DOCUMENT RESUME

ED 115 670 TM 004 940

AUTHOR Cross, Lawrence H.
TITLE An Investigation of a Scoring Procedure Designed to Eliminate Score Variance Due to Guessing in Multiple-Choice Tests.
PUB DATE [Apr 75]
NOTE 22p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, D.C., April 1975)

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS Algebra; Comparative Analysis; *Guessing (Tests); *Multiple Choice Tests; *Response Mode; Response Style (Tests); *Scoring Formulas; Senior High Schools; Statistical Analysis; Testing Problems; *Test Reliability

ABSTRACT

A novel scoring procedure was investigated in order to obtain scores from a conventional multiple-choice test that would be free of the guessing component or contain a known guessing component even though examinees were permitted to guess at will. Scores computed with the experimental procedure are based not only on the number of items answered correctly, but also on the average quality of both correct and incorrect choices as reflected in the difficulty and discrimination values associated with these choices. The scores resulting from this procedure were compared to number-right and conventional formula scores for predicting guessing-free scores, which are independently determined for the same test. These data suggest that significant increases in reliability can result if score credit is assigned only to items that were answered correctly without guessing on any type. Since the examinees were not aware that the statistical analysis of their tests included a penalty for guessing, findings should be treated with theoretical intrest. (Author/BJG)

AN INVESTIGATION OF A SCORING PROCEDURE DESIGNED TO ELIMINATE SCORE VARIANCE DUE TO GUESSING IN MULTIPLE-CHOICE TESTS[1]

Lawrence H. Cross
Virginia Polytechnic Institute and State University

Paper presented at the Annual Meeting of
the National Council on Measurement in Education
Washington, D.C., April, 1975

[1]Portions of this paper are based on the writer's Ph.D. dissertation at the University of Pennsylvania.

AN INVESTIGATION OF A SCORING PROCEDURE DESIGNED TO ELIMINATE SCORE VARIANCE DUE TO GUESSING IN MULTIPLE-CHOICE TESTS

Introduction

The conventional multiple-choice response mode requires the examinees to identify and mark the correct choice to each item. The most direct method of scoring such responses is number-right scoring, whereby every correctly marked choice in an item is assigned a score of one and each other choice is assigned a score of zero. A major limitation of the foregoing procedure is that the examiner is unable to determine whether a correct response is the result of sufficient knowledge to answer the question correctly or whether a correct response is a successful guess among two, three, or more choices. Although it is generally agreed that some attempt should be made to control the effect of guessing, to date few scoring methods other than the conventional correction for guessing have been proposed that explicitly attempt to do so, and each method has its critics.

The conventional correction for guessing simply involves subtracting from number-right scores, a quantity reflective of the number of items answered incorrectly. The fact that this procedure is not entirely satisfactory is evident from the numerous studies that have argued for or against this procedure. (See Cross, 1973, or Diamond and Evans, 1971, for a critical review of these studies.)

Coombs (1953) proposed an alternative test-taking procedure designed to assess partial information. With this procedure, examinees are directed to mark only those choices they are certain are incorrect and to leave the correct choice unmarked. A scoring rule presumed to insure this type of behavior is is imposed. On an item having 'C' choices, one point is awarded for each distracter marked, but a score penalty of (C - 1) is imposed if the correct choice is marked. Thus, an examinee is able to express and receive credit for partial information but will be severely penalized if he erroneously marks the correct choice as a distracter. Consequently, guessing under these conditions is not a profitable game to play as suggested by Coombs, Milholland and Womer (1956) and by Lord and Novick (1968, p. 315). Several studies have investigated the effect this response mode and scoring procedures have on the reliability and validity of the resulting scores (Coombs *et al.*, 1956; Collet, 1971; Koehler, 1972). The results of these studies suggest that the reliability and validity of the scores can be expected to improve or show no difference when compared to other scoring procedures. Aside from the effect this procedure has on reliability or validity, from a logical standpoint, it would seem that the elimination scoring procedure will inhibit guessing behavior more effectively than any other testing procedure. Two major drawbacks of this procedure, however, are the additional time required to administer a test and the inconvenience of having to train examinees

in the use of this response mode every time the test is administered. The Coombs response mode was selected for use in the present study to establish criterion score sets for the experimental test which would reflect varying degrees of guessing.

The present study was designed to investigate a novel scoring system that would make it possible to provide scores that closely approximate those that (a) are free from the guessing component, or (b) include a controlled guessing component as initially determined by use of the Coombs response mode. The proposed scoring system is designed to be used in conjunction with the standard response mode, and it does not require directions admonishing the examinees to refrain from guessing. Consequently, it would offer a distinct advantage over present scoring procedures which either employ directions that attempt to discourage guessing behavior and which may have an adverse effect for cautious examinees, or require the examinees to be trained in the uses of an alternate response mode every time the test is administered.

Data Collection

A series of three teacher-written algebra tests were administered to 12 sections of eleventh-grade students attending a suburban Philadelphia high school. The six participating teachers agreed to use the scores from these tests for grading purposes, and the students were so informed, thus insuring a conscientious effort. The examinees were directed to respond to each of the tests in two distinct

ways: using the Coombs response mode, which was used with appropriate directions during the first part of the testing period; and using the conventional response mode, which was used with directions that encouraged guessing during the second part of the test period. Two initial tests were designed to acquaint the students with the novel response mode and to provide feedback on their performance. Only the data from a third test (n = 230) were used for the experimental analysis. The test consisted of 20 multiple-choice items with four choices per item.

In addition to the scores resulting from the experimental test, final course grades and scores on the final examination were obtained for each examinee to be used as "external" validity criteria.

Data Analysis

Three different scoring procedures were used to score the "conventional" responses made during Part II of the test administration. First, number-right scores (NR) were computed by assigning a score credit of one to every item for which the correct choice was indicated and a score credit of zero to all other items. Second, a corrected for guessing score (NRC) was computed by subtracting from each examinee's NR score an amount equal to one-fourth of the number of items wrong. It should be noted that the NR and NRC scores are both derived from the responses made when the examinees were directed to indicate the correct choice with no penalty for guessing. Conse-

quently, the NRC scores do not reflect the influence of the directions not to guess that usually accompany formula scoring. Finally, the conventional responses were used as a basis for the proposed scoring system.

By considering simultaneously the responses made under both the conventional and Coombs response mode conditions, it was possible to compute several sets of scores, each based on the number of items answered correctly when the number of correctly guessed items is controlled. The number of choices among which guessing occurred is the distinguishing feature of these score sets. Guessing-free scores were obtained by assigning a score credit of one to an item if it was answered correctly, provided that all four distracters were identified. A score credit of zero was assigned to all other items. Thus, these guessing-free (GF) scores came only from items where it appeared the examinee knew the answer with a substantial degree of assurance.

A second set of scores was computed by assigning a score of one to all items from the GF score set and also to items for which successful guessing was limited to two choices (GF - 2 scores). Two more partially GF score sets were computed in an analogous manner for which successful guessing was limited at most to three and four choices (GF - 3 and GF - 4). It should be noted that the number of items answered correctly when guessing is free to vary among all choices is simply the number of items answered correctly, or the NR score.

The scoring rule proposed by Coombs (1953) was also used to score the elimination mode responses yielding yet another set of scores simply referred to hereafter as Coombs scores.

The experimental scoring procedure requires that scores be calculated on a set of variables for each examinee. The operational definitions for the six basic variables are presented in Table 1. The square of each of these variables and the cross-product between each pair were then computed. This resulted in a total of 27 variables. These variables were then used to predict the guessing-free and each of the partially guessing-free score sets outlined above. The forward selection program of the *Statistical Package for the Social Sciences* (Nie, Bent and Hull, 1970) was used for this purpose.

Because the proposed scoring procedure uses a multiple regression technique involving a large number of predictor variables, cross-validation of the predicted guessing-free scores was essential. To this end, the 230 answer sheets were randomly separated into two groups (groups A and B) and the b weights associated with the variables entering the prediction equation in each group were applied to the scores for the same variables in the alternate groups. The guessing-free and each of the partially guessing-free score sets served as the criterion for separate regression analysis.

The utility of the proposed scoring system for predicting each criterion was judged by comparing magnitudes of the correlation

coefficients between the cross-validated scores and each of the guessing-free criterion scores with the correlation coefficients between scores yielded by two conventional methods of scoring and the same criterion.

Since the expressed purpose of the proposed scoring system was to yield a set of scores free of a guessing component, the guessing-free (GF) score set was the criterion of central importance.

The proposed scoring system was also used to predict the two external validity measures directly. The correlation of the cross-validated scores with the final-course and final-examination grades was compared to the parallel correlations for NR and NRC score sets. Although there was no reason to expect a predictive superiority of the proposed system to predict these scores, it was of interest to determine the ability of the scoring variables to predict validity measures that exist in many practical testing situations.

Results

The means, standard deviations, and intercorrelations between the score sets generated from the experimental test are presented in Tables 2 and 3 for group A and for group B, respectively. Included in these tables are reliability estimates as well as the correlation of each test score set with the two external criteria. The matched-half reliability coefficients were computed by means of the Rulon formula

with a special splitting of the items to form halves that would measure, as nearly as possible, the same content area. There is no reliability estimate provided for the NRC scores. Because of the unusual format of the answer sheets, individual item scores could not be computed to provide a reliability estimate from a single test administration. Moreover, these scores reflect only the application of the correction formula and cannot be interpreted the same as corrected scores given with directions appropriate to them.

The descriptive statistics in these tables are presented to provide some insight into the nature of the scores being predicted or compared in the following sections.

The results of the regression analyses showed that the relative worth of the variables for predicting each of the criteria (GF, GF - 2, GF - 3, GF - 4) was quite different when compared across groups.

The cross-validated scores from the proposed scoring system were compared to the scores resulting from number-right and formula scoring of the same responses. The correlation of these scores with the GF and partially GF scores are presented in Table 4. In every case, the NR and NRC scores correlated more highly with each criterion than did the cross-validated scores.

The ability of the proposed scoring system to predict the two external score sets was compared to number-right and formula scoring as was done for the other criteria. The observed correlations are

presented in Table 5. Inspection of Tables 4 and 5 shows the correlation between the cross-validated scores and each criterion to be lower than the correlation between the NR or NRC scores with the same criteria.

Discussion

With the exception of one variable (per cent correct), each of the score variables used in the proposed scoring procedure is based on one of two basic statistics; namely, the proportion of examinees that selected each choice and the point-biserial correlation coefficient between the dichotomy of marking or not marking each choice and total scores on the test in which the choice in included. The use of such item/choice statistics to assign item scores is not novel. A scoring procedure proposed by Chernoff (1962) was based on item difficulty alone. The use of choice-total correlations is the basis of certain option-weighting procedures such as those investigated by Davis and Fifer (1959); Hendrickson (1971); Sabers and White (1969). It was thought that by including all of these choice statistics in arriving at a total score, a more effective scoring procedure would result than if just one such statistic was used. The results of this study indicate that such was not the case. It may be that the way in which these item statistics were combined in this study limited their utility for computing test scores. The choice difficulty, and discrimination coefficients associated with every choice marked by an examinee, were

summed across all items, and the mean values, for both correct and incorrect choices, were computed. For different examinees, these means were computed using different n's, depending on the number of items answered correctly and on the number of items omitted. If for discussion we can assume that more credit should be assigned when difficult items are answered correctly, a potential difficulty arises. Suppose an examinee answers every item correctly. His score for score variables V1 and V2 would be the mean correct-choice statistics for the test. However, if an uninformed examinee supplies a random guess to every item, and by chance correctly guessed, say two of the most difficult items, his score for variables V1 and V2 would be somewhat higher than the well-informed examinee. If one extends this type of thinking toward the middle range of ability, it seems reasonable that the first four score variables (V1, V2, V3, V4) may be greatly affected by chance and by the number of items over which they are computed. If there was a defensible way in which these choice statistics could be combined, and perhaps moderated by number-right scores, to assign individual item scores, perhaps more valid and reliable scores would result.

Independent of the proposed scoring system, it is of interest to consider the psychometric properties of the various guessing score sets generated in this study. Inspection of Tables 2 and 3 reveals successively higher matched-half reliability estimates for the partially-guessing-free score sets as they become more nearly

guessing-free. This finding may seem especially unusual in light of the fact that when guessing scores were computed from these same data, the split-half reliability estimates were significantly different from zero in most cases. However, the true-score model presented by Frary (1969a, 1969b) indicates that the reliability of the scores may increase or decrease when the guessing component is removed, depending on the correlation between the true and guessing components. In order that the reliability increase, this correlation must be negative and a given inequality must be satisfied. Use of the appropriate formulas presented by Frary (1969a) showed both of these conditions to be satisfied. These findings are therefore consistent with theoretical expectations and argue against the notion that score reliability can be expected to decrease when the effect of guessing is removed, even though the guessing component itself may be reliable. No systematic effect on validity was noted when the guessing component was removed as indicated by the correlation of the guessing-free and partially-guessing-free score sets with the two external criteria presented in the same tables. These data suggest that significant increases in reliability can result if score credit is assigned only to items that were answered correctly without guessing of any type. This is quite different from assigning score credit as required by the Coombs' scoring rule which did not appreciably affect the reliability or validity of the scores in this study. Of course, there is no

way to determine the number of items a student knows without imposing some type of penalty for guessing, or reward for not guessing. To advise students of a penalty for guessing and then delete the penalty in computing the scores (i.e., assign credit only to items for which complete knowledge is expressed), would be inappropriate. Consequently, these findings are only of theoretical interest at present.

At the very least, the markedly higher reliability estimates obtained for the guessing-free and partially-guessing-free score sets emphasize the potential for any scoring system that reduces or eliminates guessing.

A major assumption on which the proposed scoring procedure rests is that the alternate response mode can effectively eliminate score variance due to guessing. While this assumption holds a certain intuitive appeal, it may not be reasonable for the type of test used in this study. Most of the items used in the experimental test required the solution to an algebraic problem. The distractors represented incorrect solutions that were thought by the investigator to represent plausible errors. Consequently, if the student arrived at an incorrect answer that matched one of the choices, he may well have dismissed any doubts he had in the process of arriving at that solution and felt confident that his answer was right, since it was among the choices. In this case, he probably

would have marked all remaining choices. At the time he indicated his answer, he would not have *thought* he was guessing, even though he may have made several guesses in the process of arriving at his solution. If this hypothesis about the students' strategies is true, the various criterion scores which were thought by the investigator to reflect varying degrees of guessing may have been invalid as such. This possibility was perhaps less likely with items that did not require a solution to a problem.

REFERENCES

Chernoff, H. The scoring of multiple-choice questionnaires. *The Annals of Mathematical Statistics*, 1962, *33*, 375-393.

Collet, L.S. Elimination scoring: An empirical evaluation. *Journal of Educational Measurement*, 1971, *8*, 209-213.

Coombs, C. H. On the use of objective examinations. *Educational and Psychological Measurement*, 1953, *13*, 308-310.

Coombs, C. H., Milholland, J. E., and Womer, F. B. The assessment of partial knowledge. *Educational and Psychological Measurement*, 1956, *16*, 13-37.

Cross, L. H. An investigation of a scoring procedure designed to eliminate score variance due to guessing in multiple choice tests. Unpublished Ph.D. dissertation, University of Pennsylvania, 1973.

Davis, F. B. and Fifer, G. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, 1959, *19*, 159-170.

Diamond, J. J. and Evans, W. The correction for guessing. *Review of Educational Research*, 1973, *43*, 181-191.

Frary, R. B. Elimination of the guessing component of multiple-choice test scores: Effect on reliability and validity. *Educational and Psychological Measurement*, 1969, *29*, 665-680. (a)

Frary, R. B. Reliability of multiple-choice test scores is not the proportion of variance which is true variance. *Educational and Psychological Measurement*, 1969, *29*, 359-365. (b)

Hendrickson, G. F. The effect of differential option weighting on multiple-choice objective tests. *Journal of Educational Measurement*, 1971, *8*, 291-296.

Koehler, R. A. Coombs' type response procedures. Paper presented at the meeting of the American Educational Research Association, Chicago, Illinois, March 1972.

Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

Nie, N. H., Bent, D. H. and Hull, C. H. *Statistical package for the social sciences*. New York: McGraw-Hill, 1970.

Sabers, D. L. and White, G. W. The effect of differential weighting of individual item responses on the predictive validity and reliability of an aptitude test. *Journal of Educational Measurement*, 1969, *6*, 93-95.

TABLE 1

OPERATIONAL DEFINITIONS OF THE SIX BASIC VARIABLES

Variable Number	Variable Name	Definition
1	V1	The mean difficulty[a] of the correct choices marked by the examinee
2	V2	The mean discrimination[b] coefficient of the correct choices marked by the examinee
3	V3	The mean difficulty[a] of the distracters marked by the examinee
4	V4	The mean discrimination[b] coefficient of the distracters marked by the examinee
5	V5	The proportion of correct choices marked by the examinee
6	V6	The variance of the difficulty values for the correct choices marked by the examinee

[a]Difficulty is defined as the proportion of examinees who marked a particular choice.

[b]Discrimination coefficient is defined as the point-biserial correlation between marking or not marking a choice and total test scores uncorrected for overlap.

TABLE 2

DESCRIPTIVE STATISTICS AND INTERCORRELATIONS AMONG SEVEN SCORE SETS GENERATED FROM THE EXPERIMENTAL TEST (Group A)

	SCORE SETS						
	NR	NRC	COOMBS	GF	GF-2	GF-3	GF-4
NR		.996	.933	.895	.956	.971	.983
NRC			.946	.897	.955	.968	.980
COOMBS				.910	.953	.957	.952
GF					.951	.913	.911
GF-2						.984	.977
GF-3							.993
FE	.651	.668	.667	.648	.648	.652	.661
FG	.619	.642	.697	.653	.630	.636	.639
Mean	10.53	8.35	33.83	8.23	9.66	10.10	10.35
Standard Deviation	4.25	5.19	20.73	5.02	4.64	4.44	4.35
Split-half	.812		.815	.898	.859	.836	.814
KR-20	.790						
Alpha			.798				

TABLE 3

DESCRIPTIVE STATISTICS AND INTERCORRELATIONS AMONG SEVEN SCORE SETS GENERATED FROM THE EXPERIMENTAL TEST (Group B)

	SCORE SETS						
	NR	NRC	COOMBS	GF	GF-2	GF-3	GF-4
NR		.987	.954	.912	.962	.976	.978
NRC			.976	.928	.972	.986	.988
COOMBS				.943	.970	.967	.961
GF					.952	.930	.919
GF-2						.985	.975
GF-3							.996
FE	.623	.636	.633	.607	.606	.620	.620
FG	.640	.651	.642	.633	.625	.641	.645
Mean	10.66	8.45	34.18	8.58	9.96	10.37	10.48
Standard Deviation	4.54	5.52	22.05	5.17	4.97	4.72	4.71
Split-half	.841		.864	.904	.876	.854	.852
KR-20	.820						
Alpha			.827				

TABLE 4

CORRELATION OF SCORES RESULTING FROM THREE SCORING METHODS WITH THE GF AND PARTIALLY GF SCORE SETS

	CRITERION SCORES			
	GF	GF-2	GF-3	GF-4
	Group A			
Number-Right	.895	.956	.971	.983
Number-Right Corrected	.897	.955	.968	.980
Proposed System (Cross-validated)	.871	.945	.959	.970
	Group B			
Number-Right	.912	.962	.976	.978
Number-Right Corrected	.928	.972	.986	.988
Proposed System (Cross-validated)	.870	.948	.960	.975

TABLE 5

CORRELATION OF SCORES RESULTING FROM THREE SCORING METHODS WITH EACH OF THE EXTERNAL CRITERIA

Scoring Method	Group A		Group B	
	Final Exam	Final Grade	Final Exam	Final Grade
Number-Right	.6509	.6193	.6227	.6397
Number-Right Corrected	.6677	.6416	.6367	.6510
Proposed System (Cross-validated)	.5319	.4467	.5422	.4899