DOCUMENT RESUME

ED 060 070 TM 001 216

AUTHOR Rippey, Robert M.
TITLE Scoring and Analyzing Confidence Tests.
PUB DATE Apr 72
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April 1972

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Confidence Testing; Decision Making; *Guessing (Tests); Multiple Choice Tests; Research Design; *Scoring Formulas; Sex Differences; Student Motivation; *Test Interpretation; *Test Reliability; Test Validity

ABSTRACT

This paper examines confidence testing, and reasons for using confidence tests. Different scoring systems are studied in order to clarify the meaning of significance of the weights which subjects assign to confidence scored tests. (DLG)

SCORING AND ANALYZING CONFIDENCE TESTS

Robert M. Rippey

University of Illinois at Chicago Circle

Background

A conventional multiple choice item is a very special case of a much more ;eneral situation of decision .aking. Given a body of information, knowledge of a field, a limited number of options to choose among, a subject will in the conventional situation: consider the options, develop a set of preferences, and then make a single choice as instructed. Such a decision, however, is not particularly informative about the state of knowledge of an individual. The single, unqualified choice does not separate the confident subject from the timid one. Nor does it separate the lucky guesser from the qualified and certain expert. Furthermore, the search by test makers for questions having unique correct responses limits them to a small fraction of the possible questions which might be contrived in that broad area lying between warranted knowledge and aleatory opinion. Perhaps good guesses are as good as the same choice made with greater assurance. On the other hand, there may be some value in exploring more systematically some of the alternatives to dogamtic testing practices.

This additional information about confidence and distribution of belief may be important form several standpoints. It may lead to more accurate prediction of retention (Ahlgren, 1968). Furthermore, validity may be increased (Hambelton, Roberts, and Traub, 1970). Confidence information may be of importance in understanding the mechanisms involved in predicting performance in decision making involving complex sets of contingencies (Bruner, Goodnow, and Austin, 1956; Ward and Edwards, 1961; Kogan and Wallach, 1964).

Degree of belief in the options for an item may be represented by a set of weights. Although a set of weights contains more information than a single choice, the meaning of this additional information must be questioned, for these weights may mean different things to different subjects, and most certainly will mean different things depending upon the conditions of administration and the rewards or punishment expected as part of the testing procedure. It is perhaps because of uncertainty about the meaning of such weights that test makers have preferred almost exclusively items which had warranted unique answers. At least examples of tests of another sort are difficult to find.

By comparison with achievement tests, attitude and personality tests have eshewed unique right answers. The Strong Vocational Interest Blank, originally published in 1927, was an early example of an instrument involving response weighting (Strong, 1943). Others have followed (Swineford, 1941; Guttman, 1947). On the other hand, test makers have moved cautiously with uncertainty in the cognitive domain. Perhaps it is more acceptable to be uncertain how we feel than it is to be uncertain about what we know.

Although realizing that the area of uncertainty is always larger than the area of warranted knowledge, curriculum builders and testmakers have favored the certain side of the road, except in dire emergencies. It is interesting that much of the research in testing for uncertainty has come from the medical (Lewy & McGrire, 1966) and the military profession (Shuford, 1967).

In 1936, Soderquist suggested the application of penalties for misplaced confidence. Ebel (1965), in following this suggestion, obtained reliability increases of approximately .10.

A brief inroad into uncertainty was made by the Progressive Education Association in its Eight Year Study (Smith & Tyler, 1942). Items used in their investigation of outcomes of secondary schooling were not only classified as correct or incorrect, but also "Caution," "Insufficient data," "Beyond data," "Too certain," and "Too uncertain."

Dressel and Schmid (1935) contrasted the results obtained from a multiple choice test when administered in several novel forms. Neither of these systems actually assigned probabilities or degree of belief weights to the entire set of responses.

DeFinette (1965) has discussed a number of the consequences of utilizing different scoring systems and argues strongly for the training of subjects in a more prudent strategy for dealing with uncertainty, on the basis of utility.

> Feedback, or learning from experience, does not only concern the reinforcement of such general ideas about the profitability of an undistorted forecast or response corresponding to personal evaluation of probability: the evaluation itself becomes improved by experience. It is particularly common for untrained people to reflect in their numerical evaluations of probabilities the effect of the usual way of thinking in rough terms of 'certain,' 'impossible,' 'unknown,' or 'completely indifferent,' giving values 1.00 and 0.00 to the favored and rejected alternatives and 0.50 in the case of uncertainty between two, and so on. Experience forces them to realize how relatively often there are events which occur that can be too hastily classified as impossible, and they learn the advantage of giving these events an adequate small positive probability. It is chiefly because it provides the possibility of redressing such essential weaknesses in the machinery of human reasoning, and show how workable measurement in the fields of belief can be developed, that I have felt obliged to emphasize so strongly the desirability of training in the use of the methods described in this paper.

Shuford and Massengill take a different tack. They have argued that relaibility and validity of test can be increased using a class of scoring functions called reproducing scoring functions which maximize S's score if and only if his responses match his internal belief state. These functions were first studied by Toda (1963). The data available at this time does not conclusively support the viability of this model for responses of Ss. A purpose of the study herein reported was to clarify the meaning of significance of the weights which subjects assign to confidence scored tests.

A number of scoring functions have been developed for scoring confidence test items. Five such functions are shown on Table 1.

The probability assigned to the correct answer is the simplest and most intuitively obvious scoring function. Both the logarithmic and the spherical function possess the interesting property of allowing the student to maximize his score if and only if he does not guess (Shuford, Albert & Massengill, 1966). The Euclidean function will score items which do not have unique correct responses. Thus items can be constructed which call for answers which correspond to a distribution of preference representing the concensus of a group of experts. This function will also score items having unique correct answers. Inferred choice is analogous to conventional multiple choice scoring. According to this rule, the subject receives 1 point if his maximum confidence is assigned to the correct option. Otherwise, he receives nothing.

TABLE I

Five Scoring Functions

1. Probability assigned to correct response

$$S = r_k$$

2. Logarithmic

$$\text{If } r_k \geq .01,\ S = (2 + \log_{10}(r_k))/2$$

$$\text{If } r_k < .01,\ S = 0$$

3. Spherical

$$S = r_k / \sum_{i=1}^{n} (r_i)^2)^{\frac{1}{2}}$$

4. Euclidean

$$S = 1 - \frac{(\sum_{i=1}^{n} (r_i - k_i)2)^{\frac{1}{2}}}{\sqrt{2}}$$

5. Inferred Choice

$$S = 1 \text{ if } r_k > r_i \text{ for all } i \neq k;$$

$$S = 0 \text{ otherwise}$$

r_i = Probability assigned to the ith response

r_k = Probability assigned to the correct response

k_i = Criterion group mean probability assigned to the ith response.

Although this method of scoring simulates the performance of subjects on conventional multiple choice tests, it does not duplicate it. Some subjects, confronted by absolute lack of preference guess. Others do not. The inferred choice function simulates the behavior of the subject who never guesses when he is absolutely uncertain, but who always makes a choice, even if his preference is slight. Since there are varieties of subject behavior on tests, this function will not always give results which are identical to the choices an individual subject might make. If a subject is instructed to answer every question, and if all subjects do this, scores obtained by the inferred choice function would always be less than or equal to scores obtained by subjects under the usual choice situation, since subjects would occasionally get an additional point due to guessing. If subjects were told not to guess; and the conventional penalty for guessing were applied, scores obtained by the inferred choice function would be less predictable, and could be either greater than or less than the scores obtained by the subject responding in the conventional manner.

The seriousness of this discrepancy would be proportional to the number of instances on a test where no dominant preference was shown for a single option. In an analysis of a random sample of answer sheets for the STEP test data used in this study, such a lack of preference was found in less than 15% of the responses. Since the items were unusually difficult for the subjects by design, it is likely that the amount of guessing would be less on other tests, more appropriate in difficulty for the subjects.

On a three option test, this would suggest that scores would be approximately 5% higher in terms of S behavior as compared with inferred choice scores.

It would, of course, be possible to simulate all manner of erratic S behavior in responding to multiple choice items. However, the inferred choice function does simulate the subject who does not guess. Any other simulations, involving random or systematic awarding of points in guess situations would lead to less reliable scores than the inferred choice function. In comparing both the reliability and the validity of functions against the standard of choice, it is probably bestto use the inferred choice function as a standard since it does not contain any artificially induced error.

It would be possible, of course, to instruct S's to record both their probabilities and their choices. Thus, in the event of a split decision, the subject could flip his coin. Or the subject could also be instructed never to make his perference weights exactly equal. This should not be too unrealistic, since it is seldom, if ever, that the preference weights for a set of options would be entirely equal, no matter what the state of ignorance of S. Thus, although the inferred choice method does not simulate human behavior exactly, and this should be kept in mind, it is also unlikely that the inferred choice function produces less reliable or valid scores, or significantly lower scores, than would be given by scores obtained by conventional choice methods. Since the purpose of much research on confidence testing is to demonstrate the superiority of confidence methods over the conventional choice method, it seems that using the inferred choice function as a basis for comparison does not weaken the conclusions of such comparisons.

Although many arguments and some practice accept getting a fix on confidence, states of knowledge intermediate between certainty and chaos are not as readily accepted by some subject matter specialists. Therefore, when one examines achievement tests, it is unlikely that he will find many items dealing with incomplete information or uncertainty. The dearth of items requiring a distribution of belief over the available options may be due to a single technical consideration. Indeed, it can be argued that intrinsic

Items should not be written at all because an item which calls for a uniform distribution of confidence over all responses will not discriminate between the informed and the uninformed. Both groups would assign equal probabilities to all the options. Thus, the unweighted Euclidean function is inadequate by itself, for items not having single option responses are less efficient in detecting a state of no information than items having unique correct options. Nevertheless, this problem can be rectified by asking S for a distribution of belief and his confidence in his distribution, and subsequently incorporating both the distribution and the confidence measure into his score. The following function accomplishes this:

6. Weighted Euclidean function

$$S = C(1-2D/D_{max})$$

C = Confidence ($0 \leq C \leq 9$)
D = Distance from S's response to the criterion group response
D_{max} = Maximum distance attainable from the criterion group response

If Ss use confidence weights varying from 0 to 9, scores will vary from -9 to +9. An examinee who expresses no confidence will be neither rewarded nor penalized for his distribution of preference. On the other hand, certainty about a single incorrect response may suffer a nine-point penalty. The results of a test containing a mixture of items may be scored in at least three ways: correctness, confidence, and appropriate use of caution. The last measure will be developed in a later part of this report.

STUDIES OF DIFFERENT SCORING SYSTEMS

Subsequent to the development of the scoring programs several questions were asked. These were 1) How do the several scoring functions compare with one another and 2) Are there differences in the ways in which subjects respond to confidence scored tests?

One of the fundamental assumptions underlying the logarithmic function, and its desirable matching property of maximizing the subject's score if and only if he responds with his actual degree of confidence is the need for the subject to have feedback at test time with response to the payoffs of his set of preference weights.

Shuford has suggested several testing aids such as a computer terminal or computational devices known as scorules. Unfortunately, scoring in these ways is likely to be costly, or time consuming (Ebel, 1968). Furthermore, the necessity to provide the student with information about the scoring system is not only demanded by theory, but Rippey (1968) and Romberg and Shepler (1968) both provide data which shows that the logarithmic function may at times produce less reliable scores than conventional choice score. Furthermore, it offers the student some incentive not to guess, although not the optimal incentive promised by the logarithmic function. This leads to the question of the relative merits of various scoring functions with respect to reliability.

DESIGN

A total of 374 students, hereafter referred to as Sample A, from three Chicago suburban high schools were given three intact 30-item sections of the STEP Writing Test, Level 1[2], within schools. Tests were randomly assigned to students. No student took any form more than once.

Measures of SES and personality factors from the Personality Research Form (Jackson, 1965) were obtained for each student. The students were randomly divided into two groups - the incentive condition group was told that their scores on the test would count toward their grades while the other group was told that their scores would not.

Prior to taking the test, the subjects were given the following statement:

> Each of the questions or incomplete statements in this test is followed by suggested answers. Assign a number from 0 to 9

[2] Permission for the use of this test was granted by Educational Testing Service.

> To each suggested answer, depending on how strongly you feel the answer is correct. If you believe that only one suggested answer is correct, mark that answer with a 9 and mark the others with zeros. If you like the suggested answers equally, assign the same number to each.

Next followed several examples of how S was to distribute the numbers under various patterns of degree of belief of the correctness of the several options. Finally was the statement:

> Your paper will be scored in such a way that you will get a higher score by estimating your degree of confidence and reporting it accurately. Guessing in any form will lower your score. If you are uninformed about the question and have no preference for the suggested answer, you will obtain your highest score by honestly distributing your confidence across all the options....

Tests were subsequently computer scored and reliability was estimating using Hoyt's analysis of variance procedure (Hoyt, 1941). This procedure is suitable for confidence tests whereas a number of other procedures such as K.R. 20 are not. This is because item scores range between 0 and 1. The Hoyt method underestimates reliability on short tests. Therefore the reliabilities are all conservative.

Conclusions

1. Test reliability is proportional to the deviation of scoring function scores from simple methods which subjects anticipated.

2. The student is unfairly penalized by the scoring function which assigns him a score equal to the probability assigned to the right answer. Therefore, this function should only be used when minimal or no rewards are attached to subject performance. Otherwise the Eucidean function will not penalize the student, and will give high reliabilties in a no-feedback situation. In the event that feedback through tables, computers, or scoring aids is available, the logarithmic function is recommended.

3. On items not having unique correct responses, the weighted Euclidean function is only slightly superior to the unweighted with respect to reliability.

4. Under incentive conditions, scores on confidence tests are higher, and reliability lower.

5. Females have a greater tendency toward taking extreme positions of confidence than males, especially in the incentive condition.

6. Subjects in the incentive group liked the test better, had more of a tendency to take extreme positions, and made more appropriate estimates of their confidence.

7. Middle SES subjects, compared to both upper and lower SES subjects, made higher scores and more appropriate estimates of confidence. They seemed to be motivated more by desire for success than fear of failure.

8. High scoring subjects gambled more on difficult items under the relaxed condition, but gambled less on difficult items in the incentive condition.

9. Liking of tests was directly related to confidence.

10. There was no significant regression between confidence and the battery of personality variables, although high succorance and low harm avoidance made small contributions to prediction.

Much work remains to be done in studying confidence testing. Although it is clear that technical improvements may be made in the reliability and validity of tests through confidence scores, it is also clear that subjects do not handle their confidence uniformly. What is confidence to one may be hazard to anothe-. As Wang and Stanley (1970) state, The derivation of optimum response strategies in multiple choice testing represents an application of mathematical decision theory which underscores the decision process inherent in such tests. The success of testing procedures which attempt to control the decision process will be critically dependent on the ability of subjects to effectively use optimal strategies. It is not certain that all subjects are equally capable of learning to use such strategies.

The question of optimal strategies is likely to be perhaps the most significant outcome of further research on confidence testing. Although Bruner (1956) pointed out two basic differences in the way subjects use their confidences - the sentry condition and accuracy condition, and demonstrated empirical evidence of these two modes of behavior, there are other complex conditions which intervene between a subjective probability and a decision or action. Since it is possible, although not guaranteed that one may assess subjective probabilities accurately by means of reproducing scoring functions, two basic steps are needed. First, subjects need more experience in utilizing reproducing scoring functions. It takes a while to learn to respond intelligently to the rules of that game. Once it is possible to be confident of measures of subjective probability on a set of subjects, further study may be made of the use of optimal strategies by subjects in problematic situations. Such strategies would perhaps start with what is known about optimal search procedures in polychotomic trees (Watanabe, 1969). Although the ability to utilize optimal strategies, and the ability to make appropriate assessments of one's subjective probabilities is of value in its own right, it would perhaps be a useful next step to begin to apply information about subjecttive probabilities to the study of the structure of subject matter. This could be done through an analysis of the associative networks of highly trained subjects terms utilizing a system of analysis similar to that of Quillian (1968), substituting subjective probability in place of his all or nothing at all lines of association. Further development of such techniques, and further gathering of data on sophisticated human subjects may lead to the uncovering of most of the appropriate parameters involved in quiding decision making in problematic situations.

This goal is an ambitious one. Perhaps at a more realistic level would be the goal of increasing emphasis on the ways students react to problematic situations. Are students able to assess their state of information and respond intelligently to it? Do our teaching and testing practices make them aware that there are differences among the ways we use our information?

References

Ahlgren, A. Confidence marking and predicting retention. A paper read at the annual meeting of the American Educational Research Association, Chicago, Illinois, 1968 (mimeo).

Bruner, J.S., Goodnow, J.J. & Austin, G.A. A study of thinking. New York: Wiley & Sons, 1956.

Dressel, P.L. & Schmid, J. Some modifications of multiple choice items. Educational and Psychological Measurement, 1953. 13, 574- 595.

Definetti, B. Methods for discriminating levels of partial knowledge concerning a test item, British Journal of Mathematical and Statistical Psychology, 1965, 18, 87-123.

Ebel, R. Measuring Educatonal Achievement. New ork: Prentice-Hall, 1965.

Guttman, L. The Cornell technique for scale and intensity analysis. Educational and Psychological Measurement, 1947, 7, 247-279.

Hambelton, R.K., Roberts, D.M., & Traub, R.E. A Comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple choice test. Journal of Educational Measurement, 1970, 7, 75-90.

Jackson, Douglas Personality Research Form, Form AA, Research Psychologists Press Inc. 1965.

Kogan, N. & Wallach, M.A. Risk taking: A study in cognition and personality. New York: Holt, Rinehart, & Winston, 1964.

Lewy, A. & McGuire, C. A study of alternative approaches to estimating the reliability of unconventional tests. Paper delivered to the American Education Research Association, Chicago, Illinois, 1966 (mimeo).

Quillian, M.R. Semantic memory. In Minsky, M. (Ed.) Semantic Information Processing. Cambridge, Mass.: The MIT Press, 1968.

Rippey, Robert Probabilistic Testing, Journal of Educational Measurement, 1968, 5, 211-215.

Shuford, E.N.. Cybernetic testing. Bedford, Mass.: Decision Sciences Laboratory, Air Force Systmes Command, 1967.

Smith, E. & Tyler, E. Appraising and recording student progress. New York: Harper, 1942.

Strong, E.K. Vocational interests of men and women. Stanford, Cal.: Stanford University Press, 1943.

Swineford, F. Analysis of a personality trait. Journal of Educational Psychology, 1941, 32, 438-444.

Wang, M.W. & Stanley, J.C. Differential weighting: A review of methods and empirical studies, Review of Educational Research, 1970, 40, 663-705.

Watanabe, S. Knowing and guessing: A formal and quantitative study. New York: Wiley & Sons, 1969.