = \frac{1}{k} + \frac{k-1}{k}\left(\frac{1}{k-1}\right) + \dots \frac{k-j}{k}\left(\frac{1}{k-j}\right)$$

$$\text{or,}\quad P(C - 1) = \frac{j}{k}$$

The probability that the correct answer is not eliminated then becomes:

$$P(C = 0) = 1 - \frac{j}{k} = \frac{k-j}{k}$$

The formula can then be applied to calculate the score for each value of C at j eliminations

$$\text{At}\quad C = 0 : \text{Score} = \frac{j-0\ (k-1)}{k-1} = \frac{j}{k-1}$$

$$\text{At}\quad C = 1 : \text{Score} = \frac{(j-1) - 1(k-1)}{k-1} = \frac{j-k}{k-1}$$

The chance score for any j eliminations would be:

$$\text{Chance Score} = P(C = 0)\left(\frac{j}{k-1}\right) + P(C = 1)\left(\frac{j-k}{k-1}\right)$$

$$\text{Chance Score} = \frac{(k-j)j}{k-1} + \frac{j(j-k)}{k-1} = 0$$

Since the chance score is always zero, it is never to the subject's advantage to guess.

A tabular comparison of scores yielded at various levels of knowledge under the four scoring systems is presented in Figure 1. It will be observed that the scores of the elimination and classical model will be identical whenever four options are marked, but that the classical system makes no adjustment in scores for partial information or for levels of misinformation. The weighted-choice method can yield scores which are fairly similar or quite different, depending on the relative sizes of the weights used. The one tabulated here was used in the experiment outlined below.

EMPIRICAL ASSESSMENT

The empirical assessment of the relative efficacy of classical (C), elimination (E), and weighted-choice (W) scores as reported here derives primarily from data collected as part of a doctoral dissertation conducted at the University of Washington. However, supplementary data is provided by a partial replication of the experiment at the University of Michigan. The C and E scores in this paper were calculated by the formulas given above. W scores were obtained by first ranking the "correctness" of the decoy options according to the average score of subjects who chose the option. Plus 4 was assigned for the best answer, then +2, 0, -2, and -4 for the decoys in descending order of correctness.

In the initial study (Collet, 1968), equivalent forms (1 and 2) of a test of mental maturity were sequentially administered to two groups of subjects under each of three scoring procedures. Test sequence was balanced within each treatment by administering the test forms to one group in order 1-2 and the other in order 2-1. In addition to answering items according to the treatment instructions, all examinees were asked to attach a confidence rating to each item according to the following key:

0 -- means you have no confidence that your answer is right.

1 -- means you think the answer is right but have some doubt.

2 -- means you are moderately certain your answer is right.

3 -- means you are nearly positive your answer is right.

Finally, each student's score on the verbal section of the Washington College Entrance Test was obtained from the files to serve as a criterion of validity; these were labelled Y scores. The Y scores were analyzed in a one way ANOVA to provide a check on the initial differences among the six groups. The results are summarized in Table I. It was observed that all differences among groups were small, with F ratios near unity or below. It was concluded that the six groups were of comparable initial ability.

Three dependent-variable scores were computed for each subject. Treatment scores (X) were calculated according to the C, E, and W, formulas given above. The knowledge score (K) was the sum of the confidence ratings attached to correct answers minus the sum of confidences attached to incorrect answers. The guessing score (G) was the number of guessed items, where a guess was defined as an item in which a response was recorded with zero confidence. Note that the confidence rating was attached to the whole item: subjects in the E treatment simply indicated their confidence that all items eliminated were wrong, while those in C and W treatments indicated their confidence that they had chosen the best answer.

The results were analyzed in two stages. In stage one, each score was entered into a three way ANOVA. Factor A was scoring technique with A_1 classical, A_2 elimination and A_3 weighted choice. Factor B was test order with B_1 order 1-2 and B_2 order 2-1. Factor C was test form with subscripts corresponding to the form used.

The ANOVA of K scores is reported in Table II. It was observed that all differences failed to reach significance, with all F ratios except that for test forms being near unity or below. The ANOVA of X scores, on the other hand (see Table III), yielded a significant main effect for factor A. It was concluded that the three scoring techniques do not yield equivalent scores.

The ANOVA of guessing scores is summarized in Table IV. It was observed that there was a significant main effect of factor A ($p < .01$) A subsequent Newman-Keuls comparison among all pairs of ordered means yielded significant differences for both the E - C and E - W comparisons. It was concluded that there was significantly more guessing in both the classical and weighted-choice methods than in the elimination method.

In the second stage of the analysis, the reliability and validity of the X scores under the three scoring techniques were compared using the Gulliksen-Wilks regression test for several samples. In each of the following comparisons, subscripts are used to identify the test form on which the scores were computed. The correlations, standard errors and the chi-square values obtained from comparing SE's, regression weights, and intercepts are summarized in Table V. In all cases the direction of prediction is given by the tabular headings.

Reliabilities. It was observed that the correlation of X_1 and X_2 scores were highest under the E method (.858), next highest for C (.809) and lowest under W (.725). Since overall differences among SE's were just below significance at the .05 level, a subsequent comparison among pairs was computed. The EW difference was significant ($p < .05$) but the EC and WC differences were not.

The reliabilities of the knowledge scores (K_2 predicted by K_1) followed the same pattern, but none of the differences reached significance. In addition, it was observed that the reliability of knowledge scores was slightly lower than that of treatment scores in both E and C, and slightly higher under W. The SE's for K and X are not comparable due to the different criterion scores used.

Concurrent Validity. The prediction of summed K scores from summed X scores was used as an estimate of congruent validity. It was observed that the correlation coefficients and standard errors were in the order CEW. The C error was significantly smaller than W, but the elimination error did not differ significantly from either classical or weighted choice errors.

Predictive Validity. The predictive validity of both the X and K scores was assessed by the accuracy with which they predicted Y scores. Both X and K scores produced the highest correlations and lowest prediction errors under elimination. The C treatment was slightly better than W for X scores, but the relative position reversed for K scores. For both scores the EC and EW comparisons were significant but the CW comparison was not. It was concluded that the elimination technique produced higher predictive validity than the classical or weighted choice technique.

The close parallelism of the X and K findings suggests that the E method was operating similarly for both scores. Apparently, the subjects were able to assess the overall corrections of an item consisting of one, two, three or four eliminated options more validly than they could assess items requiring a single option to be marked. As a result of this interaction of K and treatments, the concurrent validity findings (above) are not interpretable.

REPLICATION

A partial replication of the above experiment was conducted at the University of Michigan. A class of 40 statistics students was randomly divided into two groups of twenty. Both groups were administered a thirty-five item multiple-choice midterm test, with group 1 using elimination and group 2 classical scoring techniques. The criterion of validity was the students' success or failure on a laboratory exercise which required them to compute a t test for significant differences in the means of two sets of simulated scores and to state appropriate educational conclusions. The results are summarized in Table VI. It was observed that the mean and validity of the elimination group was higher than for the classical group, although none of the differences were significant. Nevertheless, the direction of the increment in validity agreed with the finding in the previous study. Subsequently, a sign test was conducted to test the hypothesis that the elimination method made the test "easier" -- that is, it tended to increase his score. The total of obtained scores divided by the total possible score was used as an index of difficulty. Elimination yielded a higher index than classical for 22 times and a lower index for 4. This difference yielded a z of 4.08 which. was highly significant ($p < .001$). Despite the fact that the obtained t of 1.64 for the difference between means was insignificant ($p < .10$), the combined evidence was deemed sufficient to conclude that the elimination method was facilitory for this test.

SUMMARY OF CONCLUSIONS

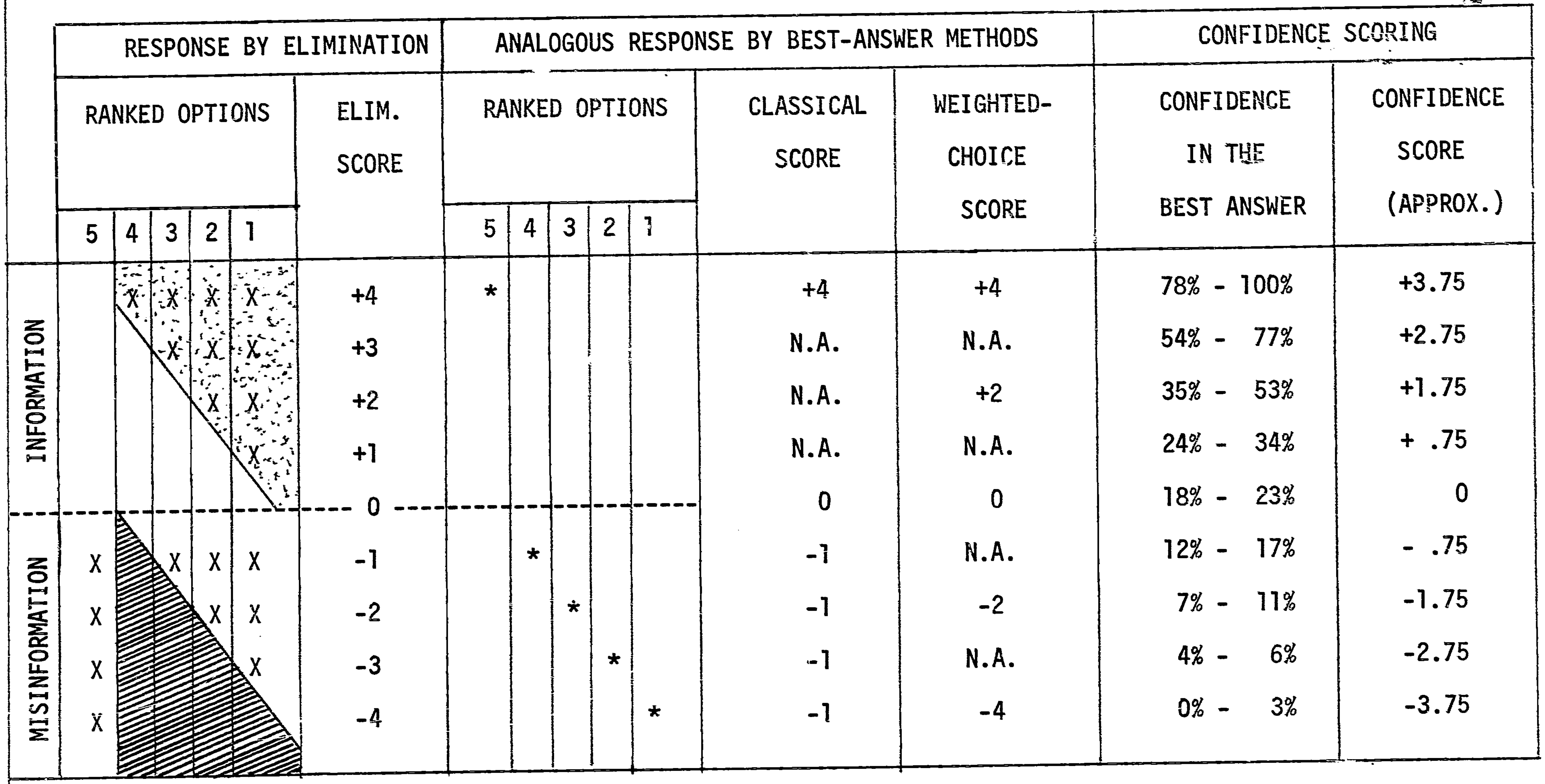
1. The effect of elimination on test difficulty seems to be somewhat equivocal since the E mean was lowest in the original data but higher in the

replication. It is possible that the test content was responsible for the shift. Perhaps the superior C scores in the original data are a result of the undercorrection for guessing -- the fact that course grades were influenced by the mid-term scores would tend to suppress guessing in the replication. In addition, there is some evidence that the somewhat slower elimination method was a disadvantage: on the average, C subjects answered 1.6 more items than E subjects. Although this difference was not significant, it seems likely that elimination would always yield lower scores under stringent time limits. It is recommended that time limits be adjusted when using elimination scoring.

2. The evidence presented supports the hypothesis that elimination scoring would reduce the number of guessed responses. The general pattern of the reliabilities and validities, plus the ANOVA of the guessing score all indicate that there were fewer random responses under elimination than under either classical or weighted-choice scoring procedures. However, it seems fairly clear that guessing per se is not a serious problem with any of the methods used: even the highest group guessed an average of only 1.82 items per subject. It is suggested that the capacity to assess the amount of misinformation is an important characteristic of a scoring system. Only E and K scores possess this capacity (see Figure 1).

3. The overall results indicate that elimination scores are more valid and at least as reliable as classical corrected-for-guessing or weighted-choice scores, with W scores generally least valid. The findings regarding W scores, however, must be restricted to the particular weighting procedure used in this study. The study did not permit a complete assessment of the relationship between elimination and confidence scores. Confidence scores seem to be more

valid when the elimination procedure is used than when they are attached to a single best answer; but, in general, the validity of the confidence scores appears to be somewhat lower than that of elimination scores. It is recommended that a direct comparison of elimination and Schuford-type confidence scores be conducted in the near future. However, until such a comparison is made, its relative simplicity would seem to favor the elimination technique.



	RESPONSE BY ELIMINATION						ANALOGOUS RESPONSE BY BEST-ANSWER METHODS							CONFIDENCE SCORING	
	RANKED OPTIONS					ELIM. SCORE	RANKED OPTIONS					CLASSICAL SCORE	WEIGHTED-CHOICE SCORE	CONFIDENCE IN THE BEST ANSWER	CONFIDENCE SCORE (APPROX.)
	5	4	3	2	1		5	4	3	2	1				
INFORMATION		X	X	X	X	+4	*					+4	+4	78% - 100%	+3.75
			X	X	X	+3						N.A.	N.A.	54% - 77%	+2.75
				X	X	+2						N.A.	+2	35% - 53%	+1.75
					X	+1						N.A.	N.A.	24% - 34%	+ .75
						0						0	0	18% - 23%	0
MISINFORMATION	X		X	X	X	-1		*				-1	N.A.	12% - 17%	- .75
	X			X	X	-2			*			-1	-2	7% - 11%	-1.75
	X				X	-3				*		-1	N.A.	4% - 6%	-2.75
	X					-4					*	-1	-4	0% - 3%	-3.75

Figure 1. Graphical representation of all possible responses under the elimination model, with comparable responses for the classical, weighted-choice, and confidence systems.

SCORING TECHNIQUE AND INSTRUCTIONS TO SUBJECTS	ORDER OF TESTS
A_1: Classical - Mark out the letter corresponding to the best answer. DO NOT MAKE WILD GUESSES. One quarter of your incorrect answers will be deducted from your score.	B_1: (1) C_1 - test 1 (2) C_2 - test 2 B_2: (1) C_2 - test 2 (2) C_1 - test 1
A_2: Elimination - Indicate your answer by blacking out the letters corresponding to the INCORRECT answers. If you are unsure of the best answer, you may receive partial credit by eliminating one or more of the incorrect options. DO NOT GUESS: score 1/4 for each elimination, but subtract 1 for eliminating the correct answer.	B_1: (1) C_1 - test 1 (2) C_2 - test 2 B_2: (1) C_2 - test 2 (2) C_1 - test 1
A_3: Weighted Choice - Mark out the letter corresponding to the best answer. DO NOT MAKE WILD GUESSES. Incorrect answer may either give 1/2 credit, 0, -1/2 or -1, depending on which wrong option you choose.	B_1: (1) C_1 - test 1 (2) C_2 - test 2 B_2: (1) C_2 - test 2 (2) C_1 - test 1

Figure 2. Experimental Design

TABLE I

ANOVA OF Y SCORES

Source	df	MS	F
Between groups	5	205.11	.84
Within groups	276	242.77	
Total	281		

Group	I	II	III	IV	V	VI
Mean	63.60	62.74	61.56	62.12	64.46	66.50

TABLE II

ANOVA OF K SCORES

Source	df	MS	F
Between Subjects	281		
A: Scoring techniques	2	1180.70	.97
B: Test order	1	984.09	.81
AB	2	1605.37	1.32
Error between	276	1214.14	
Within Subjects	282		
C: Test form	1	308.31	2.20
AC	2	47.89	.34
BC	1	24.27	.17
ABC	2	140.04	

Technique	A_1: Classical	A_2: Elimination	A_3: Weighted Choice
Means	45.55	40.93	44.91

TABLE III

ANOVA OF X SCORES

Source	df	MS	F
Between Subjects	281		
A: Scoring techniques	2	18,660.06	11.14 **
B: Test order	1	1,073.19	.65
AB	2	2,124.72	1.27
Error Between Subjects	276	1,674.64	
Within Subjects	282		
C: Test form	1	232.38	1.24
AC	2	427.03	2.28
BC	1	240.13	1.28
ABC	2	53.84	.29
Error Within Subjects	276	187.28	

Table of Differences (Row - Column)

Scoring Technique	A_3: Weighted-Choice	A_1: Classical	A_2: Elimination
Ordered Means	102.23	90.20	82.57
A_3 102.23	———	12.03 **	16.66 **
A_2 90.20		———	7.63
		$r = 2$	$r = 3$
Critical difference ($p \leq .01$)		10.37	12.14

**p < .01

TABLE IV

ANOVA OF G SCORES

Source	df	MS	F
Between Subjects	281		
A: Scoring techniques	2	66.41	5.87 **
B: Test orders	1	1.29	.11
AB	2	3.61	.32
Error between subjects	276	11.31	
Within Subjects	282		
C: Test form	1	4.98	1.09
AC	2	3.44	.75
BC	1	4.26	.93
ABC	2	5.69	1.25
Error within subjects	276	4.57	

Table of Differences (Row - Column)

Scoring Technique	A_3: Weighted-Choice	A_1: Classical	A_2: Elimination
Ordered Means	1.82	1.45	.66
A_3: 1.82	______	.38	1.16 **
A_2: 1.45		______	.79 *

	r=2	r=3
Critical difference (.05)	.68	.81
Critical difference (.01)	.89	1.01

* $p \leq .05$ ** $p \leq .01$

TABLE V

SUMMARY OF REGRESSION ANALYSES

	RELIABILITIES				CONCURRENT VALIDITY		PREDICTIVE VALIDITIES			
Prediction made →	X_1 to X_2		K_1 to K_2		ΣX to ΣK		ΣX to Y		ΣK to Y	
	r	SE_{est}	r	SE_{est}	r	SE_{est}	r	SE_{est}	r	SE_{est}
Classical	.809	21.04	.795	16.72	.929	18.15	.668	11.52	.582	12.59
Elimination	.858	16.50	.826	16.60	.922	20.94	.777	9.39	.759	9.70
Weighted Choice	.725	21.04	.759	15.73	.856	23.05	.646	12.29	.692	11.62

Difference Tested	Obtained		Chi	Square	
SE's of estimate	5.644*	.421	5.407*	7.172**	6.499**
Regression weights	.997	1.806	4.654*	N.A.	N.A.
Intercepts	8.160**	1.117	53.671***	N.A.	N.A.
Ordered SE's	E C W	W E C	C E W	E C W	E W C

Note: SE's are ordered smallest to largest. Those underscored by the same line do not differ significantly ($p < .05$).

* $p \leq .10$ ** $p \leq .05$ *** $p \leq .01$

TABLE IV

SUMMARY OF PARTIAL REPLICATION

	Test Data		Criterion Data		Test to Criterion Prediction		
	Mean	SD.	Mean	SD	r	SE_{est}	
Group 1: Elimination	85.0	19.33	.50	1.25	.22	1.48	} Chi Square < 1.
Group 2: Classical	74.5	21.34	.50	1.25	.11	1.55	

observed t for $M_1 - M_2 = 1.64$

* $p \leq .10$ ** $p \leq .05$ *** $p \leq .01$

REFERENCES

Collet, LeVerne S. *An Evaluation of the Elimination Technique for the Scoring of Multiple-Choice Tests.* Unpublished Ph. D. dissertation, University of Washington, Seattle, 1968.

Coombs, C., Milholland, J.E., and Womer, F. *The Assessment of Partial Knowledge.* Educ. and Psychol. Measurement. 1956, 16: 13-37.

Davis, F.B. *Use of Correction for Chance Success in Test Scoring.* Journal Educ. Res. 1959, 52: 279-80.

Horst, Paul *The Chance Element in the Multiple-Choice Test Item.* Journal Gen. Psychol. 1933, 24: 345-6.

Horst, Paul *Psychological Measurement and Prediction.* Wadsworth Publishing Co., Belmont, Calif., 1966.

Merwin, Jack C. *Rational and Mathematical Relationships of Six Scoring Procedures Applicable to Three-Choice Items.* Journal Educ. Psychol. 1959, 50: 153-161.

Sax, G., and Collet, L.S. *The Effects of Differing Instructions and Guessing Formulas on Reliability and Validity.* Educ. and Psychol. Measurement. 1968, 28: 1127-1136.

Schuford, E.H., Alberta, A., and Massengill, H.E. *Admissable Probability Procedures.* Psychometrika, 31: 125-145, June, 1966.

Valid Confidence Testing (kit). The Schuford Massengill Corporation. P.O. Box 26, Lexington, Mass. 1957.