DOCUMENT RESUME

ED 428 076 TM 029 476

AUTHOR	Kurz, Terri Barber
TITLE	A Review of Scoring Algorithms for Multiple-Choice Tests.
PUB DATE	1999-01-21
NOTE	21p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 21-23, 1999).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Algorithms; Guessing (Tests); *Item Response Theory; *Multiple Choice Tests; Scores; *Scoring; Validity
IDENTIFIERS	Partial Credit Model; Partial Knowledge (Tests)

ABSTRACT

Multiple-choice tests are generally scored using a conventional number right scoring method. While this method is easy to use, it has several weaknesses. These weaknesses include decreased validity due to guessing and failure to credit partial knowledge. In an attempt to address these weaknesses, psychometricians have developed various scoring algorithms. This paper provides an overview of the different scoring algorithms that correct for guessing and award credit for partial knowledge. Included in the overview is an explanation of the scoring formulas as well as a brief summary of the literature regarding the utility of each algorithm. Formula scoring methods and formula scoring with Item Response Theory are discussed. The following methods for awarding credit for partial knowledge are also reviewed: (1) confidence weighting; (2) answer-until-correct scoring; (3) option weighting; (4) elimination and inclusion scoring; and (5) multiple-answer scoring. (Contains 21 references.) (Author/SLD)

Running head: SCORING ALGORITHMS FOR MULTIPLE-CHOICE TESTS

A Review of Scoring Algorithms for Multiple-Choice Tests

Terri Barber Kurz

Texas A&M University 77843-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, January 21, 1999.

Abstract

Multiple-choice tests are generally scored using a conventional number right scoring method. While this method is easy to use, it has several weaknesses. These weaknesses include decreased validity due to guessing and failure to credit partial knowledge. In an attempt to address these weaknesses, psychometricians have developed various scoring algorithms. This paper provides an overview of the different scoring algorithms which correct for guessing and award credit for partial knowledge. Included in the overview is an explanation of the scoring formulas as well as a brief summary of the literature regarding the utility of each algorithm.

A Review of Scoring Algorithms for Multiple-Choice Tests

Multiple-choice tests are the most common format for measuring cognitive ability. This format is favored by both testing organizations and classroom teachers because these tests provide broad content sampling, high score reliability, ease of administration and scoring, usefulness in testing varied content, and objective scoring. Also, this format has great versatility in measuring objectives from the rote knowledge level to the most complex level (Sax, 1989).

One major benefit of using a multiple-choice format is the ease in scoring. These tests have traditionally been scored using a conventional number-right scoring method. Items on the test are dichotomously scored, with a value of 1 given to correct responses and a value of 0 given for incorrect responses (including blank or omitted items). With this method all items are weighted equally.

However, this method, while simple to use, has been criticized for certain weaknesses. Abu-Sayf (1979) named four such weaknesses. The first is the psychometric argument that the encouragement of guessing in the directions introduces a relatively great proportion of error variance and that the formula does not take into account partial knowledge. The pragmatic argument is that examinees who are more cautious and omit unknown answers are penalized in comparison with risk-takers. The moral argument is based on the notion that it is wrong to guess and that it is even less commendable to reward this behavior. The political argument is that encouraging examinees to guess results in having them lose their confidence in multiple-choice tests.

Dichotomous scoring has been criticized mainly because it fails to provide a direct estimate of the amount of knowledge examinees have. Most multiple-choice tests provide only ranking information. This is especially problematic in high stakes tests such as those that use test scores to discriminate between candidates who wish to enter a program or obtain licensure.

The response to these weaknesses has been to provide alternative scoring algorithms that can avoid some of these problems. These alternative scoring strategies attempt to overcome the weaknesses of conventional scoring and extract information from the examinees that would provide better estimates of their abilities. These scoring methods include those that discourage guessing and those that award partial credit for partial knowledge. In theory, these methods would increase the validity and reliability of test scores and benefit those examinees who have been penalized for not being risk takers or who are less test wise.

This paper provides an overview of the different scoring algorithms which fall under the major categories of correction-for-guessing formulas and formulas which award partial credit for partial knowledge. Included in the overview of each category will be an explanation of the scoring formulas as well as the advantages and disadvantages that have been found when studies were conducted on the different scoring formulas.

Correction for Guessing Formulas

Classical Formula Scoring

Correction for guessing formulas represent an attempt to assess examinees' true level of knowledge by eliminating from their scores correct responses that resulted from random guessing (Jaradat & Tollefson, 1988). The correction for guessing formula takes into account three possible situations: (a) the examinee knows the correct option and chooses it, (b) the examinee omits the item, or (c) the examinee guesses blindly and selects one of the responses at random. This assumption rules out the possibility that examinees sometimes answer on the basis of partial information or from misinformation (Rowley & Traub, 1977).

Proponents of the correction for guessing formula argue that this method should increase the reliability and validity of scores because the corrected score should be a better estimate of the examinee's knowledge. Because X values (i.e., the uncorrected score) are affected by random guessing, its sampling variance should be greater than the sampling variance of S (i.e., the corrected score) and the estimator with the smaller variance would be preferred since the scores are considered unbiased estimators of the same parameter (Crocker & Algina, 1986).

A correction for guessing formula is based on the assumption that all incorrect responses result from guessing. There are two models that use correction for guessing formulas. The first is the random-guessing model. This model rewards the examinee for not guessing by awarding points for omitted items. This is based on the assumption that if the examinee had attempted the omitted item, the incorrect response would be a random guess. The score formula is denoted:

$$S = R + O/k$$

where S is the corrected score, R is the number of correct answers, O is the number of omitted items, and k is the number of alternatives per item.

Abu-Sayf (1979) pointed out that strictly speaking this formula does not "correct" for guessing but aims at discouraging guessing by offering rewards for omissions. Proponents of this method believe that the psychological impact of an incentive such as the promise of a reward elicits a more favorable response in getting examinees to avoid wild guessing than does the threat of a penalty.

The more commonly used model is the rights minus wrongs correction model. This model penalizes the examinee for guessing by depriving the examinee of the number of points which are estimated to have been gained from random guessing. This is based on the assumption that each incorrect response is the result of a random guess. This formula score is denoted:

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

where S, R, and k are defined as above, and W is the number of incorrect answers.

Although these two formulas yield different numerical values, the rank order of the examinees' scores will be identical regardless of the formula that is used. In other words, if the two formulas are applied to the same set of item responses, the results will be perfectly correlated (Crocker & Algina, 1986).

Proponents of formula scoring have pointed out several advantages of this method. The principal advantage is that it discourages random guessing which can falsely inflate an examinee's score. As a result, it provides a better unbiased estimate of true knowledge based on test performance.

Some studies have indicated that correction formulas show a slight increase in validity and similar or slightly higher reliability than number right scoring. Both Diamond and Evans (1973) and Abu-Sayf (1979) have reviewed the studies addressing validity and reliability in formula scoring. While these studies tend to show increases in validity and reliability, the magnitude of the effect was very small. Lord (1975) indicated that the difference in reliability findings were due to the difference in test directions. In response to Lord's assumption, several studies were done (e.g., Bliss, 1980; Cross & Frary, 1977), however, the assumption was not proven to be correct. What these studies did find, however, was that the examinees did not behave according to the formula scoring assumptions.

Several studies have shown that examinees who are low risk takers are penalized by the formula scoring instructions (Albanese, 1988; Angoff, 1989;

Bliss, 1980; Cross & Frary, 1977; Slakter, 1968). These studies have shown that examinees who are high risk takers tend to ignore the formula scoring directions that discourage guessing and take the chance that they will guess the right answer. This increases their chances of a higher number of correct answers.

Suppose that Mike and Sarah are taking a 20-item multiple-choice test with 3 alternatives per item in which both know the answers to only 10 of the items. Sarah follows the directions and omits the 10 items she does not know. Following the formula S = R - W/k, her score will be 10 (10-0/3=10). Mike, on the other hand, guesses at the 10 questions he does not know. Mike gets 4 more answers correct through guessing. His score will be 12 (14-6/3=12). Therefore, Mike is rewarded for not following directions and guessing at the answers that he did not know. This type of outcome is exactly what formula scoring was attempting to eradicate.

Albanese (1988) found that some examinees would increase their scores by one half standard deviation or more if they answered omitted items. He also found that formula scoring slows examinee progress, leading to an increased number of trailing omits in speeded tests. In a critique of the instructions for formula scoring, Budescu and Bar-Hillel (1993) found that instructions were worded based on the theory that every test taker is an ideal test taker. An ideal test taker is one whose goal is to maximize the expected score and whose subjective probabilities are well calibrated. Since all test takers are not ideal test takers, the formula scoring instructions may not benefit an examinee who would be classified a real test taker. Formula scoring instructions are also more difficult to understand and may therefore unduly penalize low-ability examinees.

These studies point to the more serious criticisms of formula scoring. Other criticisms are that formula scoring fails to take into account partial knowledge, that the scoring formula leaves more room for computational miscalculation, and that there is a potential for negative public relations due to the penalty for guessing. Considering the criticisms of formula scoring, its use is rarely justified, except in high stakes testing situations with bright and sophisticated test takers.

Formula Scoring and Item Response Theory

In 1980 Lord described how the concept of formula scoring may be considered in estimation of true scores for tests developed with item response theory (Crocker & Algina, 1986). Item response theory is based on the probability that an examinee with ability level, θ, will answer an item correctly. An examinee's true score may be estimated by summing up these probabilities over all items. Lord indicated that this practice may need to be modified if examinees have differentially omitted items. He suggested that a number-right true score for examinee *a* could be determined by the following process:

1. Identify all items which examinee *a* answers.
2. For each of these items, obtain $P_g(\theta)$, the probability that an examinee with *a*'s estimated ability (θ) would answer the item correctly.

3. Sum these probabilities.

This process is denoted in the formula,

$$\xi_a = \Sigma^{(a)} P_g(\theta).$$

The number-right true score estimate for the examinee is then corrected for the effects of guessing by the formula

$$\eta_a = \Sigma^{(a)} P_g(\theta) - \frac{\Sigma^{(a)} Q_g(\theta)}{k-1}.$$

The use of the formula true score in item response theory is based on two critical assumptions: (a) the examinees' responses to the item are due solely to their ability levels on the latent trait, and (b) the examinees clearly understand and follow the formula scoring instructions; that is, they omit an item if and only if they have no better than random chance (1/k) of choosing the correct response (Crocker & Algina, 1986).

There are two major disadvantages to using this application. First, we can never know the examinees' true scores, therefore we must rely on estimated values of θ_a, ξ_a, and η_a. Second, we cannot know when the assumptions required for estimating the formula true score have been violated.

Considering these two drawbacks along with the complexity of item response theory, this application is rarely used.

Awarding Credit for Partial Knowledge

One of the major concerns with formula scoring is that it does not take into account partial knowledge. Partial information on a multiple-choice test item is defined as the ability to eliminate some, but not all, the incorrect choices, thus restricting guessing to a proper subset of choices that includes the correct choice (Frary, 1980). An examinee's partial knowledge affects validity since examinees who earn identical item scores on a conventionally scored multiple-choice item may have varying degrees of knowledge about that item. The recognition that examinees' levels of information fall on a continuum from complete information to complete misinformation led to a search for scoring methods that will reflect examinees' levels of information or misinformation. Scoring procedures designed to convey information about partial knowledge can be grouped into three general classes: confidence weighting, answer-until-correct, and option weighting (Crocker & Algina, 1986).

Confidence Weighting

Confidence weighting is a method of testing where weights are assigned directly or indirectly to item responses so as to reflect the examinee's belief in the correctness of the alternative or alternatives marked. When a confidence weighting procedure is used, the examinee is asked to indicate what they believe is the correct answer and how certain they are of the correctness of that answer. A right answer given confidently is given more credit than a wrong answer given without confidence. Examinees choosing the same response may receive different scores for that item because of their indications of their degrees of confidence in their responses.

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 the expected score only if the examinee reveals true degree of certainty in responding (Echternacht, 1972).

Echternacht (1972) found that the studies done testing this procedure have not shown an increase in reliability and validity coefficients and, in fact, several studies have shown a decrease in validity coefficients. In his study of confidence weighting procedures, Ebel (1965) found that a personality variable associated with general confidence, and uncorrelated with achievement, contaminates the results yielding an increase in measurement error variation. The lack of increase in reliability and validity, the personality variable, the complexity of the test-taking and test-scoring techniques, and the required test administration time are all reasons why this procedure has not been widely used.

Answer-Until-Correct (AUC)

Under an answer-until-correct scoring procedure, the examinee chooses alternatives until the correct response is selected. When the correct response is selected, the examinee is instructed to proceed to the next item. This procedure has been accomplished in the past by having the examinee erase a shield on an answer sheet, or by using a latent image answer sheet so a record is made of the number of responses attempted for each item. Today, this procedure can be readily accomplished using a computer. The traditional method of scoring AUC tests is to subtract the total number of responses made by an examinee from the total number of possible responses (Gilman & Ferry, 1972).

Hanna (1975) and Wilcox (1981) both listed several advantages to the AUC procedure: (a) the immediate feedback may promote learning, (b) it enables examinees to continue responding in a real-to-life fashion until feedback indicates success, (c) under certain assumptions, they can be used to correct for guessing without assuming guessing is at random, and (d) if examinees continue answering questions until they answer them correctly, the range of possible scores will be increased, and hence reliability and validity may be improved.

Studies have suggested that test scores using an AUC method are substantially more reliable than scores using a number-right method (Gilman & Ferry, 1972; Hanna, 1975). However, mixed results have been found in attempts to improve criterion-related validity. Hanna (1975) also found that the immediate feedback inherent in AUC media may adversely affect the performance of some anxious examinees who happen to score poorly on initial items. High administration costs due to the special media needed are also a drawback to using this method.

Option Weighting

Option Weighting is based on the assumption that item response options vary in degree of correctness and that examinees who select a “more correct” response have greater knowledge than those choosing “less correct” responses (Crocker & Algina, 1986). One method of obtaining scoring weights for the options is called rational weighting method. Judges rank order the alternatives to all test items from totally incorrect to totally correct. The judges' average rating is used as the weight for each option.

In a weighting system developed by Guttman, the weights for each option are proportional to the mean of the total test scores of the examinees who select it. Studies on validity and reliability when using Guttman weights (Hendrickson, 1971; Raffeld, 1975) have shown slight increases in reliability. Slight increases in predictive validity occurred only when constant weights are used for omissions. Option weighting could reduce a test's length while maintaining the same reliability and validity only if omissions do not exist or constant weights for omissions are used. Raffeld (1975) cautioned that the increases in reliability and validity are far from dramatic and that one would have to weigh carefully the gain in psychometric properties against the added cost of complex scoring.

Elimination and Inclusion Scoring

There are several other scoring methods available that consider partial knowledge. One of these, elimination scoring, provides a scale from complete misinformation through several degrees of partial information to complete information. In elimination scoring examinees are instructed to cross out all the alternatives that are incorrect. Inclusion scoring uses the same scale as elimination scoring, but examinees are instructed to circle the correct alternative for each item or the smallest subset of alternatives they believe includes the correct answer.

In a study by Jaradat and Tollefson (1988), comparing elimination and inclusion scoring with correction for guessing formulas, elimination scoring was found to give the most credit for partial information and permitted examinees to report their true states of knowledge on an item. Both elimination scoring and inclusion scoring have been found to produce slightly more valid and reliable scores, however test instructions were found to be confusing to some examinees which may outweigh the slight psychometric gain.

Multiple-Answer Format

A similar formula to inclusion scoring is the multiple-answer format. In the multiple-answer format, examinees are instructed that any number of the options might be correct. Each item in this format is scored by giving the number of answers correctly marked minus the incorrectly marked options.

Hsu, Moss, and Khampalikit (1984) conducted a study comparing multiple-answer formats to single answer formats. They found that multiple answer formats are more difficult, especially for below average examinees, than single answer formats, and that reliability for the two were relatively the same. The only merits found were in testing average and above average examinees giving partial credit for partially correct responses.

Discussion

Multiple-choice tests are the most common means of objectively measuring cognitive ability. Their popularity is due to the ease with which they can be administered and scored. However, the scoring of multiple-choice tests has been problematic. The most common way of grading multiple choice items is with a conventional number-right scoring method. Psychometricians have attempted to overcome the problems associated with this scoring method by the use of alternative scoring algorithms.

There have been numerous scoring formulas that attempt to correct for guessing and award credit for partial knowledge. However, results of empirical studies have not supported the theoretical rationale behind the formulas. Studies on reliability and validity show that the slight improvements over conventional scoring cannot justify the use of these methods when considering the disadvantages of the use of formula scoring. These disadvantages include complexity of administering and scoring the tests, as well as increased cost and time to administer the tests. Other concerns with formula scoring, especially with correction for guessing formulas, is that extraneous factors such as willingness to take risks and test wiseness can cast doubt on the interpretability of the scores obtained under this method.

Since the theoretical rationale behind formula scoring is sound, there continues to be research into scoring methods that are superior to conventional scoring. However, until a better method is found the conventional number-right scoring formula continues to be recommended.

References

Abu-Sayf, F.K. (1979). The scoring of multiple-choice tests: A closer look. Educational Technology, 19, 5-15.

Albanese, M.A. (1988). The projected impact of the correction for guessing on individual scores. Journal of Educational Measurement, 25, 149-157.

Angoff, W.H. (1989). Does guessing really help? Journal of Educational Measurement, 26, 323-336.

Bliss, L.B. (1980). A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 17, 147-153.

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. Journal of Educational Measurement, 30, 277-291.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Cross, L.H., & Frary, R.B. (1977). An empirical test of Lord's theoretical results regarding scoring of multiple -choice tests. Journal of Educational Measurement, 14, 313-321.

Diamond, J., & Evans, W. (1973). The correction for guessing. Review of Educational Research, 43, 181-191

Echternacht, G.J. (1972). The use of confidence testing in objective tests. Review of Educational Research, 42, 217-236.

Frary, R.B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. Applied Psychological Measurement, 4, 79-90.

Gilman, D.A., & Ferry, P. (1972). Increasing test reliability through self-scoring procedures. Journal of Educational Measurement, 9, 205-207.

Hanna, G.S. (1975). Incremental reliability and validity of multiple choice tests with an answer-until-correct procedure. Journal of Educational Measurement, 12, 175-178.

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

Hsu, T., Moss, P.A., & Khampalikit, C. (1984). The merits of multiple-answer items as evaluated by six scoring formulas. Journal of Experimental Education, 36, 152-158.

Jaradat, D., & Tollefson, N. (1988). The impact of alternative scoring procedures for multiple-choice items on test reliability, validity, and grading. Educational and Psychological Measurement, 48, 627-635.

Lord, F.M. (1975). Formula scoring and number-right scoring. Journal of Educational Measurement, 12, 7-12.

Raffeld, P. (1975). The effects of Guttman weights on the reliability and predictive validity of objective tests when omissions are not differentially weighted. Journal of Educational Measurement, 12, 179-185.

Rowley, G.L., & Traub, R.E. (1977). Formula scoring, number-right scoring, and test-taking strategy. Journal of Educational Measurement, 14, 15-22.

Sax, G. (1989). Principles of educational and psychological measurement and evaluation (3rd ed.). Belmont, CA: Wadsworth.

Slakter, M.J. (1968). The penalty for not guessing. Journal of Educational Measurement, 5, 141-144.

Wilcox, R.R. (1981). A closed sequential procedure for answer-until-correct tests. Journal of Experimental Education, 5, 219-222.

TM029476

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: A REVIEW OF SCORING ALGORITHMS FOR MULTIPLE-CHOICE TESTS	
Author(s): TERRI BARBER KURZ	
Corporate Source:	Publication Date: 1/99

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☒ ◀ **Sample sticker to be affixed to document**

Check here

Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

TERRI BARBER KURZ

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document ▶ □

or here

Permitting reproduction in other than paper copy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: X Terri Barber Kurz	Position: RES ASSOCIATE
Printed Name: TERRI BARBER KURZ	Organization: TEXAS A&M UNIVERSITY
Address: TAMU DEPT EDUC PSYC COLLEGE STATION, TX 77843-4225	Telephone Number: (409) 845-1831
	Date: 1/20/99

OVER