Proceedings of the Iowa Academy of Science

Volume 75 | Annual Issue Article 44

1968

Discrimination Weighting on a Multiple Choice Exam

Timothy J. Gannon
Loras College

Thomas Sannito
Loras College

Discrimination Weighting on a Multiple Choice Exam

TIMOTHY J. GANNON[1] and THOMAS SANNITO[2]

Abstract. A multiple choice test (59 items) was given to 141 sophomore males as a final examination in General Psychology. The students were required to indicate their degree of confidence on each item by writing to the left of their answer a *1* (pure guess), a *2* (some guessing—some certainty), or a *3* (complete certainty). The scoring involved making the confidence values positive or negative according to whether the selected answers were correct or incorrect. Three scoring procedures were then compared by correlation techniques to determine if there were any major shifts in student ranking. It was found that by scoring the number right or by the algebraic total of confidence scores there was very little change in rankings. It was recommended that the multiple choice test using discrimination-in-weighting scores had advantages over either traditional number-right scoring or formula scoring. The discrimination-in-weighting score took into account the students' degree of confidence on each item, thus allowing him to determine his own weighting without causing a major change in his class rank.

Correction factors on multiple-choice exams are intended to eliminate erroneous gains for guessing. All incorrect answers are given some negative value, usually a fraction of —1 (—¼, —½, etc.). Then the negative values are added together and deducted from the total number right. The correction fraction or negative weighting assigned to each wrong answer seems to be somewhat arbitrary and removed from the examinee's control. According to statistical probabilities (the changes of guessing correctly), the examiner decides to subtract a fraction of the total number of wrong answers from the total number of right ones. Psychometrists have never been completely satisfied with this technique.

Doppelt (1954) contends that wrong answers are not given in accordance with the laws of chance probability, because there are various degrees of guesswork on wrong answers. Hence, the assumption that the reduced score obtained by subtraction of a correction factor yields a more accurate indication of the student's knowledge might be false. Since the examiner does not know how much guessing is involved, in some cases the deduction will be too much and in others too little.

Little (1962) gave a biology examination as a pre-test, a final examination, and a post-test to 16 college students. His hypothesis was that wrong answers are not always marked in accordance with the laws of chance. Wrong answers are chosen with various degrees of guesswork. He reasoned (a) that if a student consistently made a wrong response on all three testing occasions he would not be simply

[1]Professor and Chairman; [2]Instructor and Research Supervisor—Department of Psychology, Loras College, Dubuque, Iowa.

"stabbing in the dark," and (b) that his guessing was based upon some nebulous information. He found that the *S*s were consistently wrong on the three testings on some items and wrong on other items two out of three testings. He concluded that there are various degrees of guessing on multiple choice items and that the wrong answers should not all be scored with the same correction factor. He contended that the use of such correction factors should be abolished.

Lord (1963) presents a theoretical critique of correction scoring. He concluded, like Little, that formulae or correction scorings do not deduct accurately for the various degrees of guesswork involved in a wrong response. This scoring procedure does not discriminate between the possible degrees of guessing on items and then make appropriate deductions. The amount of guessing on any item is known only by the student. All wrong answers are treated alike (assigned the same deduction) by the examiner, regardless of the degree of confidence.

The three previously cited authors seem to agree that correction factors are inaccurate because the degree of guessing on items remains obscure, and, hence, that all wrong answers resulting from different degrees of guessing are scored the same. It also seems reasonable that right answers are chosen with various degrees of assurance. These answers should likewise be assigned different scores to improve the accuracy of measurement.

The experimenters of the present study attempted to isolate the degree of certainty of *S*s of each item on a multiple choice examination. Students made gains or losses on items according to their correctness in answering and their degree of confidence. The examinees were required to choose an answer and indicate numerically their certainty from 1 to 3 (1 = complete guess and 3 = complete assurance). Each certainty value was then scored either plus or minus, depending upon whether or not the answer was correct. This departure from the traditional methods of administering and scoring multiple choice exams was intended to accomplish two things: (a) to eliminate the somewhat capricious selection of a correction formula used in scoring and (b) to shift the total responsibility to the student for determining the weighting on each item.

The main issue involved was whether this innovated technique of subjective weighting would result in a major shift in the rankings of students, compared to the usual scoring of the number right only. If the rankings showed only minor position changes, then, this new method of discrimination in weighting might influence the students to abandon the old attitude "I have nothing to lose by guessing and everything to gain."

Methods

Subjects

The subjects were 141 males enrolled in General Psychology at

Loras College. The students were from two divisions of the introductory course and were alll taught by the same person. Two semesters of introductory psychology are required for all students. Since psychology is usually taken in the sophomore year, most of the *S*s were second-year students, approximately 20 years of age.

Psychometric Device

At the end of the first semester of General Psychology, a test was constructed to measure the over-all knowledge for the whole semester's course study. The semester examination had 59 multiple choice items. Each item was reduced to three options, the right answer and two wrong alternatives. A sample item was "The chief motor pathways are found (a) in the central part of the cord; (b) in the lateral columns of the cord; (c) in the dorsal portion of the cord." It is customary in tests of this nature to include three or four distraction alternatives along with the correct answer. For simplification, only three choices were presented.

Procedure

Administration. All students in both divisions took the test at the same time. The directions were mimeographed at the top of the examination. Instructions on the test required the examinee to indicate the best alternative which would complete the statement, by placing an *a, b,* or *c* on the blank to the left. Furthermore, they were directed to indicate carefully their degree of confidence (discrimination weighting) by a *1, 2,* or *3* written to the left of the letters. The number *1* next to a student's answer would mean almost complete uncertainty on the item. The number *3* was to be used in the case of virtually absolute certainty, and *2* was meant to reflect some degree of certainty and uncertainty. Before the test administration, the *S*s were fully briefed that their scores would be determined not only by their choice of answers, but also by their confidence values. The students were informed that scoring would involve placing a plus or a minus sign next to each certainty score, depending upon whether they chose the correct answer. For a correct response, the certainty or confidence value would become positive, and for an incorrect one it would become negative. The time limit on the examination was two hours.

Scoring. The scoring of the discrimination-in-weighting (called certainty or confidence for *S*s) test was quite simple. If an answer was correct, a plus sign was placed before the student's indicated confidence value. If the answer was incorrect, his confidence number was given a minus sign. When all items on a given test were scored plus or minus, then the algebraic combination of all positive and negative numbers became the person's score. The possible range of scores on the present test was -177 to $+177$.

Statistical Analysis

Three different ways of scoring the test were intercorrelated by the Pearson Product and Spearman Rank methods. One scoring technique was to sum the total number of right answers. Another procedure involved algebraically combining positive and negative discrimination scores for a total value. The third process was to add the total number right to the total discrimination score. The means, standard deviations, z-score ranges, and correlation coefficients were computed by a computer.

Results and Discussion

The means, standard deviations, and z-score ranges for three different scoring procedures of a multiple choice examination are presented in Table 1. The smallest mean, 39.40, is that obtained with the conventional scoring method of summing the number of correct responses.

Table 1

Means, Standard Deviations, and z-Score Ranges

Scoring Procedure	Mean	S. D.	z-Score Range
Number Right[a]	39.40	6.93	5.194
Discrimination Scores[b]	56.61	34.89	5.816
Discrimination Scores[c] + Number Right	96.01	40.60	5.442

[a]This scoring procedure involved summing the number of correct answers.

[b]This scoring procedure involved making the discrimination value plus for a correct answer and minus for an incorrect one. These positive and negative values were algebraically combined to yield a total score.

[c]This scoring procedure involved adding the number of correct answers to the total discrimination score.

The mean for the discrimination scoring is 56.51. It is not surprising that the mean of the discrimination scores is higher than the average number right. The *S* was required to discriminate between three weighting-values on each item (*1*, *2*, and *3*), according to his certainty in knowing the correct answer. Then the *S*'s score is plus or minus his discrimination value on an item. If a student is careful, his correct answer should gain more for him than his incorrect one loses for him. He can gain 3 on the certain items and lose only 1 on the unsure ones.

The greatest mean, 96.01, is that for the scoring method in which the discrimination scores were added to the total number of right answers. This technique of scoring should, of course, produce the highest average value.

The standard deviations for scoring the number right, the discrimination-in-weighting scores, and the discrimination-in-weighting scores plus the number right are respectively 6.93, 34.89, and 40.60. By using either method of scoring discrimination values, in contrast to scoring the number of correct answers only, the dispersion among scores is greater. This increase in variability might make the assignment of letter grades easier.

In Table 2, estimates of relationship obtained by Pearson's correlation procedure and by rank difference method as given for each possible paring of three different methods of scoring a multiple choice examin. Both measures of correlation between the total number right and the discrimination scores are .79 (significant at the 0.01 level).

Table 2

Correlation Coefficients Between Three Scoring Procedures

Variables Correlated*	N	r	rho**
NR and Discr.	141	.79***	.79***
NR and Discr. + NR	141	.85***	.87***
Discr. and Discr. + NR	141	.99***	.99***

*Under the heading "Variable Correlated" the abbreviation "NR" represents scoring the total number of correct answers on the test; "Discr." represents the algebraic sum of the discrimination scores; "Discr. + NR" means that scoring was done by adding the total number of correct responses on a subject's test to the algebraic total of his discrimination values on the test.

**Spearman *r*.

***Significant at the 0.01 level.

The correlation indices suggest that most students do not change their rank greatly on a test when they are required to determine the weighting for each item in contrast to when they simply choose the correct answer. Hence, a confidence scoring method will, in general, maintain the same ranking of the students as will a conventional scoring method which is based on only the number right, but the former method will magnify the differences between students according to their certainties.

Since the traditional unit of measurement has been the number of correct responses, the number right was added to the discrimination-in-weighting scores as a third method of scoring. Obviously, the correlation coefficients were higher between the number right and the discrimination scores plus the number right ($r = .85$; $p = .87$) than between the number right and the discrimination scores ($r = .79$; $p = .79$). By adding the number right to the discrimination scores, the correlation coefficient is purposely increased. Although the improvement in correlation seems artificial because the number right is part of both series of numbers, it appears god design that the students' rankings will change even less from the number right than when just the discrimination scores are used without the number right added in. Hence, the change in ranks from the number right scoring to the number right plus discrimination scores is minimal, while the dispersion of scores is greater.

The correlation between the discrimination plus number of right answers and discrimination was .99 by both correlation procedures

(Table 2). That the deviation from a perfect relationship is so slight is not surprising.

Perhaps, then, a discrimination-in-weighting, "student centered," scoring method is better than the method of simply using the number of right answers and allowing for capricious selections. The results of this study seem to suggest that the best method might be a combination of the number right plus the discrimination or confidence scores, since this procedure shows the least departure in rankings from the straight number of right scoring and an increased showing of heterogeneity (scatter) based upon confidence on items.

Literature Cited

Doppelt, J. E. 1954. Test Service Bulletin of the American Psychological Corporation (No. 4). New York.
Little, E. B. 1962. J. Educ. Res., 55, 245-252.
Lord, F. M. 1963. Educ. & Psychological Meas., 23, 663-672.