DOCUMENT RESUME

ED 417 220 TM 028 177

AUTHOR	Chevalier, Shirley A.
TITLE	A Review of Scoring Algorithms for Ability and Aptitude Tests.
PUB DATE	1998-04-11
NOTE	27p.; Paper presented at the Annual Meeting of the Southwestern Psychological Association (New Orleans, LA, April 1998).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ability; *Algorithms; *Aptitude Tests; Cognitive Tests; *Guessing (Tests); Item Response Theory; Reliability; *Scoring; Validity
IDENTIFIERS	High Stakes Tests; Number Right Scoring; Partial Credit Model

ABSTRACT

In conventional practice, most educators and educational researchers score cognitive tests using a dichotomous right-wrong scoring system. Although simple and straightforward, this method does not take into consideration other factors, such as partial knowledge or guessing tendencies and abilities. This paper discusses alternative scoring models: (1) credit for omissions; (2) disproportionate correction for wrong versus omitted items (correcting for guessing); (3) scoring only for items that a given examinee is expected to get right based on one-parameter item response theory (Lawson, 1991); and (4) scoring using various partial credit models, including misinformation. The literature regarding the utility of each algorithm, including validity and reliability, is also summarized briefly. Psychologists should be familiar with alternative scoring strategies, since such strategies can be useful in the design, administration, or analysis of results from measures of cognitive abilities, especially in high stakes testing. Findings from this exploration indicate that correction for guessing formulas do not show significant benefits over conventional scoring (no correction), and while results on partial credit scoring algorithms are inconclusive, the observed slight increases in reliability and validity do not justify the additional complexity, time, and cost involved in developing, administering, scoring, and interpreting test results. (Contains 1 table and 20 references.) (Author/SLD)

Running Head: SCORING ALGORITHMS FOR ABILITY AND APTITUDE TESTS

A Review of Scoring Algorithms for Ability and Aptitude Tests

Shirley A. Chevalier

Texas A & M University 77843-4225

Paper presented at the annual meeting of the Southwestern Psychological Association, New Orleans, April 11, 1998.

Abstract

In conventional practice, most educators and educational researchers score cognitive tests using a dichotomous right-wrong scoring system. Although simple and straightforward, this method does not take into consideration other factors, such as partial knowledge or guessing tendencies and abilities. The present paper discusses alternative scoring models: (a) credit for omissions; (b) disproportionate correction for wrong versus omitted items (correcting for guessing); (c) scoring only for items that a given examinee is expected to get right based on one-parameter item response theory (Lawson, 1991); and (d) scoring using various partial credit models, including misinformation. The literature regarding the utility of each algorithm, including validity and reliability, will also be briefly summarized. Psychologists should be familiar with alternative scoring strategies. Such scoring strategies can be useful to the individuals involved in designing, administering, or analyzing results from measures of cognitive abilities, especially in high stakes testing.

Findings of this report indicate that correction for guessing formulas do not show significant benefits over conventional scoring (no correction) and, while results on partial credit scoring algorithms are inconclusive, the observed slight increases in reliability and validity do not justify the additional complexity, time, and cost involved in developing, administering, scoring and interpreting test results.

A Review of Scoring Algorithms for Ability and Aptitude Tests

The goal of cognitive measurement is to obtain an examinee's best, maximum, and highest level of performance (Hopkins & Stanley, 1981). Relevant cognitive tests include measures of achievement, intelligence, and aptitude. Uses of such test scores may range from providing information to teachers as to which students have mastered a curriculum, to such high-stakes situations as whether an examinee should be accepted into a particular university or whether a business should hire or promote a particular candidate. Because of the "high-stakes" nature of cognitive tests, issues have also been raised concerning how effectively tests measure cognitive functioning. The areas of concern include obtaining undeserved credit (not based on acquired knowledge or skill) and failure to receive credit for partial knowledge. Multiple-choice tests are the most common, and perhaps the best, tool for objective measurement of knowledge, ability, or achievement. Major weaknesses of multiple choice tests include susceptibility to guessing and insensitivity to differences between various levels of knowledge (Ben-Simon, Budescu, & Nevo, 1997).

Hopkins and Stanley (1981) noted that a general know-how of test taking can itself affect test performance. They define testwiseness as "an examinee's ability to use the characteristics and formats of the test and/or the test-taking situation to increase his/her score" (p.141). Test-wiseness is a construct known to affect the validity of test scores because test-taking skills contaminate and confound the assessment of acquired knowledge (Harmon, Morse, and Morse, 1996). Test-wiseness is logically independent of the examinee's knowledge of the subject matter for which the items are supposedly measures (Millman, Bishop, and Ebel (1995). Angoff (1989) noted that more able examinees do tend to profit from guessing. Dolly and Vick (1986) listed several studies which support test-wiseness as a source of additional variance in test scores and as a possible depressor of test score validity. Among their findings, investigators noted that (a) the subject who is low in test-wiseness is penalized for this deficit; (b) when guessing is encouraged, the average performance tends to increase; (c) test-wiseness can be learned.

Frary (1980) defined partial information on a multiple-choice test item as the ability to eliminate some, but not all, the incorrect choices, thus restricting guessing to a smaller subset of choices that includes the correct choice. If an examinee thinks the correct alternative is wrong, this is termed "misinformation". If the examinee also recognizes some of the distractors as being wrong, this becomes "partial misinformation" (Coombs, Mulholland, & Womer, 1956).

In conventional scoring of objective tests, each test score is the sum of the item scores for a given examinee, and the examinee is awarded one point for the correct item response. With this scoring rule, all items are weighted equally. This is sometimes called "number-right scoring". Although simple and straightforward, this method may be somewhat problemmatic with multiple-choice and true-false items. The conventional method also does not take into consideration the effects of test-wiseness and guessing, or award partial credit for partial knowledge.

However, there are numerous other scoring strategies, which can avoid some of these problems (Crocker & Algina, 1986; Gronlund & Linn, 1990; Pedhazur & Schmelkin, 1991; Sax, 1989; Thorndike, Cunningham, Thorndike & Hagen, 1991). For example, in addition to providing greater variance and increased score validity and reliability, some educators find other scoring models provide better feedback to students, offer more information about abilities, or are more motivational in that some of these other systems recognize and reward partial credit for partially correct choices.

This present paper summarizes fundamental information that may be beneficial to the educators involved in designing, administering, or analyzing results from measures of cognitive abilities. In an effort to provide a fundamental, but comprehensive, framework of scoring formulas, the following topics are addressed: extraneous factors which may affect performance (scores) on cognitive tests; conventional scoring methods; the utility of each scoring algorithm; advantages/disadvantages; and reliability and validity are discussed. The information presented is based on literature reviews of textbooks on educational and psychological measurement and various journal articles.

Extraneous Factors That Influence Performance on Cognitive Tests

Hopkins and Stanley (1981) noted that in addition to the trait, knowledge, or proficiency that is to be measured, many other factors may affect an examinee's performance on a test. Noteworthy extraneous factors identified include: test sophistication or test-wiseness, retest or repetition, coaching, and response styles (set). Response set includes speed versus accuracy (tendency to work slowly and carefully in some and quickly with less caution in others); positional-preference (favor certain positions of items in a list of five responses); option length set (favoring the longest option on difficult multiple-choice tests; tendency to select nontechnical options more frequently, irrespective of length), and the gambling set. However, a study by Shatz (1985) investigating the effectiveness of the guessing strategies identified by students as the most frequently used (selection of the least used letter choice, the random selection of a choice, and the selection of choice "C"), found these strategies did not produce higher-than-chance scores when test items are randomly ordered. In the case of the least-used answer strategy, significantly lower-than-chance results were obtained.

The gambling set, hereafter referred to as guessing, is a tendency which varies from the person who will not guess even when told that he/she must answer every question, to the "gambler" who attempts almost every item regardless of penalties or directions. On tests designed to discriminate among examinees, guessing would not be a problem if all students of equal ability guessed with equal frequency. On tests that do not employ a correction for chance, the "gambler" is given a special advantage over the more deliberate student. On most tests, examinees can guess better than chance because (a) they may have partial information on several items, and (b) on many items, not all distractors are plausible.

Because ability and aptitude tests theoretically evaluate student achievement, and not how test-wise a student is, Kubiszyn and Birch (1990) advocated attempting to equalize the advantages test-wise students have over nontest-wise students. Dolly and Vick (1986) defined test-wiseness as a cognitive ability or set of skills which a test-taker can use to improve a test score, and proposed a set of indicators for predicting examinees with test-wise abilities. They found evidences that test-wiseness can be predicted from pretest score, grade point average, and examinee's test-taking perceptions. To equalize test-taking advantages, they further proposed test-wiseness training to decrease the gap between the "haves and have-nots" to decrease variance (associated with test-wiseness) among test scores. Using a two-factor model proposed for the Gibb Experimental Test of Testwiseness (measures the use of secondary cues found in test items to answer the test question itself without content-specific knowledge) and confirmatory factor analysis to test for stability, Harmon, Morse, and Morse (1996) confirmed that the Gibb test is another sound measure of test-wiseness.

Other studies suggest constructing test items to preclude the effects of test-wiseness and response set. Lien (1967) offered the following suggestions for developing choices (responses): (a) correct choice should be placed at random among choices (no fixed patterns); (b) in elementary school, a minimum of three choices should be given; in high school, a minimum of four; (c) the suggested wrong choices should represent errors commonly made by the students in class discussion rather than general misconceptions; (d) the suggested choices should be as brief as possible (avoid measuring reading skills); (e) irrelevant clues should direct the examinee away from the right answer if he/she is unable to answer the problem (never direct them to the right answer). Stanley and Hopkins (1981) further recommend that, to lessen the likelihood of one being able to select the correct option by chance, the number of options should be increased. (However, it should be noted that, from an information theory perspective, Bruno and Dirkzwager (1995) indicate that three choices to multiple-choice test item appears optimal.)

To evaluate test results properly, one should not only be aware of the existence of extraneous variables, but also be able to make appropriate allowances for such factors in interpreting the results. Alternative scoring formulas are believed by some to be the solution to alleviate the effects of extraneous variables and weaknesses associated with using multiple-choice test items.

Correcting for Guessing Formula

The traditional approach to correction for guessing among examinees is through formula scoring (Hopkins & Stanley, 1991; Kubiszyn & Borich, 1990). Crocker and

Algina (1986, p. 400)) quote Rowley and Traub as identifying three possible situations taken into account by the correction for guessing formula scoring: The examinee knows the correct option and chooses it, the examinee omits the item, or the examinee guesses blindly and selects one of the item responses at random.

A correction-for-guessing formula penalizes examinees for answering questions to which he/she does not know the answer. A commonly used formula is

$$S = R - W/I$$

Where S = the examinees score corrected for chance,

R= the number of Right responses marked by the examinee

W = the number of Wrong responses, not including omitted items, and

I = the number of Incorrect options (distracters) per item

This is known as the rights minus wrongs correction. This formula theoretically reduces to zero the scores of students who, totally ignorant of the material presented in the test, guess with a chance degree of success that depends only on the number of options each item has. For example, on a five-option test, a student Larry has a chance of scoring 20-points by chance alone. Assuming Larry guessed on all 100 items and scored 20-points by guessing alone, using the rights minus wrongs formula

$$S = 20 - 80/4$$

$$S = 0$$

It is important to note that, due to penalties for guessing, using this formula,it is theoretically possible for an examinee to score less than zero. For example, considering a worstcase scenario of Larry responding to all questions but failing to answer any correctly. On a three-item multiple-choice test, he would score a negative 50-points.

Kubiszyn and Borich (1990) recommend to test administrators that, when this formula is used, examinees are strongly cautioned to use "educated guesses".

A more positive, rather than negative correction approach discourages guessing by awarding credit for omitting items (versus penalizing for guessing). The formula can be written as

$$S = R + O/A$$

where S and R are defined as above, A is the number of alternatives (options) per item, and O is the number of omitted items. It is significant to note that on a 3-options test, with 100 test items, if the examinee omitted all 100 items, he/she would still score 33 points on the test.

Hopkins and Stanley (1981) noted several studies which indicate that correction formulas show a negligible decrease in reliability and a slight increase in validity. Crocker and Algina (1986) quoted studies which show similar results. Crocker and Algina also referred to studies whose results indicate that when students answer all items, they achieve higher raw scores than when they respond under formula scoring instructions and scores are corrected for guessing. A possible explanation is that the formula-scoring model does not take into consideration partial knowledge. Considering the questionable validity and reliability results, time and effort required, and the negative public relations which can result from the "penalty for guessing", routine use of the formula is rarely justified.

Item Response Theory

Crocker and Algina (1986) noted that for test items developed using item response theory, a formula proposed by Lord may be used to estimate an examinee's true score.

Based on the probability ($P_g(\theta)$) that an examinee with ability level θ will answer item g correctly, an examinee's true score may be estimated by summing 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 this item correctly.
3. Sum these probabilities.

Written as: $\xi_a = \Sigma^{(a)} P_g(\theta)$

where ξ_a is the number-right true score for the examinee, and $\Sigma^{(a)}$ means to sum over only those items answered by the examinee. The number-right true score estimate for the examinee is then corrected for the effects of guessing by the formula

$$\eta_a = \Sigma^{(a)} P_g(\theta) - [\Sigma^{(a)} Q_g(\theta)]/k-1$$

where $Q_g(\theta)$ is $[1-P_g(\theta)]$, and k is the number of choices per item.

The use of the formula true score in item response theory is based on two critical assumptions:

1. the examinees' reponses to the items are due solely to their ability levels on the latent trait, and
2. the examinees clearly understand and follow the formula-scoring instructions; they omit an item if (and only if) they have no better than random chance (1/k) of choosing the correct response.

In actuality, we can never know examinees' true scores and must rely on estimated values of $\hat{\theta}_a$, $\hat{\xi}_a$, and $\hat{\eta}_a$. Nor can we know when the assumptions required for estimating the formula true score have been violated. Practical benefits derived from its application have yet to be demonstrated.

Awarding Credit for Partial Knowledge

When faced with a test question, the examinee is typically in one of three subjective states (Ben-Simon, Budescu, & Nevo, 1997):

1. the examinee knows the answer fully and with confidence (full knowledge);
2. the examinee knows only part of the answer or is uncertain of the answer (partial knowledge); or
3. the examinee has no knowledge of the answer (absence of knowledge).

Scoring procedures designed to convey information about partial knowledge can be grouped into three general classes (Crocker & Algina, 1986):

1. Confidence weighting: Format and instructions are constructed so that the examinees must indicate how certain they are of the correctness of each response.
2. Answer until correct (AUC): The examinee reads the multiple-choice test item, selects a response, and receives immediate feedback about the correctness of that selection. If the correct response is selected, the examinee is instructed to proceed to the next item; if an incorrect choice has been selected, the examinee is instructed to make another selection. The typical method of scoring is to subtract the total number of responses (by the examinee) from the total number of possible responses. Gilman and Ferry (1972) reported an observed increase in reliability using AUC methods over traditional right-wrong methods.

3. Option weighting: This system is based on the assumption that item response options vary in degree of correctness and that examinees who select a "more correct" response have a greater knowledge than those choosing "less correct" responses. Options of a multiple-choice item are assigned different weighted values depending on the particular option chosen. Using the GRE aptitude test conventional scoring formula as a baseline and comparing results with a priori weights (different weights for distractors developed for the test) Echternacht (1976) concluded that a priori option weighting was less reliable. He also concluded that in order for priori scoring to be cost-effective, a significant increase in reliability was required. Hambleton, Roberts, and Traub (1970) reported mixed results when comparing the reliability and validity of differential weighting and confidence testing.

Ben-Simon, Budescu, & Nevo (1997) conducted a meta-analysis of 16 studies to identify the most promising scoring methods available for accounting for partial knowledge. A brief summary of these methods is presented:

1. Item weighting: The basic principle in these methods is to "overweight" good items and to underweight "poor" items. The weights are generated from the results of an item analysis. The most commonly used weights include measures of item difficulty, validity, diagnostic ability, variance, or weights determined by experts.
2. Item structure/presentation: One method is to present the examinee with many items in dichotomous (correct/incorrect) format (i.e., the number of incorrect items identified as incorrect plus the number of correct items identified as correct). A second example is to provide multiple correct options (item has more than one correct answer). The final score is the total number of correct answers identified (in some

cases incorrect answers may be penalized). Relatively little research has been conducted on these types of items because of difficulty involved in constructing the items.

3. Matching: This is a third example of item structure, in which the examinee is presented a group of items accompanied by a long list of possible answers. This is called simple or multiple matching. The examinee is required to match the correct answers to the questions. When the number of possible answers is identical to the number of questions, the method is referred to as "simple" matching. When the number of possible answers is greater than the number of questions, the method is referred to as "multiple" matching. The multiple matching option should result in the reduced probability of guessing due to the increased number of alternative answers. However, the test constructor may encounter difficulty in producing appropriate items and longer administration time is required.

4. Self-assessment of knowledge: This method uses weights supplied by the examinees to reflect the knowledge at their command. The method is designed to reduce guessing and measurement error. Scoring differs from one option to another; main methods include:
   a. complete or partial ordering of options;
   b. confidence weighting - selecting the most likely option and indicating the degree of certainty that it is correct;
   c. probability weighting - assigning weights to each option according to the probability that it is correct;

d. selecting a subgroup of acceptable answers, or conversely; and

e. rejecting all unacceptable answers.

They are reported to be easy to implement and simple to use. They are flexible in terms of response instructions and scoring method.

8. Elimination scoring: Examinees are instructed to eliminate all distractors that they can identify as incorrect. This method has been found to discriminate between all possible levels of knowledge and discourages guessing. Studies indicate slightly improved validity and reliability over number correct and other correct-for-guessing methods. Disadvantages include: instructions and scoring are relatively complex, require longer administration time, and examinees may apply response instruction ineffectively (i.e. too conservatively).

9. Probability testing: This method allows examinees to express partial knowledge by reporting the probability that each option is the correct answer. There are 101 possible scores for each item (ranging from 0 to 1). It does not distinguish unequivocally between all five levels of knowledge. Scoring methods can be complex and may encourage examinees not to report their probabilities truthfully.

10. Confidence testing: This method is similar to probability testing; the examinee expresses his/her confidence only for the most correct option by using a C-point scale (C = confidence rating), accompanied by verbal labels, e.g. unsure.

11. Complete ordering: This method is similar to probability testing; the examinee assigns rank order to each option. This method discriminates between three levels of knowledge: partial, full, and absence. It trades precision of measurement for simplicity of administration and application.

12. Partial ordering: This method is a hybrid of elimination testing and complete ordering. The examinees are asked to rank (complete ordering) only those options that cannot be totally eliminated (elimination testing) from consideration. This method discriminates between all levels of knowledge.

Analysis of results of the meta-analysis provided mixed results regarding efficiency, reliability and validity of various scores from these various methods. To compensate for heterogeneity among the original 16 studies, Ben-Simon et al. (1997) followed up the meta-analysis with a psychometric study of the results from a single standardized test. The following is a summary of the results of comparisons of the measure of partial knowledge from this follow-up study:

- the treat of penalties induced higher levels of omission, lower mean scores, and higher variance of scores. Thus, penalizing methods should not be used if it is important to increase response rate (i.e., minimize omission rate); but should be used to increase the variance of scores;
- across tests, the highest reliabilities were obtained for the probability testing and confidence marking methods; and
- four response methods distinguished between six well-defined levels of knowledge: partial ordering, elimination testing, confidence marking, and probability testing. Thus, if interest lies in discriminating between all levels of partial knowledge, these four methods are available.

Findings also indicated that:

- Examinees had a tendency to overestimate their knowledge under all response methods. The confidence marking produced the most realistic assessments. High ability examinees were better calibrated. High ability examinees were more accurate and less biased judges of their knowledge and tended to benefit more from proposed methods;
- None of the methods examined emerged as uniformly best; and
- The model of knowledge which assumes that knowledge leads to correct responses, and a lack of knowledge leads to omissions or excessive guessing, appears invalid. Two models are presented:
  - One model assumes that knowledge about each alternative answer is dichotomous, which implies that, if the correct answer is not known to the examinees, incorrect options can be identified and eliminated; and
  - Another model assumes that knowledge is continuously distributed with regard to each alternative answer. Thus, in the absence of full knowledge, an examinee should be willing to allocate ranks, weights, or probabilities to each response alternative. Only the most likely (highest rank) option is selected with a certain degree of confidence.

In a similar study, Hsu, Moss, and Khampalikit (1984) compared the benefits of single-answer (SA) and multiple-answer (MA) scoring formulas. SA multiple-choice items are defined as consisting of a single correct option or answer and a number of incorrect options or distractors. The MA multiple-choice items are defined as having more than one correct option. If scored separately, MA items are identical to true-false items. The six formulas are summarized in Table 1. Comparisons of correction for guessing versus no correction for guessing and partial credit versus no partial credit in terms of discrimination, difficulty, and efficiency were performed. Results indicated:

- multiple answer items are consistently more difficult and more discriminating than single-answer items, however, reliability is about the same;
- single-answer items provide more information for below average examinees; multiple-answer items for average and above average examinees;
- correcting for guessing only makes the test more difficult; offers no statistically significant difference for discrimination power and reliability (over no correction for guessing);
- giving partial credit makes the test easier; partial credit formulas show only a slight increase in discrimination power and reliability over no partial credit formulas;

The conclusion drawn, is that there are some merits in using multiple-answer items, especially when testing examinees with average and above average ability. However, further investigation of an optimal scoring algorithm and test items is warranted. Application of the correction-for-guessing seems unnecessary, because no significant benefits in test score reliability and validity have been demonstrated.

Discussion

Multiple-choice is the most common, and perhaps the best, means of measuring cognitive abilities. However, there are weaknesses in the use of multiple-choice test items. These weaknesses include: extraneous variables, such as test-wiseness on the part of some examinees, which not only confound the measurement of the intended trait or characteristic, but also cause (at least a perceived) unfair advantage among test-takers (versus those without test-wiseness); a failure to measure partial knowledge; and possible threats to score validity and reliability. Possible solutions to the dilemma include improved construction of test items, testing for test-wiseness as a characteristic among students and providing training to those non-testwise students in order to "level the playing field," and identifying optimal scoring formulas (and associated test item construction) to compensate for the extraneous factors.

Numerous scoring formulas and test construction methods have been identified and studied with some methods showing promise. However, results indicate that current methods show only slight improvements in score reliability and validity while adding complexity to the test construction, administration, scoring, and interpretation with increased cost and time. To date, the benefits do not justify the effort. It is important to note, however,that most studies demonstrate no significant benefits to employing correction for guessing strategies (versus no correction). Because of the promise shown in methods awarding partial credit using multiple-answer test items, studies indicate that additional investigation in this area is warranted.

Although empirical data on the effects of extraneous factors on the validity and reliability of tests of cognitive abilities using multiple-choice items are inconclusive, most educators agree that the effects are theoretically valid and reason for concern, especially in the arena of high-stakes testing. Existing research has identified some interesting methods involving test construction and associated scoring algorithms to address the issue; however, studies reviewed in this report indicate that while those models allowing partial credit for partial knowledge (multiple-answers per test item) appear promising, further investigation is required to identify the optimal test method(s) and scoring formula(s). It appears that (a) correcting-for-guessing is unnecessary and should be avoided; and (b) future studies are needed to investigate an optimal test method and scoring formula for cognitive tests. Therefore, the conventional testing and scoring formula is recommended.

References

Angoff, W. H. (1989). Does guessing really help? Journal of Educational Measurement, 26, 323-335.

Ben-Simon, A., Budescu, D. V., & Nevo, N. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. Applied Psychological Measurement, 21, 65-88.

Bruno, J. E., & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. Educational and Psychological Measurement, 55, 959-966.

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

Dolly, J. P., & Vick, D. S. (1986). An attempt to identify predictors of test-wiseness. Psychological Reports, 58, 663-672.

Echternacht, G. (1976). Reliability and validity of item option weighting schemes. Educational and Psychological Measurement, 36, 301-309.

Frary, R. B. (1980). The effect of Misinformation, partial information, and guessing on expected multiple-choice 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.

Gronlund, N. E., & Linn, R. L. (1990). Measurement and evaluation in teaching (6th ed.). New York: Macmillan.

Hambleton, R. K., Roberts, D. M., & Traub, R. E. (1970). A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. Journal of Educational Measurement, 7, 75-81

Harmon, M. G., Morse, D. T., & Morse, L. W. (1996). Confirmatory factor analysis of the Gibb experimental test of testwiseness. Educational and psychological Measurement, 56, 276-286.

Hopkins, K. D., & Stanley, J. C. (1981). Extraneous factors that influence performance on cognitive tests. In G. V. Glass (Ed.), Educational and psychological measurement and evaluation (6th ed., pp. 141-157). New Jersey: Prentiss-Hall.

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

Kubiszyn, T., & Borich, G. (1990). Educational testing and measurement (3rd ed.). Glenview, Illinois: Brown Higher Education.

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), (1991). Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 159-168). Greenwich, CT: JAI Press.

Lien, A. J. (1967). Measurement and evaluation of learning. Dubuque, Iowa: W. C. Brown.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

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

Shatz, M. A. (1985). Students' guessing strategies: Do they work? Psychological Reports, 57, 1167-1168.

Thorndike, R. M., Cunningham, G. K., Thornkike, R. L., & Hagen, E. P. (1991). Measurement and evaluation in psychology and education (5th ed.). New York: MacMillan.

Table 1.

Illustrations and Descriptions of the Six Scoring Formulas (Hsu, Moss, and Khampalikit, 1984)

Method	I. D. #	Formula	Description
SA & MA	1	S = W, if correct	full credit only if all options correct; penalty for guessing
		= - W/k-1, if incorrect	
		= - W/25, if not all choices correct (multiple-response)	
SA & MA	2	S = W, if correct	full credit only if all options correct; no penalty for guessing
MA only	3	S = [(C - I)/K] W	partial credit if some items correct; penalty for guessing
MA only	4	S = [C/K] W	partial credit if some items correct; no penalty for guessing
MA only	5	S = [(U/R) - (V/K-R)]	partial credit; (proportion of marked answers minus proportion of marked distractors); penalty for guessing
MA only	6	S = [(C/K) - $\binom{K}{I}$/2_k] W	partial credit; all incorrect choices are guessing penalty for guessing

where, S = the obtained score for each item

W = the assigned weight for each item

R = the number of options keyed as correct (i.e., answers)

C = the number of correct choices made by the examinees (marked answers and unmarked distractors).

I = the number of incorrect choices made by the examinees (marked distractors and unmarked answers).

$K = C + I$ = the total number of options

U = the number of marked answers

V = the number of marked distractors

TM028177

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 ABILITY AND APTITUDE TESTS	
Author(s): SHIRLEY A. CHEVALIER	
Corporate Source:	Publication Date: 4/11/98

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.

[X] ← 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
SHIRLEY A. CHEVALIER
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: Shirley A. Chevalier	Position: RES ASSOCIATE
Printed Name: SHIRLEY A. CHEVALIER	Organization: TEXAS A&M UNIVERSITY
Address: TAMU DEPT EDUC PSYC COLLEGE STATION, TX 77843-4225	Telephone Number: (409) 845-1335
	Date: 2/16/98