DOCUMENT RESUME

ED 058 307 TM 001 022

AUTHOR	Echternacht, Gary
TITLE	The Use of Confidence Testing in Objective Tests.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	RB-71-41
PUB DATE	Sep 71
NOTE	32p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Confidence Testing; *Guessing (Tests); Individual Characteristics; Measurement Techniques; Multiple Choice Tests; *Objective Tests; Predictor Variables; Response Mode; *Scoring Formulas; Testing Problems; Test Interpretation; *Test Reliability; Weighted Scores

ABSTRACT

Confidence testing has been used in varying forms over the past 40 years as a method for increasing the amount of information available from objective test items. This paper traces the development of the procedure from Hevner's beginning method up to the various methods in use today and describes both the testing procedures and scoring methods used. The term confidence testing is applied to both probabilistic testing and confidence weighting procedures. Various procedures are presented and their relationship with personality factors discussed. (Author)

RB-71-41

RESEARCH BULLETIN

THE USE OF CONFIDENCE TESTING IN OBJECTIVE TESTS

Gary Echternacht

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
September 1971

THE USE OF CONFIDENCE TESTING IN OBJECTIVE TESTS

Gary Echternacht

Abstract

Confidence testing has been used in varying forms over the past 40 years as a method for increasing the amount of information available from objective test items. This paper traces the development of the procedure from Hevner's beginning method up to the various methods in use today and describes both the testing procedures and scoring methods used. The term confidence testing is applied to both probabilistic testing and confidence weighting procedures. Various procedures are presented and their relationship with personality factors discussed.

THE USE OF CONFIDENCE TESTING IN OBJECTIVE TESTS

Gary Echternacht

Lord and Novick (1968) have stated that the general problem of obtaining the maximum amount of information from a given set of items contains three major components. The first of these is the measurement procedure, or the manner in which the examinees respond to the items. The second component is the specification of an item scoring rule or formula that is used for each item. The final component is the combination of the item scores into a total score by an item weighting formula. The first two components comprise much of the subject of confidence testing. Confidence testing is a method of testing where weights are assigned directly or indirectly to item responses in such a way as to reflect the examinee's belief in the correctness of the alternative or alternatives so marked. One author (Jacobs, 1971) has made a distinction between confidence weighting procedures and probabilistic testing. This paper considers these two procedures as categories of the more general subject of confidence testing.

The purpose of this paper is to describe the various forms of confidence testing as they have been developed and provide a brief evaluation of these forms. Some of the problems associated with confidence testing are discussed with the relevant studies cited.

In the usual multiple-choice test, an examinee is given a question along with a number of possible answers to that question. He is then asked to choose an answer from those given and to indicate that choice on an answer sheet. If he chooses the correct answer, he receives a score of one; if he omits the item, he gets a score of zero; and if he guesses incorrectly or is misinformed as to the correct answer, he receives a score less than or equal to zero, as with the commonly used formula score. The total test score is then taken as the sum of the item scores.

This type of testing has the advantages of efficiency and simplicity for both the examinee and the test scorer. More items can be administered in a given period of time using this method than by any other method requiring a more complicated response, and the cost for scoring the test is also less.

Advocates of confidence testing have stated that knowledge is neither a dichotomous nor a trichotomous affair, which conventional multiple choice tests seem to imply, but is continuous in the sense that there are varying degrees of knowledge. Some contend that confidence testing discourages guessing since the scoring systems for some confidence testing systems are such that an examinee can maximize his expected score only if he reveals his true degree of certainty in responding. Lord and Novick (1968) have pointed at the strong conceptual attractiveness as being the sole recommendation for confidence testing.

In evaluating confidence testing, it is necessary to show that the procedure adds more ability variation to the system than error variation and that any increase in the amount of information gained is, in fact, worth the effort. Various kinds of costs must be kept in mind. On one hand, the case may be that the examinees are available for a long period of time and the test items are very difficult to obtain. In this case we would like to get as much information as possible from each item which confidence testing might be able to provide. On the other hand, examinee time may be scarce and items easy to come by, in which case it is probably better to increase the information by adding more test items.

Confidence Weighting Procedures

When a confidence weighting procedure is used, the examinee is asked to indicate what he believes the correct answer to be and how certain he is of the correctness of his answer. This procedure should not be confused with the personal probability approach discussed later where the examinee assigns weights to each option in accordance with the confidence he has in the correctness of each option.

Although confidence testing has a history that dates back to the early part of the century (Henmon, 1911; Hollingworth, 1913; Trow, 1923), it first began to be tried as a method for increasing the amount of information in objective tests during the decade of the 1930's. The process came about as a method for minimizing the effect of guessing on true-false type tests. Hevner (1932) reported a study evaluating the degree to which confidence testing improved the reliability of tests in aesthetics and music appreciation. True-false tests were used where the examinees chose one of two pieces of music as being more musical and then indicated their degree of confidence in their judgments on a three-point scale. Four systems of scoring were examined for differences in reliability: (1) the number right; (2) the number right minus the number wrong; (3) a weighted right-answer score, where right answers were counted as three if the highest confidence mark was indicated, two if the middle confidence mark was indicated, and one if the lowest mark was indicated; and (4) a weighted-right minus a weighted-wrong score, using the weights previously mentioned. The weighted-right method of scoring resulted in the highest reliability using the Spearman-Brown formula for a test of double length. Hevner noted that subjects in her experiment welcomed the addition of the opportunity to express a degree of confidence, especially those who felt insecure about the test. She also felt the necessity of keeping the scoring formula secret from the examinees in the case of the weighted-right method so that dishonest subjects could not artificially raise their scores.

A similar study by Soderquist (1936) described what was not a new procedure for confidence testing, but rather a different scoring system from that of Hevner. Using Soderquist's system, a student could claim special credit for an answer by indicating his confidence, the special credit amounting to four, three, two or, if no special credit was claimed, one score point. Scoring consisted of the weighted-right score minus the weighted-wrong score, where the weights for the wrong responses were double the amount of credit claimed. A true-false examination was administered using this confidence format, and the students were told of the scoring systems to be used. Tests were scored using both the weighted-right minus weighted-wrong scores and the number right minus the number wrong, ignoring the confidence responses; reliabilities were calculated using the Spearman-Brown formula on random split halves. Soderquist found a higher reliability for the weighted-right minus weighted-wrong score, a finding somewhat different from Hevner's (1932), but this difference was possibly due to the fact that Soderquist's subjects knew the scoring system used while Hevner's subjects did not. He further stated that personality factors may have influenced the result of his experiment and urged further experimentation.

Wiley and Trimble (1936) undertook a study that attempted to establish the existence of personality factors in the confidence testing procedures used earlier by Hevner (1932). When examining the correlations between four achievement tests on scores derived by counting the number of times each confidence level was used and the formula score, correlations for the confidence scores were higher than those for achievement. They concluded that personality variables could be measured in this manner, although they did not mention which personality variables were operating in the situation under study.

Swineford (1938) did identify one such personality variable as she administered a true-false test under the directions used by Soderquist. She was able to derive what she termed a "gambling" score from the test responses that was sufficiently reliable and yet independent of the right minus wrong achievement score. She subsequently (Swineford, 1941) used four other tests to measure the tendency to gamble, concluding that boys tended to gamble significantly more often than did girls, both boys and girls had a tendency to gamble more on unfamiliar material than on familiar material, and gambling scores tended to be independent of the achievement test scores. More recently, Ziller (1957) suggested an alternative method for determining a similar score from similar data.

Jacobs (1971) has questioned the use of confidence weighting on the grounds that the scoring procedure tends to be contaminated by individual differences in personality. For other studies using confidence testing in personality research see Kogan and Wallach (1964) and Slakter (1967, 1968).

Gritten and Johnson (1941) used a method of confidence testing though their objective was to relate degree of confidence in one's response to instructions of whether to guess or not. One significant aspect of the confidence testing procedures was that Gritten and Johnson used a multiple-choice test and asked for confidence responses on a five-point scale, a forerunner of the "Pick-One" confidence format later used by Echternacht, Sellman, Boldt, and Young (1971).

Interest in confidence testing seemed to diminish during the remainder of the 1940's and no significant contributions were found in the literature until the 1950's.

Dressel and Schmid (1953) experimented with various item types and scoring systems in an effort to improve the discrimination of multiple-choice items without extending the testing time. They used four modifications of conventional multiple-choice testing: (1) a free-choice test, where examinees marked as many answers as they thought were correct; (2) a degree-of-certainty test, where examinees indicated on a four-point scale their certainty in a single answer selected; (3) a multiple-answer test, where any number of alternatives might be correct and the examinee was to mark each correct alternative; and (4) a two-answer test, where exactly two of the five alternatives were known to be correct. In each case the examinees were aware of the type of testing taking place. Since only the first two experimental testing types related to confidence testing, only the scoring for the first two types will be given. The scoring for the free-choice and degree-of-certainty tests can be summarized as follows:

Free Choice

I. If the correct answer to an item was marked and the number of answers marked was	the item score was
1	4
2	3
3	2
4	1
5	0

II. If the correct answer to the item was not marked and the number of answers marked was the item score was

of answers marked was	the item score was
1	-1
2	-2
3	-3
4	-4

Degree of Certainty

If the correct answer was marked and the certainty value marked was the item score was

was	the item score was
positive	4
fairly sure	3
rational guess	2
no defensible choice	1

If the correct answer was not marked and the certainty value marked was the item score was

was	the item score was
positive	-4
fairly sure	-3
rational guess	-2
no defensible choice	-1

Dressel and Schmid (1953) found that superior students, defined in terms of traditional test scores, differed significantly from average and poor students when using the free-choice format, the difference being that high performers marked fewer answers across each of three different levels of item difficulty. It was also found that the administration time was lengthened for all students. The degree-of-certainty method, on the other hand, differentiated superior, average, and low-ability students about equally well, the confidence marks being about the same for both average and difficult items. It was also concluded that the certainty factor measured by the free-choice item was not the same as that measured by the degree-of-certainty item.

It is interesting to note that the mode of examinee response used by Dressel and Schmid (1953) was essentially identical to Gritten and Johnson's

(1941) earlier method, the only difference being that Dressel and Schmid used a four-point scale rather than a five-point scale. Also, in the same volume of Educational and Psychological Measurement, Coombs (1953) described a method of confidence testing that was similar in intent to Dressel and Schmid's free choice. Although Coombs' method may be more accurately classified as a method of differential weighting rather than confidence testing, it is included here as it has influenced other confidence testing authors to a considerable extent.

Coombs (1953) was concerned more with the assessment of an examinee's partial information rather than assessing a degree of confidence measure. He made the psychological assumption that partial knowledge exhibited itself in recognizing some wrong answers. The examinee was instructed to cross out each of the alternatives he believed to be incorrect, rather than indicate all of the possibly correct alternatives, as Dressel and Schmid (1953) instructed. For each wrong alternative correctly crossed out, one score point was granted, but if a correct answer was crossed out, a penalty of 1 - k (k being the number of alternatives) was given.

Coombs, Milholland, and Womer (1956) used Coombs' method on 855 juniors and seniors in high school. They found an increase in reliability equivalent to a 20 per cent increase in the length of a conventional test of the same type. The authors pointed out that as the difficulty of the test increased, the reliability of this test procedure also increased. The same items tended to discriminate well when administered as multiple-choice and experimental formats. Students involved in the study preferred the experimental tests to the multiple-choice tests and thought them to be more fair, as indicated by their questionnaire responses.

Archer (1962) compared both the free-choice and the elimination type of confidence testing with conventional testing, using reliability and validity criteria. He found the two confidence measures to be only slightly more reliable than the conventional and the conventional to correlate higher with the criterion (student's rank within his class) than either of the two confidence methods. He also reported that reliability per unit of testing time favored the conventional, while the elimination method proved more reliable on difficult items. The rank-order relationship of item difficulties under the different response methods was high.

Under the general topic of confidence testing in a beginning psychological measurement textbook, Ebel (1965b) described what amounts to a modification of Soderquist's (1936) method using true-false items with a more refined scoring system. Ebel's motivation, like that of the early pioneers in the field, was to reduce the error in testing due to guessing. His items were placed in a true-false format, each presenting a statement and asking the examinee to indicate the truthfulness or falsity of that statement. The examinee had five possible responses; those responses, with the scoring, were as follows:

Response Number	Significance	Right	Wrong	Omit
1	The statement is probably true	2	-2	
2	The statement is possibly true	1	0	
3	I have no basis for response			.5
4	The statement is possibly false	1	0	
5	The statement is probably false	2	-2	

Ebel (1965a) gave reliability data for three different classroom tests. The reliabilities were improved by the use of his scheme in each case. The improvement factors expressed in terms of how many times as long a conventional test would have to be in order to be as reliable as the original test were 1.84, 1.48, and 1.72 for each of the classroom tests.

Ebel (1965a) concluded that confidence testing could be effective if the more capable students were also more discriminating in choosing their level of confidence responses although experimental evidence suggested that this was not necessarily so. A general attitude of confidence, uncorrelated with achievement, was also found, as did Swineford (1938), as a factor affecting the examinee's score. To neutralize this effect, Ebel suggested specifying in advance the proportion of answers that must be given confidently.

The confidence weighting methods described thus far have been of two types: (1) an examinee indicates his answer and then indicates his confidence in that answer, and (2) an examinee indicates any number of answers as being plausible or not plausible. The methods of scoring these procedures are rather arbitrary in nature with the characteristic that the right answer given confidently is given more credit than a wrong answer given without confidence. A personality variable, uncorrelated with achievement defined by the traditional scoring methods, has been found in the systems and that, combined with the required test administration time, has proven to be the major obstacle in the adoption of those techniques.

Personal Probability Approaches

During the early 1960's, the notion of subjective probability made its intrusion into the behavioral sciences. The entry of the concept into the field of confidence testing resulted in the development of what has often been termed probabilistic testing. The concept of subjective probability served as a basis for probabilistic testing through work abroad and in the United States sponsored principally by the Department of Defense.

Before going any further, something should be said about subjective or personal probability. The notion of probability in general can be represented in four different ways. Some may think of probability in terms of relative frequencies in a population of trials with heuristic limiting arguments (Cramer, 1946), others may think of probability as an axiomatic theory (Kolmogorov, 1933), others as logical probabilities (Keynes, 1921), and still others as personal probabilities. De Finetti, Jeffreys, Lindley, and Savage have played a prominent role in the development of personal probability over the past years. Personal or subjective probability is defined to be a psychological attribute of an individual who is forced to make a judgment by consideration of a wager; that is, the amount p which he is indifferent to betting on an event in order to win $1 - p$. He is willing to bet that amount, but is indifferent between making the bet or not betting at all.

Although these ways of thinking of probability are conceptually different, fortunately, they all agree on certain basic properties and all arrive at the same definition of conditional probability. From there on the development is very much the same for each view.

De Finetti (1965) applied his knowledge of subjective probability to some of the problems of objective testing. He examined the problem from the examinee's point of view by posing the question of how an individual should behave when required to choose one of r permitted alternative answers to a test question. If the examinee were certain of the correct answer, the best response was just that, and the problem disappears; but oftentimes, especially if the item was difficult, the examinee had a degree of uncertainty about his action. The wise examinee wanted to respond in the most advantageous way in the face of uncertainty of a given action. To do this, de Finetti advocated the use of decision theory in light of the examinee's degree of belief; that is, personal probability.

Six preliminary assumptions were made which constituted the underlying philosophy of the approach. These were:

1. The scoring method and permitted modes of responding must be known to the subjects, the subjects fully understanding the implications in the face of uncertainty.
2. The subjects must be keenly interested in scoring high.
3. The subjects must be trained to understand the correspondence between their own belief and the numerical probabilities to which these are translated.
4. The total knowledge and belief of a given subject about a question and its alternatives must be expressed and fully represented by numerical probabilities he attaches to each of the alternatives.
5. The scores using any scoring method can be divided so as to determine the partial information of a subject from his responses.
6. The evaluation of this procedure should concern how well the scoring method describes the subject's beliefs and its value to him.

Up until this time most of the scoring methods for confidence testing were quite arbitrary. De Finetti (1965), primarily a statistician, attempted to provide some rationale for behavior as he presented and discussed various scoring schemes.

The continuous method of scoring was presented as the most powerful. The examinee was to write down his personal probabilities attached to each alternative. The score was constructed so that the subject must reveal his true beliefs since any falsification would turn out to his disadvantage if he were unsuccessful in his falsification. The score S was given for an r alternative item by

$$S = 2p_h - \sum_{d=1}^{r} p_d^2 \quad ,$$

where p_d was the probability assigned by the examinee to the d^{th} alternative, and alternative h was the correct answer. In every case the score was maximal where all the probability was concentrated on the correct answer ($p_h = 1$) and minimal when all the probability was concentrated on one wrong alternative ($p_r = 1$, r/h).

De Finetti (1965) recognized that assignment of exact probabilities to the alternatives was a very difficult task for the examinee taking the test, and he therefore discussed a wide variety of alternate methods. The other methods discussed did not involve the subject's recording his personal probability for each alternative or, for that matter, any single alternative. What De Finetti did try to do was to estimate a subject's personal probability under different response schemes. These estimates were made in terms of ranges and intervals.

Some of the more practical methods discussed were the purely rank-order methods, where the subject might rank all or some of the alternatives as to their correctness or mark a number of alternatives as being favorable and crossing out others as being unacceptable. Flexible schemes with two or three permitted levels of response were also presented, such as a method of mark one or none, crossing out as many alternatives as wished (i.e., Coombs' (1953) method) and various combinaticns of marking and crossing out freely.

A new system that may be credited to de Finetti (1965) was developed and termed a strict least-distance method. By least-distance methods de Finetti meant those derived as simplifications of the continuous method he first presented. For example, rather than requiring the examinees to record their exact subjective probabilities, examinees might be restricted to a finite set of probability responses such as multiples of .2 units of probability. Strict and flexible least distance methods were distinguished, strict methods being such that the probabilities used by the examinees summed to one. The most notable system was termed the five-star system and was used with five-alternative choice items. Using this system, the examinee was given five stars or weights and was asked to place them on the alternatives in such a way that the weights indicated his relative strength of belief about the alternatives. Complete sureness in an alternative would require that all five stars were assigned to that alternative; complete ignorance would require one star allocated to each alternative. An item score scale was set with range from 0 to 25 so that only small integers and no fractions entered in the scheme. It was assumed that one star was worth .2 units of probability. Table 1 shows the scores obtained for different numbers of stars attached to the correct alternative.

Insert Table 1 about here

For example, if an examinee places four stars on one alternative and one star on another alternative, the item score would be given by type 4-1. If he placed the four stars on the correct alternative, he would receive a score of 24. If he placed only one star on the correct alternative, he would receive a score of 9. On the other hand, if he placed no stars on the correct answer, he would receive a score of 4. The general procedure is to identify the distribution of stars used by the examinee and locate that distribution with respect to the type. Then determine the number of stars assigned to the correct answer and locate the score corresponding to that number.

The importance of de Finetti's (1965) contribution was that he introduced a high degree of mathematical sophistication into the subject of confidence testing and based his method on assumptions of examinee behavior. It also marked the entrance of decision theory and personal probability into the area of confidence testing. He realized thaτ examinees had to be prepared for this new type of test and urged an easing-in process in which the examinee could gradually grasp the notion of attaching a quantity to his belief in the truth of the alternative. Winkler (1967a,b) has discussed the problem of the subject's ability to assess personal probabilities in the face of uncertainty.

The major drawback of this study, for which de Finetti (1965) cannot be honestly criticized since he was writing theoretically only, was a lack of consideration for operational and psychological factors. The scoring table appears to be extremely complex, especially for the examinee who is not likely to understand fully the consequences of his responding. Nothing is mentioned of the lengthening of test administration time, time required for hand scoring, the ability of the examinees to understand the directions, or the various extraneous psychological factors possibly contaminating the score. Some recognition was made of the fact that it may not be in the best interest of the examinee to answer honestly; for example, if every student answered honestly, the ranking of test scores would approximate the ability ranking for whatever factor the test was measuring. If only the top 10 per cent of the examinees were to be rewarded, a low ability examinee would find it only to his advantage to falsify his probabilities in order to fall in the upper 10 per cent based on the test scores.

The impetus for the confidence testing procedures developed in the United States was provided by the Air Force at the Decision Sciences Laboratory.

Groundwork for the development of this type of testing was done by Toda (1963), who experimented with various scoring schemes including what is known as the quadratic and logarithmic scoring schemes, and also by Roby (1965), who discovered a spherical scoring system. It is also noteworthy that Van Naerssen (1961) independently discovered both the quadratic and logarithmic scoring functions.

The problem of confidence testing was discussed by Shuford, Albert, and Massengill (1966). Their objective was to extract a larger portion of the available information from objective test items. For them, this information was contained in the student's degree-of-belief probabilities or personal probabilities concerning the correctness of the various possible answers. They recognized that to measure these probabilities successfully a scoring system must be devised so that any student, whatever his level of knowledge or skill, could maximize his expected score if, and only if, he honestly reflected his personal probabilities. Scoring systems that made use of this property and were understood by the examinees were termed admissible probability measurement procedures. The more commonly used measurement procedures were not admissible according to the authors.

Shuford et al. (1966) further introduced the concept of a scoring system with a reproducing property; a scoring system was reproducible when the personal probabilities possessed by the examinee were identical to the probabilities with which he responded. They derived necessary and sufficient conditions for the reproducibility of a test item with two possible alternatives. They further showed the class of reproducible scoring systems to be virtually inexhaustible and demonstrated a method of construction. Under most circumstances it was thought desirable to have a scoring system so that an examinee's score did not depend on the position of the correct alternative in the order of choice; thus, symmetric reproducing scoring systems were presented with quadratic and spherical examples.

The question was then considered of constructing a reproducing scoring system in which the examinee's score depended only on the probability assigned to the correct answer and not on the probabilities assigned to the distractors. After some consideration, one such scoring scheme emerged which was termed logarithmic. The logarithmic function did have one difficulty, though, because the values of the scoring function were unbounded and thus impossible to use in practice; that is, when an examinee assigned a probability of zero to what was the correct answer, his score was minus infinity.

There did appear to be an approximate solution. Shuford et al. (1966) suggested using a truncated logarithmic scoring function. In particular

$$g(r_k) = \begin{cases} 1 + \log_{10} r_k & \text{for} \quad .01 < r_k \leq 1 \\ -1 & \text{for} \quad 0 \leq r_k \leq .01 \end{cases}$$

where r_k was the probability assigned to the correct answer. The authors concluded their paper by discussing the case of the fill-in type items providing methods for valid confidence testing using these item types. In a more recent development Winkler (1969) has shown that a logarithmic payoff function is necessary if the subjective probabilities serve both to keep the examinee honest and evaluate the examinee. Shuford et al. further showed that the logarithmic scoring function was the only such system when more than two alternatives were considered.

Shuford and Massengill then left the Decision Sciences Laboratory and formed the Shuford-Massengill Corporation where their prime interest was in improving confidence testing procedures. A kit was developed for what they term "valid confidence testing" which included answer sheets, scoring tables, class analysis forms, and a SCoRule response aid. The SCoRule was a mechanical device used by the examinee to aid him in marking his personal probabilities. It used the truncated logarithmic scoring function truncated at .1 rather than at .01, as had been earlier suggested. The examinee determined his personal probabilities by manipulating the length of various lines for each alternative--the longer the line the greater the probability for that corresponding alternative. Having completed this, the examinee then recorded, on his answer sheet, numbers from the SCoRule corresponding to the lengths of the lines. These numbers were the scores for the item, given the particular alternatives were correct.

Ebel (1968) has reviewed the kit and criticized the complexity of the task, estimating the administration time to be about double that of standard testing. He also was critical of the fact that no direct evidence of increases in reliability and validity using the process was presented. Ebel concluded by acknowledging the logic of the procedure and stated that valid confidence testing does, in fact, provide more information per item, but he questioned whether that gain in information justified the increases in costs, administration time, and scoring time.

Shuford and Massengill have tried their procedure a number of times using personnel from the Air Force (Massengill & Shuford, 1969; Shuford & Massengill, 1968, 1969). The validity of this procedure was discussed (1969) in terms of item confidence distributions. The authors concluded that about 4/5 of the students were able to complete the test in the required time limits using confidence testing. Scatterplots were also made plotting the percentage "Z" correct (the percentage of correct responses to an item where complete confidence was placed on one response) against the regular multiple choice score. Increases in reliability have been noted along with increased correlations among subtests of various test batteries.

Each of Shuford and Massengill's studies seem to suffer from the same basic deficiency; that is, lack of control. Their studies usually consisted of taking a group of subjects, administering a traditional multiple-choice test as a confidence test, plotting the percentage "Z" correct against the indexed multiple choice score, obtaining a scatterplot for the ratio of the number of times an item was correct for one of 26 intervals and the number of times that interval was used correctly versus the hypothetical proportion under true examinee subjective probability responses, and, finally, obtaining scatterplots for valid confidence score versus inferred choice score. The inferred choice scores were found by using the alternative with the highest probability assigned to it as an inference of what the examinee would have indicated had he been taking the test in a multiple-choice format.

The findings in Shuford and Massengill's studies usually demonstrated the percentage "Z" answers correct to be higher than the inferred expected percentage correct answers. This was not an unexpected finding, as examinees who took a test consisting only of items about which they were sure would be expected to have a higher percentage correct than examinees who took those same items in addition to other items about which they were unsure of the answer.

The scatterplots of the above-mentioned ratios versus the hypothetical proportions (termed external validity graphs) did not show that confidence testing yielded increased information as they had contended. Those scatterplots demonstrated only how realistic a given subgroup of examinees' response probabilities were, a test of de Finetti's (1965) third assumption.

When confidence test scores were plotted against inferred test scores, confidence test scores were, in general, higher which caused the authors to conclude that more information was present in confidence tests. If the expected score using the truncated logarithmic scoring function under the condition of complete ignorance is compared to the expected score under the same conditions for rights only scoring, the confidence test expected score is higher than the rights only expected score. From this, one could conclude that confidence test scores were higher due in part to the scoring scheme used rather than information increased.

Another deficiency in Shuford and Massengill's studies was the lack of a large sample size. The sample size for the study at the Air University was reported to be 26, while the reported sample size for the Officer Training School was 96. Those sample sizes may have been constraints on the experiments imposed by the Air Force, but their size should point to the need of further research. Other discussions of both Shuford and Massengill's and De Finetti's confidence systems appear in Stanley and Wang (1970).

In a study aimed at investigating the contribution of the psychological factors in probabilistic testing, Hansen (1971) studied the relationship between the degree to which examinees display certainty in their responses and certain personality variables. Hansen was able to derive an index of an individual's tendency to show certainty which was related to scores on both a modified F-Scale and Kogan and Wallach's (1964) Choice Dilemmas Questionnaire. Subjects displayed a characteristic tendency to be either certain or uncertain which was relatively stable from one exam to the next and which could not be fully accounted for on the basis of the stability of knowledge. This tendency further appeared to be only slightly related to the knowledge possessed by the subjects. Hansen finally concluded that training with the confidence system did not improve the accuracy of the scoring system as Shuford and Massengill have claimed.

Michael (1968) has also experimented with confidence testing in the form of a personal probability approach. She required subjects to allocate 10 points to the various alternatives of a given question and scored the item by using the proportion of points assigned to the correct alternative. She obtained both higher reliabilities and lower standard errors when the process was used. The major advantage in her approach was the ease in scoring and the ease with which the directions can be understood.

Rippey (1968) experimented with scoring probabilistic tests by logarithmic, spherical, and Euclidean scoring schemes noting that increases in reliability were not found automatically. Probabilistic-scored items were found to provide a different type of information that was attributed to the personality factor of general confidence. In a later study (Rippey, 1970), the simple scoring system used by Michael (1968) was advocated on the basis of high reliability and ease in scoring.

Hambleton, Roberts, and Traub (1970) compared a form of confidence testing similar to Shuford and Massengill's (1968) with a form of a priori differentially weighted alternative items and traditional multiple-choice testing. The response format used was similar to Michael's allocating 10 points. Confidence testing was found to yield the most valid and least reliable scores.

Boldt (1971) has devised a method of testing that combines a confidence weighting procedure with a subjective probability based scoring system. The method, termed Pick-One confidence testing, required the examinee to first choose the alternative that he believes most likely to be correct from a list of k alternatives. He then records that alternative and indicates on a five-point scale his sureness of his answer. Although the examinee does not respond with actual personal probabilities, the directions for this type of testing imply personal probabilities for each level of sureness possible.

The scoring for this type of confidence testing is based on subjective probability and depends only on the number of alternatives. Two constraints are placed on the scoring system, the score is 0 where the lowest sureness level is indicated and the maximum item score is 1. Two scoring functions are utilized, both quadratic, one for the case when the alternative chosen is correct, the other for the case when the wrong alternative is chosen. A scoring table for four values of k (k indicating the number of alternatives) and various subjective probability levels is given in Table 2. Various probability levels are given to illustrate the function. It is understood that it is the duty of the test constructor to create examples that illustrate the correspondence between the points and the probabilities.

Insert Table 2 about here

One primary advantage of the Pick-One confidence testing method is that it is adaptable to standard machine scorable answer sheets for scoring by a digital computer thus enabling the procedure to be used in mass testing cases. The technique also has a simple format for the examinee to understand and is relatively easy to score by hand when the scores are rounded.

The procedure is designed so that the test administrator has a degree of control over the levels of subjective probability used. He must provide directions to the examinees that give them an intuitive grasp of the probability level associated with each level of sureness. This has the advantage of flexibility of response modes, yet may cause some confusion both to the examinee if he becomes accustomed to this type of testing and possibly to the test administrator who must rationalize his selection of various response probabilities.

The number of scale points allowable is variable, the scoring system being independent of the number of scale points. Five points were used in the referenced study, this number being chosen for adaptation to machine scorable answer sheets.

On the negative side, the technique does not require responses for the alternatives not chosen, thus some information of the degree of belief in the distracters is lost. One can only conclude that the degree of belief in any other alternative is less than or equal to the approximate probability indicated. Another disadvantage lies in the fact that different scoring functions must be used when items with varying numbers of alternatives are included in the test due to the constraint that the omission and random response scores be identical.

Conclusions

In many of the studies conducted to evaluate confidence testing, the criterion of increase in reliability has been used for evaluation. It should be pointed out that this is not necessarily a particularly good criterion. Where tests have been established with reasonable standards of reliability, the desire is not to increase the reliability but to shorten the test and keep the same level of reliability. Confidence testing is not required to do this. For example, consider an achievement test in history where multiple-choice items are used with rights only scoring. If the test is reduced in length by considering only a homogeneous group of questions, say questions about one topic in history, the reliability of the test is likely to remain the same or even increase although we can recognize that this has probably not improved the test in terms of validity.

A final evaluation of confidence testing must weigh the gain in ability variation against the gain in error variation. The studies of Jacobs (1971), Hansen (1971), and Swineford (1938, 1941) lead one to conclude that there is a personality factor of "general confidence" operating in the confidence testing procedure which contaminates the results yielding an increase in error variation. Shuford and Massengill (1969) have claimed that this factor can be eliminated or at least considerably reduced with practice using their SCoRule device although there exists little data to support such a claim. The question remaining to be answered is whether this personality effect is greater for one particular procedure or another. Can a system be modified to reduce the effect of "general confidence"? If confidence testing does supply greater information about examinees, can that information be used for better selection, placement, or diagnosis? How is the increase in effort for scoring worth the increases in reliability or decreases in the number of items required for the test? Might the technique be better adapted to computer testing? What are the examinee's attitudes toward the process, especially one who feels uncomfortable in the standard test taking situation? What is the relationship between the examinee's confidence test score and his true score?

References

Archer, N. S. A comparison of the conventional and two modified procedures for responding to multiple-choice items with respect to test reliability, validity, and item characteristics. Unpublished doctoral dissertation, Syracuse University, 1962.

Boldt, R. F. A simple confidence testing format. AFHRL-TR-71-31. Lowry AFB, Colo.: Technical Training Division, Air Force Human Resources Laboratory, 1971. (Also Research Bulletin, RB-71-42, Princeton, N. J.: Educational Testing Service, 1971.)

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

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

Cramer, H. Mathematical methods of statistics. Princeton, N. J.: Princeton University Press, 1946.

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.

Ebel, R. L. Confidence weighting and test reliability. Journal of Educational Measurement, 1965, 2, 49-57. (a)

Ebel, R. L. Measuring educational achievement. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1965. (b)

Ebel, R. L. Valid confidence testing-demonstration kit. Journal of Educational Measurement, 1968, 5, 353-354.

Echternacht, G. J., Sellman, W. S., Boldt, R. F., & Young, J. D. *An evaluation of the feasibility of confidence testing as a diagnostic aid in technical training*. AFHRL-TR-71-33. Lowry AFB, Colo.: Technical Training Division, Air Force Human Resources Laboratory, 1971. (Also Research Bulletin, RB-71-51, Princeton, N. J.: Educational Testing Service, 1971.)

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

Hambleton, 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*, 75-82.

Hansen, R. The influence of variables other than knowledge on probabilistic tests. *Journal of Educational Measurement*, 1971, *8*, 9-14.

Henmon, V. A. C. The relation of the time of a judgment to its accuracy. *Psychological Review*, 1911, *18*, 186-201.

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.

Hollingworth, H. L. Experimental studies in judgment. *Archives of Psychology*, 1913, *29*, 1-119.

Jacobs, S. S. Correlates of unwarranted confidence in responses to objective test items. *Journal of Educational Measurement*, 1971, *8*, 15-19.

Keynes, J. M. *A treatise on probability*. London, England: Macmillan, 1921.

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

Kolmogorov, A. N. *Grundbegriffe der wahrscheinlichkeitsrechnung*. Berlin: Springer, 1933. (English translation by N. Morrison. *Foundations of the* theory of probability. New York: Chelsea, 1956.)

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

Massengill, H. E., & Shuford, E. H. Confidence testing at the academic instructor course of the Air University: August and September 1968. Lexington, Mass.: Shuford-Massengill Corporation, 1969.

Michael, J. J. The reliability of a multiple-choice examination under various test-taking instructions. Journal of Educational Measurement, 1968, 5, 307-314.

Rippey, R. M. Probabilistic testing. Journal of Educational Measurement, 1968, 5, 211-215.

Rippey, R. M. A comparison of five different scoring functions for confidence tests. Journal of Educational Measurement, 1970, 7, 165-170.

Roby, T. B. Belief states: A preliminary empirical study. ESD-TDR-64-238. Bedford, Mass.: Decision Sciences Laboratory, L. G. Hanscom Field, 1965.

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

Shuford, E. H., & Massengill, H. E. Final report of work performed under contract number AF 49(638)-1744 and under amendment number 1. Lexington, Mass.: Shuford-Massengill Corporation, 1968.

Shuford, E. H., & Massengill, H. E. Confidence testing at the officer training school, Lackland Air Force Base: September 1968. Lexington, Mass.: Shuford-Massengill Corporation, 1969.

Slakter, M. J. Risk-taking on objective examinations. American Educational Research Journal, 1967, 4, 31-43.

Slakter, M. J. The effect of guessing strategy on objective test scores. Journal of Educational Measurement, 1968, 5, 217-222.

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. Weighting test items and test item options, an overview of the analytic and empirical literature. Educational and Psychological Measurement, 1970, 30, 21-35.

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, 348-444.

Toda, M. Measurement of subjective probability distributions. ESD-TDR-63-407. Bedford, Mass.: Decision Sciences Laboratory, L. G. Hanscom Field, 1963.

Trow, W. C. The psychology of confidence. Archives of Psychology, 1923, 65, 1-47.

Van Naerssen, R. F. A scale for the measurement of subjective probability. Acta Psychologica, 1961, 19, 159-166.

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.

Winkler, R. L. The assessment of prior distributions in Bayesian analysis. Journal of the American Statistical Association, 1967, 62, 776-800. (a)

Winkler, R. L. The quantification of judgment: Some methodological suggestions. Journal of the American Statistical Association, 1967, 62, 1105-1120. (b)

Winkler, R. L. Scoring rules and the evaluation of probability assessors. Journal of the American Statistical Association, 1969, 64, 1073-1078.

Ziller, R. C. A measure of the gambling response-set in objective tests. Psychometrika, 1957, 22, 289-292.

Table 1

Scores Obtained for Different Numbers of

Stars Attached to Correct·Alternative

Type	Distribution of Stars	Corresponding Probabilities	Score	Type	Distribution of Stars	Corresponding Probabilities	Score
5	* * * * *	1.0	25	2-2-1	* *	0.4	18
		0.0	0		* *	0.4	18
					*	0.2	13
						0.0	8
4-1	* * * *	0.8	24				
	*	0.2	9				
		0.0	4				
				2-1-1-1	* *	0.4	19
					*	0.2	14
3-2	* * *	0.6	21		*	0.2	14
	* *	0.4	16		*	0.2	14
		0.0	6			0.0	9
3-1-1	* * *	0.6	22	1-1-1-1-1	*	0.2	15 } five
	*	0.2	12		*	0.2	15 } times
	*	0.2	12			0.0	10
		0.0	7				

Table 2

Scoring Table for Pick-One Confidence Testing

Probability Corresponding to the Selected Alternative	Alternatives							
	2		3		4		5	
	If Correct	If Wrong	If Correct	If Wrong	If Correct	If Wrong	If Correct	If Wrong
.2							0	0
.25					0	0	.12	-.04
.3					.03	-.05	.23	-.08
.333			0	0	.21	-.09	.31	-.11
.35			.04	-.01	.25	-.11	.34	-.13
.4			.18	-.11	.36	-.17	.44	-.19
.45			.32	-.21	.46	-.25	.53	-.25
.5	0	0	.44	-.31	.56	-.33	.61	-.33
.55	.19	-.21	.54	-.43	.64	-.43	.68	-.41
.6	.36	-.44	.64	-.56	.72	-.53	.75	-.50
.65	.51	-.69	.72	-.70	.78	-.64	.81	-.60
.7	.64	-.96	.80	-.85	.84	-.76	.86	-.70
.75	.75	-1.25	.86	-1.02	.89	-.89	.90	-.82
.8	.84	-1.56	.91	-1.19	.93	-1.03	.94	-.94
.85	.91	-1.89	.95	-1.38	.96	-1.17	.96	-1.07
.9	.96	-2.24	.98	-1.57	.98	-1.33	.98	-1.20
.95	.99	-2.61	.99	-1.78	1.00	-1.49	1.00	-1.35
1.00	1.00	-3.00	1.00	-2.00	1.00	-1.67	1.00	-1.50