DOCUMENT RESUME

ED 056 098 TM 000 910

AUTHOR	Boldt, Robert F.
TITLE	A Simple Confidence Testing Format.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Air Force Human Resources Lab., Lowry AFB, Colo.
REPORT NO	RB-71-42
PUB DATE	Sep 71
NOTE	20p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Guessing (Tests); *Multiple Choice Tests; *Response Style (Tests); *Scoring Formulas; Statistical Analysis; Test Interpretation
IDENTIFIERS	*Confidence Testing

ABSTRACT

This paper presents the development of scoring functions for use in conjunction with standard multiple-choice items. In addition to the usual indication of the correct alternative, the examinee is to indicate his personal probability of the correctness of his response. Both linear and quadratic polynomial scoring functions are examined for suitability, and a unique scoring function is found such that a score of zero is assigned when complete uncertainty is indicated and such that the examinee can expect to do best if he reports his personal probability accurately. A table of simple integer approximations to the scoring function is supplied (Author)

RESEARCH BULLETIN

A SIMPLE CONFIDENCE TESTING FORMAT

Robert F. Boldt

Research sponsored by the Technical Training Division, Air Force Human Resources Laboratory, Lowry Air Force Base, Colorado, Contract No. F-41-609-70-C-0044. This report was issued by the Air Force as AFHRL-TR-71-31.

Educational Testing Service
Princeton, New Jersey
September 1971

A SIMPLE CONFIDENCE TESTING FORMAT

ABSTRACT

This paper presents the development of scoring functions for use in conjunction with standard multiple-choice items. In addition to the usual indication of the correct alternative, the examinee is to indicate his personal probability of the correctness of his response. Both linear and quadratic polynomial scoring functions are examined for suitability, and a unique scoring function is found such that a score of zero is assigned when complete uncertainty is indicated and such that the examinee can expect to do best if he reports his personal probability accurately. A table of simple integer approximations to the scoring function is supplied.

A SIMPLE CONFIDENCE TESTING FORMAT[1]

Test takers and test developers have long been aware that multiple-choice item format has certain presumed deficiencies. Among these is the presumed anxiety generated by the need to indicate either/or conclusions about the correctness of the item. There is a lack of ability of the scorer to differentiate between answers that are a product of knowledge and those that are largely a product of uncertainty. While it is true that the traditional methods work and have not as yet been improved upon in a way that demonstrably upgrades their utility to the score user, one might still be willing to accept some additional complication in mass processing if the testing process could be made more palatable to the examinee. One way to do this is to make some provision for the test taker to communicate the fact that he is uncertain to some extent of the correctness of the response that he is making. In this way and with a reasonable scoring procedure one can reassure the person tested that hesitant choices among responses will not incur large score differences. Thus the intensity of the conflict encountered in this risky decision situation should be reduced, and the testing process should become rather more comfortable. This, at least, is one kind of rationale for allowing the test taker to communicate his degree of uncertainty about his response.

Various ways of allowing for uncertainty have been made ranging from the garden variety formula score, which merely eliminates the advantage a guesser has if a rights-only score is used, to the more elaborate subjective probability methods introduced by De Finetti (1965), who requires that a scoring method oblige the examinee "to reveal his true beliefs, because any falsification will turn out to be to his disadvantage." Stanley (1968) has described a variety of methods allowing for uncertainty including those where the main motivation is to eliminate advantages due to guessing. These methods apparently do not always yield gains in reliability and do nothing for the expression of degrees of confidence. Confidence testing is discussed by Lord and Novick (1968) and studies of effects on reliability are summarized by Echternacht (1971). However, a method suggested by Dressel and Schmid (1953) is one which is virtually identical with that favored in this paper. A forerunner of the Dressel-Schmid format was introduced by Hevner (1932), and work using formats highly similar to that of Dressel-Schmid was done by Soderquist (1936), Wiley and Trimble (1936), Swineford (1938, 1941), Gritten and Johnson (1941), and Frederiksen, Jensen, and Beaton (1968). These studies are discussed by Stanley (1968) and by Echternacht (1971). They had the examinee mark the correct alternative and, in addition, assign a confidence weight, ranging from one to four in the case of Dressel and Schmid, in accordance with their degree of certainty as to the correctness of their choice. To anticipate later development in this paper, it may be noted that in a sense the present paper presents a scoring rationale and weighting scheme for the Dressel-Schmid confidence format, based on modern notions of subjective probability. It should be understood that the author is not endorsing the uncritical acceptance of confidence testing practices. It has its probable drawbacks, some of which are discussed in the last section of this paper. What is intended is that the use of the confidence testing be made easy, still retaining the desirable requirement of de Finetti as enunciated below.

Shuford, Albert, and Massengill (1966) have defined a "reproducing scoring system" in the spirit of de Finetti as follows: Let $\varphi_h(R)$ be a function of the vector, R, of responses to a multiple-choice item when alternative h is the correct one, and let p_i be the test taker's personal probability that the ith response is the correct one. In this vein R is a vector with nonnegative elements r_i which sum to unity, and the p_i are also nonnegative and sum to unity. The distinction is made that R is the vector of responses actually made which may or may not correspond to the p's. This lack of correspondence might arise through some idiosyncratic notions about test taking strategies, and the intent of the scoring system is to produce a situation where the subject can do his best by revealing the p's as accurately as he can. This is to be done by taking as an objective function the examinee's expected score, S, with respect to his own personal probability and choosing φ so that S is at a maximum when the r's equal the corresponding p's. That is, choose φ so that

$$S = \sum_h p_h \varphi_h(R)$$

is at a maximum when $r_h = p_h$ for all admissible sets of p's , and subject to constraints that the r's must be nonnegative and must sum to unity. Such a scoring system is called a "reproducing scoring system" by Shuford et al. because if the examinee does in fact knowingly behave so as to maximize S , his responses will reproduce his subjective probabilities.

Note that the functions φ_h have as arguments the elements of the vector R and hence require the recording of a response for each alternative. Thus the task of the examinee is to choose for each item a vector, R, by estimating the relative strength of one's subjective attitudes toward the alternatives or according to some personal strategy. This task may be too difficult for the examinee, and be carelessly done, and also may be prohibitively expensive to score. Hence the simplicity of the Dressel-

Schmid format, together with a rationale using subjective probability notions to develop the reproducing property, is appealing. Admittedly, the Dressel-Schmid format will not be fully reproducing since the entire vector R is not developed--only the largest element in R is recorded and therefore the term "quasi-reproducing" is used subsequently and refers to the reproduction by the response made of the corresponding underlying subjective probability. The utility of the format is also limited in that one does not expect more than minor increases in reliability from its adoption over the standard formula score. Its main advantage seems to the author to be its improved credibility and the attractiveness of the scoring rationale, i.e., the situation is structured so that the optimum strategy is the honest expression of the answer and one's confidence in its correctness. It is felt that there are situations, [illegible]larly those wh[illegible] test anxiety seems high, where these advantages may be compelling.

The present paper is concerned with a simplification wherein the examinee rates his response to only one alternative, the alternative rated indicating his choice of the best alternative and the rating indicating his degree of confidence in that alternative only. The response will be scored on whether the correct alternative was marked and how confident the examinee is in his choice. One would like a scoring scheme in which wrong opinions confidently expressed incur large penalties, frank guesses or near guesses are only mildly punished or rewarded if at all, and confidently expressed correct opinions are greatly rewarded. Two scoring functions will be used, one if the correct alternative is marked and another if the incorrect alternative is marked. Both will be monotonic functions of the level of confidence expressed and it will turn out that the scoring function for the correct alternative will be monotonically increasing while the scoring function for an incorrect alternative response will be monotonically decreasing.

If the level of confidence recorded is x, $f(x)$ is the scoring function if the correct alternative is marked, $g(x)$ is the scoring function if an incorrect alternative is marked, and p is the examinee's subjective probability that the response he marked is, in fact, correct; then the objective function becomes

$$S = p\ f(x) + (1 - p)\ g(x)$$

and one wishes to choose f and g so that S is at a maximum if x equals p for all admissible p. Various constraints can be imposed on the f and g yielding different scoring functions. In this paper linear and quadratic functions will be examined--if more requirements seem needed, higher order polynomials could be adopted.

The Linear Case

Assume $f(x) = ax + b$ and $g(x) = Ax + B$.

Then $S = p(ax + b) + (1 - p)(Ax + B)$.

Since in this case S is linear in x, it follows that x should take on an extreme value, since the function S has no relative minimum in the interval zero to one. Hence it is not possible to get a quasi-reproducing scoring system with a linear scoring function in the "pick one" format. To avoid forcing the candidate to express certainty when he does not feel so certain, set both a and A equal to zero. Further we set S equal to zero when p is one divided by the number of alternatives (the examinee has no preferred answer) because it seems reasonable to have an expected score equal to the omit score when complete uncertainty prevails. Omits will be given a zero so

$$S = (b/k) + (k - 1)B/k = 0$$

and

$$b = -(k - 1)B$$

where k is the number of alternatives. If we take b as positive, the examinee will respond to that alternative for which his subjective probability is the highest since he will have the most to gain (we are defining "good" scores on S as being in the positive direction). Since the response made is the one with the highest subjective probability and since the subjective probabilities must add up to one, it follows that the examinee who marks the answer with a confidence of $1/k$ is completely uncertain. That is, if the highest of a set of p's equals $1/k$, then $\Sigma p \leq k(1/k) = 1$. But, Σp must equal one so $p \geq 1/k$. Clearly the lowest possible value for p is $1/k$, and p takes this value only when all p's are equal, again because Σp must equal one. Hence the substitution of $(1/k)$ for p indicates correctly a state of complete uncertainty--the one we want to receive the same score as an omit. It remains only to take b as unity to yield the standard formula score. While this score is not quasi-reproducing, neither does it force the student to over- or underexpress his certainty when that certainty is elicited. The rather surprising result here is that if a linear scoring system were to be used, the confidence elicited should not be scored (a and A are zero). Further, since most writers agree that it is important to inform the examinee carefully about the scoring system, one would elicit the confidence response and then carefully inform the examinee that it would be ignored! It is concluded therefore that unless one is prepared to use nonlinear functions of the confidence expressed, one should not attempt to introduce confidence scoring.

The Quadratic Case

More useful results obtain in the case of the quadratic scoring function. Here we define

$$f(x) = bx^2 + cx + d$$

and

$$g(x) = Bx^2 + Cx + D$$

to obtain

$$S = p(bx^2 + cx + d) + (1 - p)(Bx^2 + Cx + D)$$

and choose b, B, c, C, d, and D so that S is at a maximum for all admissible p , and so that $f(1/k) = g(1/k) = 0$. It will be seen that these requirements impose five conditions on the six constants leaving an arbitrary choice of a sixth condition. For this condition we choose $f(1) = 1$. To maximize S, equate

$$dS/dx = 2pbx + pc + (1 - p)2Bx + (1 - p)C ,$$

evaluated at the point p to zero to obtain

$$dS/dx\Big|_p = p^2(2)(b - B) + p(c + 2B - C) + C = 0 .$$

Setting coefficients of the powers of p to zero, obtain

$$b = B,\ C = 0,\ c - C + 2B = 0.$$

Thus,

$$f(x) = bx^2 - 2b + d$$

and

$$g(x) = bx^2 + D.$$

Then $f(1/k) = 0$ implies that

$$d = -b/k^2 + 2b,$$

and $g(1/k) = 0$ implies that

$$D = -b/k^2.$$

Thus,

$$f(x) = b[x^2 - 2x - (1/k^2) + (2/k)]$$

and

$$g(x) = b[x^2 - (1/k^2)].$$

Note that

$$df(x)/dx = b(2x - 2)$$

which takes on the opposite of the sign of b since $2x$ must be less than 2 unless x equals or exceeds one (which it cannot). It is desirable to have the derivative of $f(x)$ with respect to x be nonnegative and therefore the sign of b should be negative.

If b is chosen to be negative, then $g(x)$ will be monotonically decreasing with increasing x ; and if x is not less than $1/k$, the reward for responding honestly to the subjectively most probable of the correct answers will always be greater than any other course of action provided the least certainty the examinee is allowed to express is complete uncertainty, that is $1/k$. This caution is introduced because under certain conditions the value of S will be greater if the examinee indicates a very small subjective probability for an alternative he is virtually certain is incorrect than if he marks an alternative he is moderately sure is correct. This possibility is to be avoided because it is relatively difficult to avoid having at least one bad distractor, and it will be shown that if allowed the examinee should mark the wrong distractor with a lower subjective probability than a right one, unless he is pretty sure it is right. To show this, suppose that the examinee is certain that an alternative is incorrect and he marks it zero. Then his payoff is

$$S_e = 0 \cdot f(0) + 1 \cdot g(0) = b(-1/k^2)$$

if according to his hypothesis he marks the wrong one zero. However, if we want him to mark an alternative that has a chance of being correct, his probability may be as low as $1/(k - 1)$ and according to his hypothesis his payoff would be

$$S_h = \frac{1}{(k - 1)} f\left(\frac{1}{k - 1}\right) + \left(\frac{k - 2}{k - 1}\right) g\left(\frac{1}{k - 1}\right)$$

$$= \frac{-b}{k^2(k - 1)^2} .$$

Clearly, $S_h = \frac{1}{(k - 1)^2} S_e$ and is less than S_e.

Therefore if the candidate knows the payoff system, he should in this case indicate that the erroneous distractor is incorrect rather than making the best guess he can about which alternative is correct. This can be avoided by limiting the range of responses he can make from $1/k$ to one since in this range,

$$S = pf(p) + (1 - p)g(p),$$

if p is used for x,[2] and has a first derivative equal to

$$dS/dp = 2(-b)(p - k^{-1})$$

which is clearly positive if b is negative.

Finally, for the sake of definiteness, we choose b so that $f(1)$ equals one. That is

$$1 = b(1 - 2 - k^{-2} + 2k^{-1}) \qquad \text{or} \qquad k^2 = -b(k - 1)^2$$

hence

$$f(x) = k^2(k^{-2} + 2x - x^2 - 2k^{-1})(k - 1)^{-2} \quad \text{and}$$

$$g(x) = k^2(k^{-2} - x^2)(k - 1)^{-2}.$$

Using Discrete Values

The intent of the above analysis is to arrive at a scoring function which is reproducing at least in the sense of eliciting an honest expression of confidence about the response made, which requires only a simple response from the examinee, and which is easy to process. A scoring system which requires that the response be recorded as a number for one or all alternatives requires data processing steps to get from the recorded response to a machine-processable record. These steps can be avoided using a discrete rating system, of which De Finetti has discussed a number.[3] By using the Dressel-Schmid format with a discrete multi-level confidence rating scale, one allows the examinee to make a very simple response which through mark sensing or optical scanning is directly available for quasi-reproducing scoring using digital processing.

Table 1, which could serve as a basis for choosing scores for discrete responses, contains the scoring system for common numbers of alternatives. Note that in this table the scores are not defined for confidence levels below 1/k. It can be seen that in all cases $f(x)$ has a positive slope and a negative acceleration. Since the two functions take on the same value when their argument equals 1/k, they diverge as x increases as does the risk of expressing an increased degree of certainty. However, note that the values of the objective function, S, are increasing as confidence increases so the examinee can indeed expect to be rewarded on the average by expressing

Insert Table 1 about here

his certainty when he feels it. Also contained in Table 1 are f(x), g(x), and S, for the limiting condition of k equals infinity (free response scored right or wrong).

It is felt that the scoring procedure can very well be approximate as long as some provision for expressing confidence is made and the scoring system in Table 1 is roughly reproduced. Hence the following method for obtaining scoring alternatives is suggested: (a) using five responses, describe verbally one extreme response as absolute certainty and the other as absolute uncertainty. Then the scoring for these extremes can be 0 and 10 (or 100), if the response is correct. If it is wrong, the scores are 0 and 10 (or 100) times the entry in Table 1 appropriate to the number of distractors; (b) state verbally that the middle categories represent equal intervals of uncertainty (or certainty) about the answer. If it's a "push," use the middle interval. If not, use one of the other two to show the strength of certainty. This kind of language may be taken as justification for assigning to the categories the scores from one-sixth, one-half, and five-sixths (the category midpoints if the interval is equally divided into thirds) of the distance from complete uncertainty to certainty.

For example, if a true-false test (k = 2) were given, the lower and upper category boundaries are .5 and 1 , respectively. Then the two middle category boundaries are

$$(1/k) + \frac{1 - (1/k)}{3} = .5 + \frac{1 - .5}{3}$$

and

$$(1/k) + \frac{2(1 - [1/k])}{3} = .5 + \frac{2\ (1 - .5)}{3} \quad .$$

The category midpoints are, then,

$$.5 + \frac{1 - .5}{6} ,$$

$$.5 + \frac{1 - .5}{2} ,$$

and

$$.5 + \frac{5(1 - .5)}{6} .$$

Table 2 gives the tabled weights. The true-false and free response scores are given on a correct score scale of 0 to 100, rather than 10, to avoid identical weights for different responses. Its entries can easily be displayed on an answer sheet or provided to an examinee as ancillary material. Finally, it can be adapted to a four-alternative answer sheet by instructing the examinee to omit the item if he has no preference among any of the alternatives.

Insert Table 2 about here

Perspective

Confidence testing seems to hold promise for the person who is concerned about certain anxiety-producing aspects of usual formula score testing. Certainly, one is inclined on the face of it to suppose that there is a difference in knowledge between persons who are confident that wrong answers are correct and those who express wrong responses diffidently. One is certainly very interested in finding some way to improve both the task of the examinee, as well as that of the one who must interpret his performance. Hence, the present paper presents a way of accomplishing confidence testing which it is hoped is relatively easy to use and which has an appealing rationale.

However, the sentiments expressed above are not intended to convey a belief that confidence testing in any format known to the author is in all situations anxiety reducing, nor would the author be willing to claim a reduction in anxiety by the use of the method in any given situation at the present time --nor an increase, for that matter. It has also been pointed out that confidence testing is not expected to make major increases in reliability or validity. In fact, Swineford (1938, 1941) presents evidence that the tendency to claim extra credit under conditions of risk is quite unrelated to other variables and suggests a possible contamination of scores based on confidence techniques due to irrelevant personality trends.

When converting from a standard multiple-choice test to a confidence format, one should at least consider the assessment of a response style with respect to risk in order to determine whether some allowance should be made for that style. However, response styles and personality factors may be operative under current testing modes as well as under confidence testing. It is not that one is "right" but that both may be used, and, when they are, possible moderation by personality scores could well be considered. And when such consideration is given, the method presented herein, being as simple to use and score as the author can make it, is recommended.

REFERENCES

De Finetti, B. Methods for discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 1965, 13, 87-123.

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

Echternacht, G. The use of confidence testing in objective tests. Research Bulletin 71-41. Princeton, N. J.: Educational Testing Service, 1971.

Frederiksen, N., Jensen, O., & Beaton, A. E. (Contr. by Bloxom, B.). Organizational climates and administrative performance. Research Bulletin 68-41. Princeton, N. J.: Educational Testing Service, 1968.

Gritten, F., & Johnson, D. M. Individual differences in judging multiple-choice questions. Journal of Educational Psychology, 1941, 32, 423-430.

Hevner, K. A. A method of correcting for guessing in true-false tests and empirical evidence in support of it. Journal of Social Psychology, 1932, 3, 359-362.

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

Shuford, E. H., Albert, A., & Massengill, H. E. Admissible probability measurement procedures. Psychometrika, 1966, 31, 125-145.

Soderquist, H. O. A new method of weighting scores in a true-false test. Journal of Educational Research, 1936, 30, 290-292.

Stanley, J. C., & Wang, M. D. Differential weighting, a survey of methods and empirical studies. New York: College Entrance Examination Board, 1968.

Swineford, F. Measurement of a personality trait. Journal of Educational Psychology, 1938, 29, 295-300.

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

Wiley, L. N., & Trimble, O. C. The ordinary objective test as a possible criterion of certain personality traits. School and Society, 1936, 43, 446-448.

FOOTNOTES

[1]This research was supported by the Technical Training Division, Air Force Human Resources Laboratory, Lowry Air Force Base, Colorado. This report was issued by them as AFHRL-TR-71-31.

[2]Since it is known that the scoring system is quasi-reproducing, it is proper to use p instead of x as arguments of f and g.

[3]The task of the subject in the Dressel-Schmid format is like that of method B-1 of de Finetti except that more confidence levels are allowed. The scoring rationale here is also different.

Table 1

Scores[a] for Common Numbers of Alternatives as a Function of Expressed Confidence Levels

% Confidence Level (100x)	Number of Alternatives: 2			3			4			5			[illegible]		
	f(x)	-g(x)	S	f(x)	-g(x)	S	f(x)	-g(x)	S	f(x)	-g(x)	S	f(x)	-g(x)	S
0													0	0	0
5													0.1	0.0	0.0
10													0.2	0.0	0.0
15													0.3	0.0	0.0
20										0	0	0	0.4	0.0	0.0
25							0	0	0	0.1	0.0	0.0	0.4	0.1	0.1
30							0.0	0.05	0.0	0.2	0.1	0.0	0.5	0.1	0.1
33 1/3				0	0	0	0.2	0.1	0.0	0.3	0.1	0.0	0.6	0.1	0.1
35				0.0	0.0	0.0	0.25	0.1	0.0	0.3	0.1	0.0	0.6	0.1	0.1
40				0.2	0.1	0.0	0.4	0.2	0.0	0.4	0.2	0.1	0.6	0.2	0.2
45				0.3	0.2	0.0	0.5	0.25	0.1	0.5	0.25	0.1	0.7	0.2	0.2
50	0	0	0	0.4	0.3	0.1	0.6	0.3	0.1	0.6	0.3	0.1	0.75	0.25	0.25
55	0.2	0.2	0.0	0.5	0.4	0.1	0.6	0.4	0.2	0.7	0.4	0.2	0.8	0.3	0.3
60	0.4	0.4	0.0	0.6	0.6	0.2	0.7	0.5	0.2	0.75	0.5	0.25	0.8	0.4	0.4
65	0.5	0.7	0.1	0.7	0.7	0.2	0.8	0.6	0.3	0.8	0.6	0.3	0.9	0.4	0.4
70	0.6	1.0	0.2	0.8	0.9	0.3	0.8	0.8	0.4	0.9	0.7	0.4	0.9	0.5	0.5
75	0.75	1.25	0.25	0.9	1.0	0.4	0.9	0.9	0.4	0.9	0.8	0.5	0.9	0.6	0.6
80	0.8	1.6	0.4	0.9	1.2	0.5	0.9	1.0	0.5	0.9	0.9	0.6	1.0	0.6	0.6
85	0.9	1.9	0.5	0.95	1.4	0.6	1.0	1.2	0.6	1.0	1.1	0.7	1.0	0.7	0.7
90	1.0	2.2	0.6	1.0	1.6	0.7	1.0	1.3	0.7	1.0	1.2	0.8	1.0	0.8	0.8
95	1.0	2.6	0.8	1.0	1.8	0.9	1.0	1.5	0.9	1.0	1.35	0.9	1.0	0.9	0.9
100	1	3	1	1	2	1	1	1 2/3	1	1	1 1/2	1	1	1	1

[a]Where decimals are given, rounding is to the nearest low order position. Figures without decimals are exact.

Table 2

Approximate Scores for Responses to

Confidence Items on Dressel-Schmid Format

	Credit if Right					Loss if Wrong				
Category k[a]	2	3	4	5	α[a]	2	3	4	5	α[a]
Absolutely Certain	100	10	10	10	100	300	20	17	15	100
Certain	95	9	9	9	88	225	14	12	11	43
Middle Certain	75	7	7	7	75	125	7	5	5	25
Somewhat Uncertain	35	3	3	3	28	20	2	2	1	2
Completely Uncertain	0	0	0	0	0	0	0	0	0	0

[a]One hundred point scale used to avoid duplication from rounding.