Partial-Credit Scoring Methods for Multiple-Choice Tests

Robert B. Frary

*Virginia Polytechnic Institute and State University*

This review covers multiple-choice response and scoring methods that attempt to capture information about an examinee’s degree or level of knowledge with respect to each item and use this information to produce a total test score. The period covered is mainly from the early 1970s onward; earlier reviews are summarized. It is concluded that there is little to be gained from the complex responding and scoring schemes that have been investigated. Although some of them have confirmed potential to increase internal-consistency reliability, this outcome is often obtained only at the expense of validity. Also, the extra responding time required by some methods would permit lengthening a conventional multiple-choice test sufficiently to obtain the same reliability improvement. Partial-credit response and scoring methods that continue to be used will probably earn this status due to secondary characteristics such as providing feedback to enhance learning.

For free-response tests, it is commonplace for scorers to give partial credit for answers that are incorrect but nevertheless reveal some relevant knowledge on the part of the examinees. Moreover, a scorer would probably be considered inflexible or possibly negligent if this practice were not followed. Therefore, it is not surprising that, for as long as multiple-choice tests have been in general use, there seems to have been a widespread, nagging uneasiness about the all-or-nothing character of conventional number-right scoring. Another way this uneasiness has been expressed is in

Requests for reprints should be sent to Robert B. Frary, Learning Resources Center, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.
terms of the potential for guessing correctly, either after elimination of some but not all wrong options, or when completely ignorant. Even today, some educators and segments of the general public dismiss multiple-choice testing as inherently deficient, claiming that lucky and test-wise examinees gain unwarranted advantage from guessing. At the same time, psychometricians wish to eliminate score variance due to guessing to enhance reliability and validity.

In part, these concerns are addressed by multiple-choice response and scoring methods that attempt to capture information about an examinee's degree or level of knowledge with respect to each item. The methods devised for this purpose are so diverse that some categorization is needed to facilitate reviewing them. The following basis for categorization is used in this review:

1. Methods simply requiring the examinees to mark what they believe is the best answer. Included in this category is answer-until-correct testing, in which examinees mark the single option they believe to be correct, but must repeat this action until the correct option is marked. These methods are referred to as direct response methods.

2. Methods requiring the examinees to judge the extent of their knowledge concerning an item. These methods may require an examinee to mark options believed to be clearly wrong or may require some direct expression of examinee confidence that a marked option is correct. These methods are referred to as examinee judgment methods.

Strictly speaking, conventional formula scoring, which instructs examinees not to guess when guessing would be at random among all options, belongs in the second category. This is because examinees must then judge whether they have any basis forguessing other than at random and, subject to the capriciousness of luck, tend to receive partial credit for guessing among fewer than all the options. Nevertheless, to avoid undue complexity and in consideration of space limitations, this review does not cover formula scoring. Also not covered are methods that simply transform the number right and possibly the number omitted to yield estimates of parameters such as the proportion of options the examinee can independently classify correctly as either right or wrong (Garcia-Perez, 1987) and the proportion of wrong options the examinee can recognize (Reid, 1977). (See also, Duncan, 1974, for an interesting approach to correcting for misinformation.) Two earlier reviews are summarized briefly at this point.

An extensive review by Wang and Stanley (1970) covered item weighting, as well as topics related to the present review. Item weighting refers to awarding differing point values (fixed across examinees) for correct answers to each test item. For example, a correct answer to some items might be
worth 1 point, whereas for other items, it might be worth 2 points. From
this simple kind of a priori weighting scheme to ones involving weights
determined by multiple-regression techniques, the results from empirical
studies were uniformly disappointing. As pointed out by Wang and Stanley
(see also Stanley & Wang, 1970), this outcome was only to be expected for
tests of more than a few items, given limitations imposed by the natural
tendency for responses to test items to be somewhat intercorrelated.
Although thus effectively "burying" item weighting, Wang and Stanley held
out hope for methods assigning item scores that could vary for a single item
on the basis of how the examinees responded or, in other words, provide
partial credit. Their review covered both direct response and examinee
judgment methods.

Echternacht (1972) reviewed only examinee judgment methods. This
review ended on a pessimistic note, and, indeed, the results of the empirical
studies reviewed could only be classified as mixed at best. However,
Echternacht's pessimism arose not only from these unpromising results but
also from psychological studies suggesting that confidence is a personality
trait influencing behavior independently of other stimuli.

The present review is mainly devoted to covering developments since
publication of these two reviews. However, some older studies are described
to provide perspective or if they contribute directly to the conclusions of
this review.

DIRECT RESPONSE METHODS

Option Weighting

When option weighting is used, examinees are instructed simply to mark a
single option for each item. Varying point values are assigned based on
which option is chosen, including the option of omitting. The maximum
value per item is typically uniform. Option weights (the point values
associated with each option) may be determined by expert judgment or
empirically.

Weights may be determined from judgments by a panel of experts or by
a single individual. For example, the options for each item might be ranked
or rated as to their adequacy as answers, and then weights proportional to
the ratings or ranks (average ratings or ranks if provided by a panel) would
be assigned. Nedelsky (1954) pioneered this approach with a simplified
method for determining scores and reported higher reliability compared to
number-right scoring. However, this finding was not supported by a test of
statistical significance. Since that time, results from the use of judgment-
based choice weighting have been varied and inconsistent. For example,
Patnaik and Traub (1973) used Thurstone's rank-order method of paired-comparison scaling to determine option weights based on rankings provided by a panel of 61 members. One version of this weighting scheme had varying maxima across items, lending an aspect of item weighting to this study. However, the results were the same when a uniform maximum weight per item was imposed—namely, that scores arising from use of the option weights were significantly more reliable and valid. In contrast, Kansup and Hakstian (1975), using a weighting scheme that provided integral weights, found no advantage from judgment-based option weighting, as did one phase of studies by Cross, Ross, and Geller (1980) and by Downey (1979). Even worse, one phase of a study by Echternacht (1976) found judgment-based option weighting inferior to other scoring methods both with respect to reliability and validity.

Option weighting based on the responses of the examinees themselves is called empirical option weighting. An early study of this type was that by Davis and Fifer (1959), one of whose methods was to weight options in proportion to the point-biserial correlations between choosing each option and total score. Thus, an examinee who chose an option popular among low scorers would receive a low item-score. Their results showed significant improvement in internal-consistency reliability with no meaningful decrease in validity compared to number-right scoring.

Subsequent studies have used a variety of weighting methods. Most have used for an option weight the average z score of examinees selecting that option as determined over the other items of the test. If this procedure is repeated iteratively until coefficient alpha stabilizes, the result is equivalent to that derived by Guttman (1941) for maximizing internal consistency. Bejar and Weiss (1977) used simulation to study three empirical option-weighting methods and found that their relative effectiveness for improving coefficient alpha was related to the level of interitem correlation. Claudy (1978) used Brogden biserial correlation coefficients between option selection and total score on the remaining items, finding it superior to other methods with respect to internal-consistency reliability. Nevertheless, the differences due to weighting methods have been small and are probably not of much practical importance.

Although improvement in internal-consistency reliability has been a typical outcome of empirical option-weighting studies, most have either not evaluated validity or have found validity to be weakened (e.g., Hendrickson, 1971; Reilly & Jackson, 1973; Waters, 1976). Echternacht (1976) did not find validity weakened by empirical option weighting but used a criterion so similar to the predictor test that the reported validity coefficients might better be regarded as parallel-form reliability coefficients. Another exception (not covered by Wang & Stanley, 1970) was reported by Sabers and White (1969), who obtained nearly identical
reliability and validity estimates for number-right versus option-weighted scores. This study and the Davis and Fifer (1959) study differed from the other studies involving validity in that no admonition against random guessing was given. The others were administered under conventional formula-scoring instructions and found that the weights assigned to the option of omitting were almost always the lowest for each item. This outcome raised the question of whether those who complied with the formula-scoring instructions were penalized by option-weighting compared to those who guessed when totally ignorant and received generally higher item scores for doing so.

Raffeld (1975) attacked this problem by eliminating examinees who omitted from the option-weighting procedure and assigning the average of the weights thus determined to the omissions. This procedure was compared to the usual option-weighting procedure, and it was found that the new procedure yielded lower internal-consistency reliability estimates but much higher validity estimates, slightly better than those from number-right scoring. Reilly (1975), apparently independently, proposed essentially the same approach as Raffeld and anticipated the same results based on a reanalysis of data from Reilly and Jackson (1973). To investigate this question further, Cross and Frary (1978) applied empirical option-weighting to groups taking tests under number-right and formula-scoring instructions. They also found very low item-scores for omissions under formula scoring. Both sets of option-weighted scores had better reliability and validity estimates than their conventional counterparts. However, these estimates were nearly equal for the two sets of option-weighted scores. In a subsequent study, Cross et al. (1980) used only number-right instructions and found the validity of the empirically option-weighted scores to be only about equal to that of the number-right scores. However, internal-consistency reliability was again found to be superior to that for number-right scores. Similar results had been observed by Downey (1979), who also used only number-right scores for comparison.

Explaining option weighting, either judgmental-based or empirical, to examinees could be difficult. Cross et al. (1980) reported success in their efforts to do so with college classes in chemistry and psychology, citing questionnaire results heavily in favor of option weighting. However, probably thinking of a less select examinee population, Echternacht (1976) followed his partially positive results with the question, “... how do you explain the scoring to an examinee?” (p. 309). At the same time, it would seem that option weighting offers no consistent basis for improving the psychometric quality of test scores unless there is a problem with respect to internal-consistency reliability. These factors along with the cost and complexity of option weighting have apparently contributed to its decline; there have been no major journal articles in this area for several years now.
Answer-Ulti Rect Procedure

This response mode requires the examinee to continue selecting options until the answer is chosen. The original medium was a board covered by a paper answer sheet with holes under the space for each option. The examinee punched out the paper until a hole painted red at the bottom was revealed for each item. Later versions have included erasable shields over indicators of whether each option is correct or incorrect and chemically treated answer sheets marked by sensitizing felt-tipped pens that reveal the same information for each option. Of course, these media make scoring extremely laborious. No way has ever been devised to score answer-unti correct responses using an optical mark reader. Computer presentation of answer-unti correct testing is altogether feasible but also somewhat costly in terms of the labor required to supervise administration. Apart from psychometric considerations, answer-unti correct testing has been recommended for its ability to provide immediate feedback to examinees, thus enhancing the learning process (Beeson, 1973; Pressey, 1950). However, other investigators have noted (relatively) lower test scores for group using answer-unti correct responding as compared to conventional responding (Montor, 1970; Strang & Rust, 1973). These differences may have been due to the unsettling effect of having one's ignorance revealed or could reflect inexperience with the method of responding.

A scoring rule for this method was first proposed by Brown (1965). Effectively, Brown's rule results in item scores equal to the number of options not chosen. Thus, for a 5-choice item, an examinee who chose the answer on the first attempt would receive 4 points, on the second attempt 3 points, and so on. It should be noted that this scoring rule rewards partial information more generously and penalizes misinformation (classifying the answer as a distractor) more harshly than number-right or formula scoring (see Frary, 1980). As noted by Israel (1979), an infinite number of other scoring rules for answer-unti correct testing are possible, depending on how one wishes to reward partial information and penalize misinformation, subject only to the restriction that it must not profit the examinee to make another response after identifying the answer. Musser and Thompson (1977) provided an example of use of a scoring rule other than Brown's.

Perhaps due to the scoring labor or the need for computerization associated with answer-unti correct testing, it has been involved in rather few empirical studies. Gillman and Ferry (1972) reported substantially higher internal-consistency reliability estimates for answer-unti correct scores compared to number right, as did Evans and Misfeldt (1974). In two studies, Hanna (1975, 1977) reported improved reliability but no differences for validity. The same result was reported by Whetton and Childs (1981). A simulation study by Frary (1982) showed only modest improvement in
internal-consistency reliability and no meaningful validity differences for answer-until-correct scores compared to number-right.

In the studies just reviewed, no attention was given to a characteristic of the test items having a substantial bearing on the effectiveness of answer-until-correct testing. Based on suggestions by Brown (1965) and Merwin (1959), Kane and Moloney (1978) showed analytically that the quality of answer-until-correct scores is enhanced by items that, if administered conventionally, would be rather difficult. In addition, there should be some gradation of the options, such that the best examinees would tend to choose the answer, followed by successively weaker groups of examinees attracted to each of the remaining options. Of course, this requirement places a substantial additional burden on the item-development process.

This result, along with difficulty of administration or scoring, has evidently led to a lack of enthusiasm at the practical level for answer-until-correct testing. At the same time, research by Wilcox (1981, 1982a, 1982b, 1983a, 1983b) has provided a different perspective from which to consider answer-until-correct testing. Most of what Wilcox did is not really germane to this review. Rather, he has shown various ways to use responses from an answer-until-correct test administration to address questions concerning a variety of psychometric properties of item and total scores from the same tests administered conventionally. He also derived two scoring methods for answer-until-correct responding, each of which can provide estimates of the number of items known outright by examinees (as opposed to giving partial credit when this level of knowledge is absent).

Item Response Theory (IRT) Methods

IRT models for weighting options were first developed by Samejima (1969). Her models are suitable only when the responses can be ranked in terms of correctness. A similar limitation applies to the adaptation of the Rasch model proposed by Masters (1982). Studies based on this line of research were the focus of a recent special issue of *Applied Measurement in Education* (Vol. 1, No. 4, 1988) and are not covered in this review.

Bock (1972) proposed an IRT model without the restriction that options be ranked and applied the model to responses to a vocabulary test, finding increased accuracy of ability estimation over the lower half of the score range. Thissen (1976) applied Bock's model to the Raven Progressive Matrices and found similar results but no meaningful increase in reliability overall. Huynh and Casteel (1987) compared pass–fail decisions by teachers with those determined by number-right scores and scores from Bock's model. They found no meaningful differences for tests of moderate length but, for very short tests, found substantial differences, with number-right scores much more closely coinciding with teacher determinations.
Multiple Correct Options

Many multiple-choice stems might be compatible with more than one correct option. If the number of correct options varies from one item to another, the examinee must evaluate every option as to whether it is correct or incorrect, in which case each option effectively becomes a true–false item. One way to score such tests is simply to count the correctly classified options. Applying formula scoring makes no sense in this case because every option is viewed as having been classified; if marked, the examinee classified it as true, if not, as false. (Some items may have no right answer, or the examinee may believe this to be true.) Alternatively, the examinee may actually respond true or false to each option with the possibility of omitting if uncertain. In this case, formula scoring may be applied. This approach to testing should provide better evaluation of an examinee’s partial knowledge than conventional responding, because the expected score gain over the options of an item would be proportional to the number on which the examinee is truly informed. This expectation was confirmed by Frisbie and Sweeney (1982), who reported higher reliability compared to number-right scoring when testing time was equalized for the two methods.

Serlin and Kaiser (1978) devised a weighting scheme applicable to responses to items with varying numbers of correct options. These weights are the eigenvalue elements for the first principle component of the intercorrelation matrix for all test options (scored 1 if correctly classified, 0 otherwise). Effectively, this method must be viewed as item weighting, with each option serving as an item. Willson (1982) applied the Serlin and Kaiser method and reported high internal-consistency reliability compared to that from a parallel test form with conventional items.

A variant form of items with differing numbers of correct options is one that lists options and then specifies combinations of these as the actual responses—for example, the first actual option might read, “A and B are correct,” the second, “B and D are correct,” and so forth. Albanese (1982) compared items of this type with direct responses to the primary options and found that providing combinations of responses spuriously inflated the scores. Mueller (1975) looked at item-analysis statistics for combinations of options and reported that their presence tended to make items more difficult but not more discriminating.

What has just been outlined should not be confused with the case in which each item has a fixed number of correct options and the examinees are informed of this number. Typically, the examinee responds by marking the same number of options as the number known to be correct. (A scoring method that could lead to marking more or fewer than this number would depend on examinee evaluation of partial information and would place the procedure under the examinee judgment methods to be reviewed.) T. C.
Hsu, Moss, and Khampalikit (1984) gave a good discussion of the scoring methods that may be used for tests of this type. They evaluated six methods empirically and found that those allowing partial credit yielded more reliable and valid scores. Ebel (1978) reported negative results for a very restricted case of this testing method. He referred to his approach as "multiple true-false," but this should not be confused with what was described at the beginning of this section (i.e., multiple correct options within a content-unified item). Instead, he had groups of statements, unrelated by content, each of which had one correct (true) statement. Examinees made one mark per "item" (group of statements) and received credit only if they marked the true one. L. M. Hsu (1979) examined the use of grouped true-false items theoretically and found that groupings of two or three true-false items would be more effective for discriminating among examinees of lower ability than the same items presented singly. However, he did not find this to be true for examinees of middle or higher ability.

Duncan and Milton (1978) considered the testing methods described in this section from a board theoretical perspective forming the background for several interesting results. They derived general classes of scoring rules to account for misinformation as well as partial information, embracing some of the examinee judgment methods (to be reviewed) as well as direct response methods. A minimax response strategy for examinees is derived for the case of a known number of correct options per item, along with an appropriate examiner response strategy. The examinee strategy for the case of an unknown number of correct options per item is derived under the assumption that the examinee has some prior belief (in the Bayesian sense) about the likely number of correct options.

EXAMINEE JUDGMENT METHODS

Confidence Testing

Although Echternacht (1972) included all examinee judgment methods under the term "confidence testing," the present review includes only those requiring the examinee to express some specific level or degree of confidence in addition to answering an item. Early studies of confidence testing used essentially arbitrary rules assigning response categories for levels of confidence and using these responses in the determination of item scores. Echternacht pointed out that most of these methods are flawed in that, for examinees who can really evaluate how likely they are to answer correctly, there are optimal ways of responding that will not be generally obvious. The results from the earlier studies of this type reviewed by Echternacht were generally unpromising.
Confidence testing was given new life by the work of Shuford, Albert, and Massengill (1966). They devised scoring systems with the characteristic that examinees would maximize their scores if and only if they responded in a manner consistent with their personal evaluations of their probabilities of answering correctly. Such a scoring system is said to have the reproducing property. Independently, DeFinetti (1965) suggested theoretical bases for reproducing scoring systems. Echternacht (1972) gave an excellent summary of these developments, and Lord and Novick (1968, pp. 321–324) presented additional theoretical details.

Abu-Sayf (1979) provided a good update and supplement to Echternacht's (1972) review. As noted there, additional investigations have confirmed Echternacht's concern about the extent to which examinees can objectively evaluate their confidence as to the correctness of options (see Echternacht, Boldt, & Sellman, 1972; Koehler, 1974). In addition to this concern, Linn (1976) also questioned the extent to which examinees can be properly motivated to do their best under confidence testing.

Comparing confidence testing with other modes has yielded mixed results. Kansup and Hakstian (1975) instructed examinees to distribute a "kitty" of 10 points over the options of each item in proportion to how strongly they felt that each might be the answer. Five scoring methods were applied, the simplest of which was to assign as an item score the number of points assigned to the answer. This method was at least as effective as the others, but none was clearly superior to number-right scoring with respect to reliability and validity. Hakstian and Kansup (1975) reported additional use of the method just described. Compared to number-right scoring, they noted improved reliability, but the additional testing time required for the confidence testing would have permitted lengthening the number-right test sufficiently to overcome this difference. Pugh and Brunza (1975) also found improved reliability but not validity from confidence testing. Poizner, Nicewander, and Gettys (1978) compared number-right scoring with a reproducing scoring method and one for which the examinees ranked the options according to perceived likelihood of being the correct answer. In this latter method, they received for item scores the ranks given the correct answers. One phase of this study was especially interesting in that the extent of partial information available to the examinees was controlled. Outcomes are consistent with the levels of partial information, and both confidence methods yielded somewhat higher reliabilities than number-right scoring.

The difficulty of entering response data for the computer analysis required by virtually all confidence-testing procedures has limited practical use of this method to computerized testing. Due to the attendant costs, this testing has been limited to high-budget educational enterprises such as those in medicine and the military. Rippey has been a leader in this area and pioneered the development of the necessary computer systems (Rippey &
Donato, 1978). Rippey and Smith (1979) reported an interesting procedure by which they measured the realism of examinees in evaluating their probabilities of answering correctly. This led to adjustment of confidence test scores for realism (Rippey & Voytovich, 1982), with attendant improvement of reliability. Rippey and Voytovich (1983, 1985) also related various aspects of confidence testing to instruction-related phenomena for medical students, such as the tendency to jump to conclusions in diagnosis or to entertain obvious misinformation. Thus, in summary, it appears that confidence testing has found a narrow niche in the world of measurement. Moreover, it seems that its use is more appreciated for secondary qualities than for possible improvement of the psychometric characteristics of scores.

Subset-Selection Methods

Subset selection refers to examinee designation of a subset of the options of an item with the proviso that the examinee knows how many options are keyed as correct. (See the Multiple Correct Options section, presented earlier, for the case in which the examinee does not know.) Instructions may call for the subset to include all the options the examinee believes to be distractors or they may call for inclusion of all options the examinee believes necessary to ensure inclusion of the correct option(s). These two approaches are entirely analogous logically; the options not in the subset believed to be distractors are simply the ones the examinee believes must include the answer(s). Studies based on subset selection were first reported by Coombs (1953) and by Dressel and Schmid (1953), neither referring to the other's work. Coombs stated that the idea was not original with him but did not attribute a source. Coombs's method called for examinees to mark all options they were sure were distractors, with the award of 1 point for each so marked. However, if an examinee inadvertently marked the answer (assuming one correct option per item), \( C - 1 \) points were deducted from the number of distractors marked \( (C = \text{number of options}) \). The Dressel and Schmid study instructed examinees to mark as many options as they felt necessary to include the (single) correct answer, receiving 1 point for inclusion of the correct answer and \(-1 / (C - 1)\) points for each distractor marked. For a given state of knowledge (e.g., knowing only that a single option is a distractor), the Coombs item score can be obtained by multiplying the Dressel and Schmid item score by \( C - 1 \). Hence the two methods are equivalent. They penalize an examinee for guessing after assured identification of one or more distractors and, like answer-until-correct scoring, reward partial information more generously than number-right or formula scoring (see Frary, 1980). Early subset-selection studies reported in the reviews by Wang and Stanley (1970) and Echternacht
(1972) were fairly uniform in reporting improvements in internal-consistency reliability but no improvement in validity compared to number-right or formula-scored versions of the same tests. Later studies using the scoring methods of Coombs or of Dressel and Schmid have reported a considerable variety of results with respect to reliability and validity (see Bradbard & Green, 1986; Collet, 1971; Hakstian & Kansup, 1975; Koehler, 1971; Sax & Collet, 1968). A simulation study by Frary (1982) produced results similar to most of the empirical studies—namely, improved reliability but not validity with respect to number-right or formula scoring.

Plake, Thompson, and Lowry (1981) evaluated responding by the Coombs method and under number-right scoring with respect to (a) order of item difficulty (random or progressively more difficult), (b) examinee information level (whether or not examinees were informed of the item difficulty order), and (c) varying levels of test anxiety. They found no significant differences with respect to the mode of responding and any of these independent variables. Jaradat and Sawaged (1986) compared scores from the Dressel and Schmid method with number-right and formula scores, taking into consideration examinees' achievement levels and propensity to take risks. They reported better reliability and validity for the Dressel and Schmid scores and found that, unlike formula scores, they did not tend to favor high risk-takers.

Other studies have used the responding method of Coombs or Dressel and Schmid but other scoring rules. Arnold and Arnold (1970) used the Coombs response method and assigned for an item score the expected formula item-score arising from the state of knowledge suggested by the examinee's Coombs response. For example, suppose an examinee identifies two distractors for a 4-choice item and has no idea which of the remaining two is the correct answer. Under formula scoring, if this examinee guesses, the expected item-score is \( \frac{1}{3} \) point, and this was assigned by Arnold and Arnold if the examinee marked the two distractors and left blank the other two options. This method does not (on average) penalize for guessing as is the case for Coombs or Dressel and Schmid scoring. Arnold and Arnold used this method for classroom testing and reported good student acceptance but did not investigate the psychometric characteristics of the scores. Their method was derived anew by Austin (1981) using less restrictive assumptions. He also showed that under certain conditions the Arnold and Arnold item-score variances would be less than those arising from conventional testing with formula scoring. This characteristic should lead to improved reliability, which was confirmed in the simulation study by Frary (1982). Jacobs (1975) compared the Arnold and Arnold method with that originated by Coombs and found that examinees tended to identify more options as distractors under the Arnold and Arnold instructions than under
the Coombs instructions. Diamond (1975) instructed examinees to mark as many options as needed to ensure inclusion of the answer, with scoring a multiple of the Dressel and Schmid method if the answer was included but simply zero if it was not. In addition, examinees were asked to rank the options from *most likely* to *least likely* to be the answer. The rankings were a numerical expression of confidence as required for inclusion of this study with the confidence testing studies already reviewed. However, the rankings were not used for scoring; those receiving a mark of 1 were used to determine number-right scores, which were compared with the others for reliability and validity. The subset-selection scores had higher reliability but about the same validity as the number-right scores.

In addition to the theoretical work of Duncan and Milton (1978), previously mentioned with respect to direct response methods, two other theoretical articles bear on subset-selection methods. Gibbons, Olkin, and Sobel (1979) derived expected means and variances of item scores arising from responding under Coombs or Dressel and Schmid conditions. This derivation is based on assuming average probabilities (for any examinee) of including the answer in a subset \( p \) and of eliminating each distractor \( p' \). The analysis then continues to seek the number of items needed to estimate the examinee's true state of knowledge, defined as \( (p + p') / 2 \), to any desired degree of accuracy. This process is then repeated for number-right responding. Surprisingly, it turns out that the relationship between the numbers of items needed under subset-selection versus number-right conditions is independent of the degree of accuracy. Not surprisingly, subset selection is always the more efficient method. This process is also carried out for items with two correct choices.

Hutchinson (1982) proposed theoretical bases for evaluating partial knowledge that are entirely distinct from any others reviewed here. His approach is based on continuous random variables reflecting the inappropriateness of each option as an answer. It is further assumed that each examinee has a threshold such that an option will be classified as an answer if the threshold exceeds the value of the inappropriateness function. If more than one option has this characteristic, the one exceeding the threshold the most is chosen. Depending on the inappropriateness functions used, several very different ways of evaluating latent ability arise. (Hutchinson chose different functions for right versus wrong options.) All these allow for the response of omitting when the threshold does not exceed any of the values of the inappropriateness functions across the options. Although this approach is not within the bounds of this review, it is reported due to the interesting analysis by Hutchinson of how the various subset selection and confidence methods relate to his theory. Specifically, he noted that the various subset-selection methods are inconsistent with some versions of his
clearly plausible theoretical formulation. Further consideration of the interaction between Hutchinson's theory and the subset-selection methods may be very revealing.

CONCLUSIONS

What characterizes the empirical studies reviewed here perhaps the most is their frequently contradictory results. What appears to work well in one setting does not in another or in a replication. In many cases (often unavoidably), the study designs did not allow for sample splitting or other means of cross validation, but, even so, after so many studies, a picture emerges. It appears that the methods covered by this review singly or jointly provide no panacea or even consistent potential for worthwhile improvement of the psychometric properties of test scores. Even the confirmed ability of some methods to enhance internal-consistency reliability is devalued by one or more of the following: reduced validity, increased time required for testing, scoring complexity, difficulty of explaining the scoring to examinees and other users, and difficulty of explaining the response mode and training examinees in its use.

This negative outcome should not have been entirely unexpected. After all, number-right and even formula scoring do not leave partial information unrewarded. Their way of rewarding it is different from and more subject to chance fluctuation than the partial credit methods reviewed here. Therein lies any hope for partial credit scoring. It may be less subject to chance error and may be able to take advantage of any unusual relationship between partial information and a criterion. Unfortunately, the empirical studies have not found this potential to be widespread, although a few studies may have happened on special circumstances in which partial credit scoring may be desirable. Many of these cases will probably involve secondary characteristics of partial credit scoring, such as its ability to provide various kinds of examinee feedback that may enhance learning (e.g., confidence and answer-until-correct testing). In general, however, the search for better test-score quality might better be directed toward areas such as adaptive or tailored testing (if computerized administration is feasible) or toward the use of classical and IRT methods to upgrade score quality in tests administered and score conventionally.

REFERENCES


Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.


Echternacht, G. J., Boldt, R. F., & Sellman, W. S. (1972). Personality influences on


reasoning by computer-assisted confidence testing. *Journal of Computer-Based Instruction, 9*, 88–97.


