Alternative Response and Scoring Methods for Multiple-Choice Items: An Empirical Study of Probabilistic and Ordinal Response Modes

Sharon B. Poizner, W. Alan Nicewander, and Charles F. Gettys
University of Oklahoma

Binary, probability, and ordinal scoring procedures for multiple-choice items were examined. In a situation where true scores were experimentally controlled by the manipulation of partial information, it was found that both the probability and ordinal scoring systems were more reliable than the binary scoring method. A second experiment using vocabulary items and standard reliability estimation procedures also showed higher reliability for the two partial information scoring methods relative to binary scoring.

The usual goal of a performance test is to estimate accurately how much the examinee knows at the time of examination. In the case of multiple-choice tests, examinees choose those answers which they consider best; and each test item is scored correct or incorrect. Very often examinees know something about a question, but are not completely sure of the answer. In this situation examinees may not receive credit for their partial information if the item is answered incorrectly, or they may receive more credit than they deserve if their answer to the item is correct. On any particular item, therefore, performance when examinees are incorrect is indistinguishable from performance when examinees have no information; and when examinees are correct, their performance cannot be distinguished from those who have perfect information. States of partial knowledge cannot be measured and revealed by binary scoring, a method which seems neither to meet the needs of the examiner completely nor to be fair to the examinee. There are, however, other alternatives which potentially can remedy the imprecision of the binary procedure. Two of these alternatives, which will be discussed here, are the probabilistic response and scoring system and an ordinal response and scoring system.

A probabilistic system has been suggested for multiple-choice items; it has the potential for extracting more information about the ability of the respondents than the usual responding and scoring procedures (de Finetti, 1965; Shuford, Albert, & Massengill, 1966). The probability system requires examinees to assign a personal probability estimate to each alternative of a multiple-choice item, indicating their degree of belief regarding the correctness of the alternative, so that the sum of the probabilities for each item equals one. For example, suppose probability scoring is used on a multiple-choice item which asks one to choose the word which is closest in meaning to "whelp," where the alternatives are puppy, snail, scratch, grindstone, duck. Examinees can reveal their degree of belief by assigning, for example, .60 to "puppy," .40 to "duck," and .00 to the remaining alternatives. In this in-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 2, No. 1 Winter 1978 pp. 83–96

stance, the probabilities would indicate that the examinee has eliminated all but two alternatives and believes "puppy" to be more correct than "duck."

Thus, the examinee's state of partial knowledge on each item can be expressed by the subjective probabilities assigned. If the examinees accurately assign subjective probabilities to the alternatives of multiple-choice items, the examinee who thinks that three alternatives are implausible on, for example, a five-choice item will score higher than the examinee who eliminates only two. Not only does the probability system extract partial knowledge, but, in theory, it also eliminates guessing. Through personal probability estimates examinees can express their exact state of knowledge about an item. The problem is not to draw a line between knowledge and guessing, but to ascertain the examinees' degree of knowledge as reflected by their statements of certainty or uncertainty (de Finetti, 1965). It is not only important to measure what examinees know on a collection of items, but also important to know how they perform on an item-to-item basis. Probability scoring is technically termed a reproducing scoring system, because it reproduces the examinees' present state of knowledge about each item when it is used accurately.

Although a probability response system is the most informative response and scoring procedure, it is also the most technically complex. It requires detailed instructions on responding and scoring, the conceptualization and understanding of these procedures, and the consequences of the responses on the part of the examinee. Also, if the score the examinee earns bears a linear relationship to the probabilities, it can be shown that examinees can maximize their expected score by not reporting their subjective beliefs. Rather, they can assign probabilities of 1.0 to the alternative they believe to be most correct and zeros to the remaining alternatives, thereby destroying the reproducing properties of the scoring system. It then becomes necessary to introduce a pay-off function which persuades the examinees to be honest.

Theoretically, one of the most defensible and practical scoring formulae for a probability system is the truncated logarithmic function of Shuford et al. (1966), where:

$$\text{Item Score} = F(x) \qquad [1]$$

$$= \begin{cases} 1 + \log x & \text{if } .01 \leq x \leq 1 \\ -1 & \text{if } 0 \leq x < .01 \end{cases}$$

where x is the probability assigned to the correct alternative. Suppose an examinee has assigned .80 to the correct alternative; by applying the scoring rule in Equation 1, $f(.8) = 1 + \log(.8) = .92$. For that item the examinee's item score would be .92. Using this pay-off function, it can be shown that examinees will maximize their expected score (with respect to personal probability distributions) only when they assign honest probability estimates to each alternative.

Underlying a probability response system is a theoretical continuous distribution of states of knowledge, and a probability procedure has the potential for extracting the exact state of knowledge. However, as Lord and Novick (1968, ch. 14) have noted, the complexity of the probability system may introduce more error variance than ability-related variance. In theory, a probability system has the potential of avoiding the measurement error inherent in a binary system. In theory, probability scoring is perfectly reliable, because it reproduces perfectly what a person knows; thus it reduces the error variance in the observed score to zero. In practice, however, probability scoring may introduce undesirable variability through errors in using the system, misunderstanding of instructions, or failing to realize the consequences of one's response. A probability system is likely to be somewhat difficult to teach to most examinees and could not be taught to some. All of the above practical problems would contribute to error-related variance. Therefore, if the examinee does not understand and correctly utilize the scoring procedure, it is clear that test reliability will be less than perfect.

Although the probability scoring system is theoretically optimal, practical difficulties in its implementation may be such that a theoretically sub-optimal system will yield higher reliabilities. Such a sub-optimal system is an ordinal response and scoring system (de Finetti, 1965). Because it is somewhat less complex than the probability system, the examinee should have less difficulty in using it efficiently. As a result of the reduction in complexity compared to probability scoring, there is a concomitant reduction in the amount of information potentially available from the examinee. The ordinal nature of such a system destroys the distance information which the examinee may possess and which he/she can reveal through a probability response. Using a simple version of this discrete scoring system on a five-choice item, examinees "rank" the alternatives from 5 to 1, assigning 5 to the alternative they believe to be most correct and 1 to the alternative they believe to be least correct.

An item of this type is most simply scored by recording the number the examinee assigned to the correct alternative. Consequently, the possible scores for a five-choice item are 5, 4, 3, 2, and 1. Under an ordinal system, examinees can employ no strategy to maximize their expected score other than honest response; and no pay-off function is necessary. Ordinal responding and scoring is obviously much easier to teach to examinees and can be taught to younger children. Guessing is still possible with the ordinal method and is, therefore, a source of error variance. Since binary scoring captures none of the partial information possessed by examinees and ordinal scoring does, the total error variance may be less for ordinal than binary scoring. The loss in ability information of ordinal scoring relative to probability scoring may be offset by reduction in error variance due to the complexity of the probability method.

The binary system is the least complex, most often used, and least informative response and scoring system. A major disadvantage of binary scoring is its inability to distinguish between partial and complete information. The binary procedure is a response system which is extreme in its information-destruction compared to probability and ordinal scoring. Another severe drawback to the binary method is the problem of guessing. One attempt to deal with this problem has been the introduction of corrections for guessing. Many of these corrections are based on the unrealistic assumption that in the absence of perfect knowledge, the examinee guesses randomly among the alternatives (de Finetti, 1965). Examinees who have either partial information about an item or misinformation do not respond at random, and alternatives do not appear equally attractive to the examinee. A simple correction formula for guessing based on the random guessing model would be inappropriate in these cases (Lord & Novick, 1968, ch. 14).

The three scoring systems previously mentioned have been discussed theoretically (de Finetti, 1965; Shuford et al., 1966); but their psychometric properties have not been studied empirically, although Coombs, Milholland, and Womer (1956) have studied other related procedures. The crucial issue of the psychometric properties of these three scoring systems needs to be empirically addressed. As Lord and Novick (1968, p. 314) point out, "in evaluating any new response method, it will be necessary to show that it adds more relevant ability variation to the system than error variation, and that any such relative increase in information retrieved is worth the effort." The present study, through a series of three experiments, attempts to address these issues by investigating the efficacy of using the three systems—probability, ordinal, and binary—in a testing situation. An empirical evaluation of the three scoring systems has the unique advantage of investigating the tradeoff between the potential theoretical advantages of some scoring systems and the complexity of response modes which can reduce any potential advantage.

Experiment I was an evaluation of the set of instructions designed to teach probability scoring. Experiment II used an experimental ap-

proach to obtain empirical data on the three scoring systems. A large number of subjects were assigned and trained for one of three experimental conditions, each condition being one of the three response and scoring systems. Each subject's level of knowledge was controlled experimentally. Identical tests were administered to all subjects who, using one of the three scoring systems, responded with the controlled amount of knowledge available. Because of the nature of this experiment, normally unobservable information became available, and reliabilities could be computed for each scoring system.

Due to the somewhat artificial nature of a controlled experiment, Experiment III was designed to study the three scoring systems in a realistic testing situation. The vocabulary section of the Lorge-Thorndike Intelligence Test was administered to students trained in all three scoring and response modes. Reliabilities were estimated and again a comparison was made among the probability, ordinal, and binary systems.

Method: Experiment I

This experiment was a preliminary and crucial part of the study. The complexity of the probability system necessitated a detailed set of instructions for the subjects. Without adequate instructions in the probability response method, deficiencies in the probability scoring system could be attributable to either poor instructions or the method itself. Since a lack of understanding of the scoring procedure would lower reliability, the aim of this experiment was to try to eliminate misunderstanding as a possible source of error variance. For these reasons Experiment I was designed to pretest the instructions for probability scoring.

Subjects

Subjects for the pretest were 32 students drawn from the pool of introductory psychology students at the University of Oklahoma.

Materials and Procedure

Materials consisted of a four-page set of instructions on the probability system (Poizner, 1974). The instructions included a description of the drawbacks to a binary system, a detailed explanation of probability estimation, a description of the necessity of being honest, and the consequences of the payoff function associated with the probability method. A table was also provided listing possible probabilities assigned to the correct alternative from .00 to 1.00 (in steps of .05), along with the corresponding score obtained from the truncated log function. Two exemplary multiple-choice items were then presented with typical answers and obtained scores for three hypothetical students. A discussion followed concerning how these students should have responded, given certain assumptions about their states of knowledge. Finally, the subjects were instructed to provide probability estimates for two items. After being given the correct answer by the experimenter, subjects were told to look on the previously described table for their obtained scores.

An eight-item performance test was then administered to test the subjects' knowledge of a probability estimate, the necessity of responding honestly, and the payoff function.

Results and Discussion

The mean number of items that were answered correctly by the 32 subjects was 7.59 out of a possible 8; the modal number of correct responses was 8. It was concluded that the instructions were sufficient and understandable—a necessary condition for further work using these instructions.

Method: Experiment II

A unique approach in the empirical study of response and scoring systems is the use of a controlled experiment. Experiment II employed a test covering obscure topics in which the subject's level of knowledge was experimentally

manipulated by providing varying amounts of information sufficient to eliminate a controlled number of item alternatives. Each controlled knowledge level created a state of partial knowledge. Subjects, assigned to partial knowledge levels and trained in an assigned responding and scoring system, took an eight-item test and responded with the appropriate response method. Because partial knowledge was controlled, the mathematics for computing the expected reliabilities of the three scoring systems in the experimental setting was quite tractable. Therefore, the obtained reliabilities of the three scoring procedures could be compared with the expected reliabilities.

Subjects

Subjects were 507 students from introductory psychology at the University of Oklahoma. The subjects who participated in Experiment I were not eligible for this experiment.

Materials

Materials for this experiment consisted of a set of instructions on the scoring system to be used by the subject. For those using the probability method, the same four-page instructions were used as were pretested in Experiment I. In addition, a 16-item test was constructed which covered four obscure topic areas; information about the topic areas was provided on fact sheets. There were four different forms of each fact sheet, each providing different amounts of information necessary to answer the related questions. In some instances the information provided was bogus, and in some it was actual. The obscure topics were Advanced Neuroanatomy, History and Construction of Kentucky Rifles, the Life and Works of Charles Sherrington, and the Nature and Type of Nervous Disorders of Soldiers Who Fought in the Korean War. The topics were selected in order to minimize or eliminate any previous knowledge of the topics.

Procedure

One hundred and sixty-nine subjects were assigned to each of three groups, P, R, and B, where Group P responded under probability scoring, Group R under ordinal scoring, and Group B under binary scoring. Each group of 169 subjects was run in subgroups of no more than 20 subjects each. There was a common set of instructions for the test. In addition, there was a set of specific instructions detailing the appropriate scoring and response modes which were read by the subjects and recited by the experimenter. Each subject in each of the three groups was provided with a fact sheet which placed the subject in one of four predetermined states of knowledge or partial information —called Θ_1, Θ_2, Θ_3, or Θ_4. Subjects assigned to a knowledge level of Θ_i had fact sheets which gave them information sufficient to eliminate definitely i of the five alternatives as incorrect. Only 8 of the 16 items on the test were scored. The remaining 8 "filler" items, for which the subject's fact sheets gave more or less knowledge than the particular condition to which he or she was assigned, were included to prevent subjects from realizing that for all items a fixed number of item alternatives could definitely be eliminated as incorrect. Therefore, the design for the experiment was a completely crossed 3 × 4 design with three scoring procedures and four levels of partial knowledge within each scoring method. It should be emphasized that all subjects completed the same test, but under different scoring conditions and under differing amounts of information or knowledge. A sample test item actually used is provided below with the corresponding facts for the four levels of partial knowledge.

Example: Sample Question and Corresponding Facts for Four Θ Levels

Test question for all information levels. A Kentucky Rifle can be identified by three unique

features. These unique features are:

a. Rifled bore, fancy patchbox, and maple stock
b. Flint ignition, fancy patchbox, and long barrel
c. Fancy patchbox, maple stock, and flint ignition
d. Rifled bore, fancy patchbox, and flint ignition
e. Flint ignition, long barrel, and maple stock

Facts given to groups (each Θ represents the information given to a particular group).

Θ_1: (information sufficient to eliminate one alternative) One of the three unique features of a Kentucky Rifle is its flint ignition.
Θ_2: (information sufficient to eliminate two alternatives) Two of the three unique features of a Kentucky Rifle are a flint ignition and fancy patchbox.
Θ_3: (information sufficient to eliminate three alternatives) One of the three unique features of a Kentucky Rifle is its rifled bore.
Θ_4: (information sufficient to eliminate four alternatives) There are three unique features about Kentucky Rifles—rifled bores, fancy patchboxes, and flint ignition.

In each of the three groups of 169 subjects, 15 subjects were randomly assigned to knowledge level Θ_1, 33 to level Θ_2, 106 to level Θ_3, and 15 to Θ_4, so that the (simulated) marginal probability distribution of Θ_1, Θ_2, Θ_3, and Θ_4 was .0888, .1953, .6272, .0888, respectively. The number of subjects assigned to each of the four partial knowledge conditions were chosen to reflect a probability distribution for Θ which, in theory, should produce distinctly different reliabilities for the probability, ordinal, and binary testing procedures (Nicewander, 1973). This distribution was also chosen because it produces a "moderately difficult" binary item.

Scoring procedures. For the probabilistic scoring method, items were scored by applying the truncated logarithmic function in Equation 1 to the subjective probabilities assigned to the correct alternatives. Under the ordinal scoring procedure, subjects assigned the integers 5, 4, 3, 2, and 1 to item alternatives, with 5 assigned to the "best" alternative, and 1 to the "worst" alternative. Each item score was determined by the integer assigned to the correct alternative. The standard scoring procedures were followed with binary scoring—items were scored 1 if correct and 0 if incorrect.

Derivation of Expected Reliabilities

It is important to know the reliabilities that would theoretically be obtained if the subjects were capable of understanding and using the scoring systems optimally and utilizing the partial information properly. Such theoretical reliabilities are the expected reliabilities under the assumption of optimal usage of information provided and optimal usage of each scoring system. In this experiment control of the subjects' knowledge through utilization of fact sheets implies that the true scores of the subjects were also experimentally controlled.

Following the classical theory of test scores (Lord & Novick, 1968), a true score for an individual is defined as the expected value (over replications) of his/her observed score. In Experiment II the levels of partial knowledge were the only source of true variance among the subjects (assuming that they had no previous knowledge of the obscure topics tested and that all the subjects used the scoring systems in the optimal way). Therefore, fixing level of partial knowledge fixed the true scores for all individuals assigned to a given level of partial knowledge.

The varying number of subjects assigned to each partial knowledge level for Groups B, P, and R generated a probability distribution for Θ_i. Hence, the true scores corresponding to Θ_i had exactly the same probability distribution as Θ. For all subjects in Groups B, P, and R it was assumed that those assigned to Θ_i eliminated (on the basis of their fact sheets) i of the five alterna-

tives to the multiple-choice items as incorrect; they were assumed to be indifferent regarding the $5 - i$ remaining alternatives. It was further assumed that the correct alternatives were never eliminated as incorrect. Given these assumptions, the conditional probability distributions of item score and information level for all three scoring methods can be easily derived.

For example, consider information level Θ_2 across all scoring procedures. These subjects could eliminate two alternatives as incorrect; they were indifferent regarding the remaining three alternatives. Therefore, for information level Θ_2 under binary scoring, subjects should have randomly selected one of the three non-eliminated alternatives as the correct response. The probability of obtaining a score of "0" would then have been .3333. The probability of obtaining a score of "0" is simply 1 minus the probability of obtaining a score of 1, given Θ_2 or .6667. For ordinal scoring under Θ_2, subjects using the procedure optimally should randomly assign integers of 1 and 2 to the two alternatives they could eliminate as incorrect. Since it was assumed the correct alternative was never eliminated, the probability of obtaining a score of 1 or 2 is zero (i.e., these integers would never appear on the correct alternative). Left indifferent among the remaining three, one of which was the correct answer, the probability of obtaining the scores of 5, 4, or 3 were all equal to .3333. Under the assumption of optimal usage of probability scoring, each subject in Θ_2 should have eliminated two alternatives by assigning a probability of zero to each and should have assigned a probability of .33 to the remaining three. There should be no variability among subjects' responses under probability scoring in this experiment. All subjects in condition Θ_2 should have assigned a probability of .33 to the correct alternative, which translates into an item score of .5229 using Equation 1. The deduction of all other conditional probabilities, $p(x_j|\Theta_i)$, followed this logic and are presented in Table 1 along with the true scores for each scoring method.

Due to the nature of the experiment, there were only four values that true scores could assume under each scoring system, i.e., the four fixed levels of partial knowledge, Θ, had a single associated true score for each scoring system. The values of the true scores for each scoring system were computed as the expected values of the conditional probability distributions. For each scoring system, the joint probability distributions of observed item scores and partial knowledge level were computed from the relationship,

$$p(x_j,\Theta_i) = p(x_j|\Theta_i)p(\Theta_i) \qquad [2]$$

$$i = 1, 2, 3, 4$$

$$x_j = 1, 0 \text{ for binary scores}$$

$$x_j = 1, 2, 3, 4, 5 \text{ for ordinal scores}$$

$$x_j = .3979, .5229, .6990, 1.0000 \text{ for probabilistic scores}$$

where $p(x_j, \Theta_i)$ is the joint probability of the item score x_j and partial knowledge level Θ_i; $p(x_j|\Theta_i)$ is the conditional probability of the item score x_j, given knowledge level Θ_i; and $p(\Theta_i)$ is the marginal probability for partial knowledge level Θ_i. Notice in Equation 2 that the probabilistic scores are restricted to only four values: .3979, .5229, .6990, and 1.00. In theory, the probabilistic score is continuous in the interval (0, 1); however, in the present experiment, partial knowledge was controlled so that optimal use of the information provided would yield only four scores. For example, subjects assigned to knowledge level Θ_3 were able to eliminate three of the five alternatives for any item as incorrect, leaving them indifferent among two—one of which was the correct alternative. Optimal usage of the probabilistic method would dictate that these subjects assign probabilities of 0 to the three incorrect alternatives and .5 to the two plausible alternatives. Therefore, a personal probability of .5 should appear on the correct alternative; and .5 transformed using Equation 1 yields an item score of

Table 1

Expected Conditional Probability Distributions for Item Scores and Information Levels along with Expected True Scores

	Information Level			
Item Score	Θ_1	Θ_2	Θ_3	Θ_4
Binary Scoring				
1	.25	.33	.50	1.0
0	.75	.67	.50	0
(True Scores)	(.25)	(.33)	(.50)	(1.0)
Ordinal Scoring				
5	.25	.33	.50	1.0
4	.25	.33	.50	0
3	.25	.33	0	0
2	.25	0	0	0
1	0	0	0	0
(True Scores)	(3.5)	(4.0)	(4.5)	(5.0)
Probability Scoring				
1.0000	0	0	0	1
.6990	0	0	1	0
.5229	0	1	0	0
.3979	1	0	0	0
(True Scores)	(.3979)	(.5229)	(.6990)	(1.0)
Marginal Probability Distribution for Information Levels (Θ_i)	.0888	.1953	.6272	.0888

Note. These probabilities and true scores are those expected under optimal usage of both the information provided and the scoring method.

.6990. Similar logic was used to produce the remaining three values of the probabilistic score corresponding to the three other values of Θ_i.

Table 2 presents the joint and marginal probability distributions for the three scoring methods. Again, all probability distributions are those expected under optimal usage of both the partial information provided and the scoring system.

The expected marginal probability distributions for the observed scores for each scoring method presented in Table 2 were computed by summing the rows of the joint distributions of the observed scores and information levels. All the necessary information is now available for computing the expected observed and true-score variances for an item under each scoring method. The observed score variance for any of the scoring methods was computed as

$$\sigma^2(x) = \sum_j p(x_j)x_j^2 - \left[\sum_j p(x_j)x_j\right]^2 , \quad [3]$$

where $p(x_j)$ is the marginal probability for the item score x_j. Similarly, the true score variances for each scoring method were computed as

$$\sigma^2(T) = \sum_i p(T_i)T_i^2 - \left[\sum_i p(T_i)T_i\right]^2 , \quad [4]$$

or

$$\sigma^2(T) = \sum_i p(\Theta_i)T_i^2 - \left[\sum_i p(\Theta_i)T_i\right]^2 , \quad [5]$$

since the marginal probabilities for the true scores, $p(T_i)$, are equal to the marginal probabilities for the information levels, $p(\Theta_i)$. Finally, the reliabilities for a single item under each scoring method were computed by taking the ratio of true to observed score variance.

As the eight items composing the complete tests were statistically parallel given the assumptions, the Spearman-Brown equation was used to compute the predicted reliabilities for eight-item tests administered under the three scoring methods. The predicted reliability was .7345 for the ordinal scoring procedure and .5577 for the binary scoring. If the probabilistic system were used optimally by the subjects in this experiment, every subject at a given level of partial knowledge should receive exactly the same score, e.g., subjects in Θ_1 could eliminate one alternative to which a probability of zero should have been assigned. All remaining alternatives (one of which was the correct answer) should have been assigned probabilities of .25. Under the assumption of optimal usage of the probability system, the variances of all conditional distributions for the observed score are zero. As a consequence the error variance is zero, which implies perfect reliability. From this argument, the predicted reliability of the probability scoring was 1.0.

Results and Discussion

The obtained reliabilities for the eight-item test were then computed from the squared correlation between the obtained and true scores for the eight-item tests and are presented in Table 3, along with the expected reliabilities. Under probability scoring, reliabilities were computed for the 169 subjects and also for 167 subjects, eliminating two extreme scores for subjects who, on the basis of their responses, clearly misunderstood the system or did not cooperate. The elimination of these two subjects substantially increased the obtained reliability of probability scoring.

As indicated in Table 3, there was a substantial gain in reliability of both the probability and ordinal scoring systems over the binary system. When the two aberrant subjects in the probability group were dropped, the reliabilities of the rank-order and probability methods were approximately the same. Using the Spearman-Brown equation to predict the reliabilities for a 25-item test, the gain in reliability by probability and ordinal scoring was very evident. The length of the binary test would have to be increased 1.52 times (adding 13 items) to attain the reliability of the test using ordinal scoring and 1.63 times (adding 16 items) to attain the reliability of the test using probability scoring. Although

Table 2

Expected Joint and Marginal Probability Distributions for Item Scores and Information Levels

Item Score	Joint Probability Distributions of Item Score and Information Level				Marginal Probability Distributions for Item Scores
	Information Levels				
	Θ_1	Θ_2	Θ_3	Θ_4	
Binary Scoring					
1	.0222	.0651	.3136	.0888	.4897
0	.0666	.1302	.3136	0	.5104
Ordinal Scoring					
5	.0222	.0651	.3136	.0888	.4897
4	.0222	.0651	.3136	0	.4009
3	.0222	.0651	0	0	.0873
2	.0222	0	0	0	.0222
1	0	0	0	0	0
Probability Scoring					
1.0000	0	0	0	.0888	.0888
.6990	0	0	.6272	0	.6272
.5229	0	.1953	0	0	.1953
.3797	.0888	0	0	0	.0888

Note. These probabilities are those expected under optimal usage of both the information provided and the scoring method.

Table 3

Theoretical and Obtained Reliabilities for the Eight-Item Test

	Scoring Method		
Reliability	Probability	Ordinal	Binary
Theoretical	1.0	.7345	.5577
Obtained	.5912 (.6795)[a]	.6639	.5655

[a]Reliability eliminating two deviant subjects.

somewhat more time per item is required with the more complex scoring methods, the slight increase in testing time produced a substantial improvement in reliability over the binary method.

This experiment supports the feasibility of the alternative response and scoring procedures as desirable alternatives to the conventional binary method. Taking into account the ease of teaching, grading, and understanding, an ordinal response and scoring system seems to fare quite well. It is also important to point out that both the probability and ordinal methods were very palatable to the subjects. There are limitations to this study, however. The simulated situation is obviously not entirely representative of reality. There are an infinite number of levels of knowledge, and this study only used four. In this limited artificial situation, the subject was given the necessary information to eliminate certain alternatives and remain indifferent among the remaining alternatives. In reality, the behavior of an examinee is not so easily represented. An examinee may be able to definitely eliminate certain alternatives as incorrect and may feel other alternatives to be fairly probable or improbable, with differing degrees of certainty. Therefore, a third experiment studied the three scoring systems in a more realistic testing situation.

Method: Experiment III

As in other controlled laboratory studies, the aim of this study was to be able to generalize the results of Experiment II to applied settings; and in Experiment III this goal was attempted. Three randomly parallel forms of a vocabulary test were administered to subjects. Each form was administered under a different scoring method. Reliabilities of the three scoring systems were then estimated and compared.

Subjects

Subjects for this experiment were 63 upper division psychology students at the University of Oklahoma. Approximately 20 subjects were randomly assigned to each of three groups, and the data were collected in small groups of approximately 10 subjects each.

Materials

Materials consisted of a set of instructions for each subject describing all three scoring methods. The instructions were taken from the three separate sets of instructions used in Experiment II. The three tests were constructed using a pool of 24 items taken from the vocab-

ulary section of the Lorge-Thorndike Intelligence Test. Eight items were randomly assigned to three forms.

Procedure

Each subject took all three forms of the test. A Latin Square design was used to randomly assign a different scoring system for each form and a random order of the forms. The set of instructions were administered to a group, and all three scoring systems were explained. Each person received three forms in the predetermined order with eight items per form. Immediately preceding each form were the instructions designating which scoring system was to be used on that form and a concise summary of that scoring system.

Results and Discussion

Reliability estimates were computed for the three scoring systems for each group, and an average was taken across groups. These average reliabilities are presented in Table 4. Table 4 also shows the average item difficulty for each form, along with the reliabilities of each scoring system computed on that form. The reliability estimate used was Guttman's λ_3 (Guttman, 1945; Lord & Novick, 1968, p. 94) and was chosen because it is a slightly better reliability estimator than the more commonly used Coefficient Alpha.

Again, probability and ordinal scoring surpassed the binary method in terms of reliability, with the probability method yielding the highest average reliability. In general, as the difficulties of the forms increased, the reliabilities of the ordinal and binary methods increased, and the reliability of probability scoring decreased. There are, however, some considerations in interpreting these results. In Experiment III upper division psychology students were used as subjects and seemed to better understand the probability scoring system on first exposure. The higher reliability of the probability method may in part reflect greater sophistication of subjects. Also, using the Latin Square assignment procedure, individual reliabilities were estimated from only 20-23 subjects, and the average reliabilities are based on 63 subjects; therefore, care should be exercised in interpreting the results.

Experiment III does seem to support the conclusions of Experiment II. Additionally, Experiment III demonstrates the feasibility of using the alternative scoring methods in realistic testing situations.

The reliabilities for the rank and probability

Table 4

Average Item Difficulty and Corresponding Reliabilities for Each Form

Form	Average Item Difficulty	Reliability by Scoring Method		
		Probability	Ordinal	Binary
A	.83	.40	.66	.63
B	.63	.60	.35	.21
C	.66	.70	.37	.19
Average reliabilities for eight-item test		.5658	.4593	.3409

scoring systems may increase with greater familiarity with the systems. This should certainly be the case for probability scoring where the subjects in this experiment were attempting to master a difficult scoring system for the first time. If either ordinal or probability scoring were used frequently, the reliabilities for these systems might show an even greater advantage over the reliability of the traditional response and scoring method.

Conclusions

The results of these studies support the contention that the probability and ordinal scoring methods may prove superior to binary scoring for some purposes and for some examinees. These results, while not definitive in many respects, do suggest that alternatives to the binary scoring system are worthy of further study. The question of which scoring system to use in a given situation is still open. However, some tentative recommendations can be made on the basis of these results in the absence of more definitive information.

The choice of a particular response mode would seem to be governed by the interaction of the following factors: (1) the potential gain or loss of reliability associated with the response mode; (2) the characteristics of the examinees, such as age and intelligence; (3) the time required to master the response mode; (4) the palatability of the response mode to the examiner and the examinees; and (5) the availability of computerized scoring. Each of these considerations will be discussed in turn, and interrelationships among them will be noted.

The potential gain in reliability afforded by the probability and ordinal response modes is considerable in the college population studied. With more training in the probability response mode, college students would likely show an even greater gain in reliability. The ordinal mode shows near optimal reliability in the second experiment (theoretical reliability = .73, obtained = .66); consequently, there is little room for improvement. Examinees were using the ordinal system about as well as they could. One would expect, however, that with less sophisticated examinees the potential superiority of the probability mode would not be realized; older grade-school children might be expected to master the ordinal mode easily, but might well be incapable of understanding probability responding. In this case the reliability of the probability mode might be lower than even the binary method. Both the probability and ordinal response modes were readily accepted by the examinees during the experiment.

One potential difficulty with both the probability and the ordinal scoring modes is the increased labor in scoring the results of the examination. Manual scoring of probability responses is almost prohibitively time-consuming. Even though ordinal responses can be scored fairly easily, the scoring takes considerably more time than manual scoring of binary responses. However, the advent of on-line computerized testing makes both the probability and ordinal procedures feasible, as the computer can easily score either response mode.

Our general conclusions, although very tentative, would be to recommend the ordinal scoring mode for use with adults or secondary school students, if the gain in reliability so obtained outweighs the additional effort of scoring. In a computerized testing situation, either the ordinal or the probability response mode might be useful. The ordinal response is so easily taught that it would be best for episodic testing. In contrast, the potential gains in reliability afforded by the probability mode might best be exploited in computerized instruction, where the time spent teaching the mechanics of a probability response could be prorated over many testing situations.

References

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

de Finetti, B. Methods for discriminating levels of

partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology,* 1965, *18,* 87–123.

Guttman, L. A basis for analyzing test-retest reliability. *Psychometrika,* 1945, *10,* 255–282.

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

Nicewander, A. *A theoretical comparison of the reliability of right-wrong scoring of multiple-choice tests vs. a rank-order responding and scoring system.* Paper presented at the meeting of the Midwestern Psychological Association, Chicago, 1973.

Poizner, S. *Alternative responding and scoring methods for multiple-choice tests.* Unpublished master's thesis, University of Oklahoma, 1974.

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

Author's Address

Alan Nicewander, Department of Psychology, University of Oklahoma, Norman, OK 73019.