DOCUMENT RESUME

ED 073 172 TM 002 422

AUTHOR Echternacht, Gary
TITLE A Comparison of Various Item Option Weighting Schemes.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RB-73-6
PUB DATE Jan 73
NOTE 18p.; A Research Bulletin draft

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Item Analysis; Measurement Techniques; Multiple Choice Tests; Performance Criteria; *Scoring; *Scoring Formulas; Standardized Tests; Tables (Data); Technical Reports; Test Construction; Test Interpretation; *Test Reliability; Test Results; Test Validity; *Weighted Scores

ABSTRACT

This study compares various item option scoring methods with respect to coefficient alpha and a concurrent validity coefficient. The scoring methods under consideration were: (1) formula scoring, (2) a priori scoring, (3) empirical scoring with an internal criterion, and (4) two modifications of formula scoring. The study indicates a clear superiority of the empirically determined scoring system with respect to both coefficient alpha and the concurrent validity. (Author)

RESEARCH BULLETIN

RB-73-6

A COMPARISON OF VARIOUS ITEM OPTION WEIGHTING SCHEMES

Gary Echternacht

Educational Testing Service
Princeton, New Jersey
January 1973

A COMPARISON OF VARIOUS ITEM OPTION WEIGHTING SCHEMES

Gary Echternacht

Educational Testing Service

Abstract

This study compares various item option scoring methods with respect to coefficient alpha and a concurrent validity coefficient. The scoring methods under consideration were: (1) formula scoring, (2) a priori scoring, (3) empirical scoring with an internal criterion, and (4) two modifications of formula scoring. The study indicates a clear superiority of the empirically determined scoring system with respect to both coefficient alpha and the concurrent validity.

A COMPARISON OF VARIOUS ITEM OPTION WEIGHTING SCHEMES[1]

Gary Echternacht

Educational Testing Service

One of the essential goals in measurement systems research is to extract as much information as possible from a given set of items. This allows the test constructor to use fewer items in a test, while retaining a previously set reliability standard. This, in turn, is especially desirable in the case where items are difficult and/or expensive to construct. The general problem of increasing the amount of information from an item requires examining one or all of three components: (1) how the examinee is to respond to the item, (2) how an item is scored, and (3) how the items are put together to form a total score.

If one assumes that the multiple-choice format that now exists will be continued in use for some time in the future and considers that past research with weighting items differentially has proven unfruitful (Stalnaker, 1938; Wilks, 1938), one concludes that the most productive area of research lies with investigating various scoring methods or, in other words, differential weighting of options of an item. There are two different general methods of weighting item options most often accepted. One involves empirically weighting options using some internal or external criterion, the other an a priori weighting of the options.

Weighting by using some internal or external criterion dates back to the 1920's when Strong began work on his interest inventory (Strong, 1943). This type of criterion keying, usually with an external criterion,

---

[1]This study was sponsored by the Graduate Record Examinations Board.

was used mostly with self-report types of items. Strangely, the question of differentially weighting the options of achievement and ability items to improve reliability has received little attention. This has been true since: (1) the use of an external criterion with achievement and ability is time consuming, expensive, and somewhat prone to error in the criterion measure; and (2) obtaining weights with minimal sampling variance requires a large amount of data and much computation.

As might be expected, a priori weighting of test items (with differing weights for distractors) has not been widely practiced. Gage (1957) and Yee and Kriewall (1969) have used a priori scores on the Minnesota Teacher Attitude Inventory with an effectiveness equal to that when the more elaborate criterion keying was used. Davis and Fifer (1959) used both forms of option weighting in raising the cross-validated comparable-forms reliability of a specially prepared arithmetic-reasoning test. Other than that, there appear to be few noteworthy attempts to use a priori option weighting. Although not generally thought of as being a priori, because equal weights are given to all distractors, the usual formula score, along with its modifications, is an a priori system.

In contrasting empirical weighting of options and a priori weighting of options, weighting empirically seems to suffer from one major difficulty. The examinee does not know the consequence of his action when responding to any given item or, in other words, he does not understand the scoring system being used, a defect that appears to this writer to be somewhat unethical. One would be inevitably asked the question why person A received a score of X on a given item and person B received a score of Y on the same question even though both answered incorrectly. One

would be hard pressed to answer with anything satisfactory to the examinee.

It seems to this writer that the most fruitful search for a simple, easily understood, ethical system of differential weighting lies with a priori weighting. This process has some problems of its own though. For example, most a priori option weighting studies have utilized a panel of judges for supplying the weights, which introduces a further source of error into the weighting system and can be somewhat expensive. It would seem more desirable if the test item writer could specify the weights in some predetermined manner as he developed the various distractors. The work of Elizur (1970) and Guttman (1965) with facet design has provided an indication that this might prove to be a fruitful method for constructing a priori weights. Also, there is some question as to whether item writers can construct items using facet design though that question was not investigated in this study.

Purpose and Procedure

The purpose of this study is to determine the effectiveness of various item option scoring schemes, especially empirical and a priori schemes, in relation to formula scoring and some of its modifications. Since the ultimate goal is to shorten the test and retain the same degree of reliability, the reliability of the test under these various scoring schemes becomes the prime measure of effectiveness. Thus, the reliabilities obtained under the different scoring schemes will be of prime importance.

Of secondary importance (only in this instance) is the question of validity. In a study such as this, it was not feasible to collect any completely adequate criterion measure although a similar (not parallel)

test of greater length was thought to be useful for obtaining a concurrent validity statistic. Such was possible in the operational structure of this study, and the correlation between the experimental test score using the various scoring methods and the longer test served as a validity check. This study was conducted through the operational framework of the Graduate Record Examinations (GRE) program, with the experimental test embedded within the GRE Aptitude Test, which served as one of the regular pretest sections. The items were quantitative in nature, and the longer, similar test mentioned above consisted of the regular GRE quantitative test section.

In order to determine the effectiveness of the scoring system, six random samples were drawn from the total number of examinees taking the specially designed test form. Values of coefficient alpha were calculated for each scoring system on each sample. In addition, correlations between the main section quantitative test score and the special test score were obtained for each scoring system.

Test Construction

A 30-item quantitative pretest section was constructed especially for this study. This pretest section appeared in the June 1972 administration of the GRE Aptitude Test. The 30 test items were written completely by the Educational Testing Service Test Development Division. According to specifications provided by the study director, they were instructed to construct items with one correct answer, two distractors differing from the correct answer in only one aspect (one error in logic or operation) and two distractors differing from the correct answer in more than one aspect. The distractors with only one error were termed "first order" distractors, while the remaining were termed "second order" distractors. The item

writers kept a log of the time required to write and review the items, so the additional costs for writing such items could be computed.

In the a priori scoring scheme, no attempt was made to differentiate between the two first order distractors. The same was true for the second order distractors.

Scoring Systems under Consideration

The usual scoring system for GRE tests is to give one point for a correct answer, zero for an omit, and -1/4 for an incorrect answer. Thus, the formula scoring system becomes a baseline system for making comparisons. Since it is extremely easy to construct, a rights only scoring system was also used.

The a priori scoring system was developed with the following properties in mind: (1) the scoring system should use integer scores; (2) the expected score under random guessing should be zero; and (3) the intervals between the scores should be equal, excluding the omit score. Thus, a scoring system was used that gave the score of 6 to a correct answer, a score of 1 if a first order distractor were chosen, and a score of -4 for selection of a second order distractor. All omits were scored as zero.

The procedure for obtaining empirical scores for each item option, including omit, was to use the keying for internal consistency procedure found in Reilly and Jackson (1972) which is similar to that of Hendrickson (1971). The computational details will not be given here, but basically the process consists of first scoring the test using the conventional scoring formula (rights - 1/4 wrongs); secondly, assigning the weight determined by the mean standard score on the remaining items for all persons choosing that option; and finally, computing coefficient alpha. The

procedure can be used iteratively until coefficient alpha appears to stabilize, although Reilly (personal communication) notes that such iterations fail to change coefficient alpha by any sizable amount, at least when the test is already fairly reliable to begin with.

Although the expressed purpose of this study was to compare a priori, empirical, and formula scoring methods, it was relatively easy to add two others that were modifications of formula scoring. The motivation for including these scoring systems in this study was that they had recently appeared in the literature, and there were no empirical results where these systems were used. It was also very inexpensive to incorporate these systems into the design. The two systems were recently developed by Zinger (1972) and are termed Z1 and Z2 . The Z1 scoring system gives a score of one to a correct answer and a score of $-c$ to an incorrect answer. The value c is determined by

$$c = \sum_{i=1}^{a-1} n_i^2 \Big/ \Big(\sum_{i=1}^{a-1} n_i\Big)^2 \qquad (1)$$

where a indicates the number of alternatives, n_i the number of examinees responding to the ith distractor, and the summations are taking over the distractors.

The Z2 system gives a score of $1 + b$ to a correct answer, $-c$ if the answer is incorrect. The value b is determined by

$$b = \sum_{i=1}^{a-1} (n_i - \bar{n})^2 \Big/ \Big(n_T \sum_{i=1}^{a-1} n_i\Big) \qquad (2)$$

where n_T indicates the number answering correctly and

$$\bar{n} = \sum_{i=1}^{a-1} n_i \Big/ (a - 1) \quad . \qquad (3)$$

These two scoring systems are based on the concept of "ideal" items as presented by Weitzman (1970) and provide a correction for guessing that takes into account a nonuniform distribution of wrong answers assumed by the formula scoring. In essence, the distractors are weighted more negatively as the distribution becomes more nonuniform. In both these systems omitted items were given a score of zero.

For the empirical weights and the Z1 and Z2 weights an initial sample is needed for calculation of the item option weights. The weights thus obtained are then used in each subsequent scoring replication.

Sampling

As stated previously, the pretest section was spiralled and thus was taken at most test centers across the country. Since seven independent samples were required, it was decided to use a two-stage process in making sample selections. The first stage consisted of selecting test centers, while the second stage consisted of selecting students within a test center. Actually, a test center represents a fairly good primary sampling unit as the students in these centers tend to be somewhat homogeneous with respect to undergraduate institution and geographic region.

Further, it was decided to balance the sample with respect to ability level as measured by the number correct on the regular quantitative test section. Thus, cutting scores were developed, using the entire sample, for classifying any individual into the lower, middle, or upper third in quantitative ability as measured by the number correct on the regular section. Also, it was of interest to have some samples completely female and others completely male in makeup. A two by three table (sex X ability level) was conceptualized for further selecting test centers.

A count of the number of test centers having X individuals in each cell of the table described above was made. The number of test centers was further broken down by geographic region (Census Bureau classification) for each value of X . X varied from one until it was so large that no test center had at least X in each cell.

The first sample was designed to be a base sample for calculating the various weights involved in the empirical scoring and the Z1 and Z2 systems. This sample had to be at least 2,500 in number since it was desired to have the standard deviations of the estimated proportion of people responding to a particular alternative be less than .01. It was also desirable to select as many centers as possible to make up this sample in order to represent as wide a range as possible. Therefore, all centers having at least two candidates per cell were selected for the base sample or sample 1. There were 231 such centers. Within each center, candidates were classified into the six cells and two candidates were selected using simple random sampling. The resulting sample size for the base sample was 2,772.

Six samples of size approximately 1,000 were to be selected for computing efficiency. Of these, two were to consist entirely of females, two entirely of males, and the remaining two balanced. Sample 2 (female) and sample 4 (male) were selected by sampling 112 of the 186 test centers having at least three candidates per cell. This sampling of test centers was carried out using proportional allocation over the four geographic regions. Sampling within test center was accomplished using simple random sampling as before. The resulting females were termed sample 2; the resulting males, sample 4.

Samples 3 (female) and 5 (male) were obtained as were samples 2 and 4, only centers having at least four per cell were selected. A total of 85 out of the 151 possible test centers were so selected.

Samples 6 and 7 were mixed with equal numbers of males and females selected from the five and six per cell centers. A total of 34 of 130 five per cell centers were selected, while 28 of 109 six per cell centers were selected.

Results

The GRE aptitude test is a moderately speeded test (Swineford, 1968), which creates a number of problems in determining empirical weights. The problem usually occurring is that the omit score becomes extremely large negative and the validity of the test is reduced (see Reilly & Jackson, 1972) even though alpha is increased. In examining a preliminary item analysis of the special test section, it became apparent that the special test was also speeded. In fact only about 17% of the examinees finished the 30 items!

In order to reduce the effect of speed it was decided to eliminate some of the items from the special test for the analyses. A response rate of 90% for the entire test was felt necessary. By examining the item analysis, it was determined that 92% of the examinees finished the first 18 items. Thus, only the first 18 items of the special test were scored.

The resulting values of coefficient alpha for the scoring systems under study on each replication appear in Table 1. As can be seen the

---

Insert Table 1 about here

---

maximum coefficient alpha is obtained when empirical weights are used. This is not unexpected since the empirical weights tend to maximize coefficient alpha. What is important is that these values are substantially higher for empirical weights, equivalent to increases of 31.6%, 32.5%, 30.8%, 35.7%, 34.9%, and 32.5% in the test length when formula scoring is used. On the other hand, the a priori and Z1 and Z2 systems did not equal the performance of formula scoring

The correlations with the main section quantitative score closely resemble the results of the coefficient alpha calculations at least in pattern. These correlations are presented in Table 2.

Insert Table 2 about here

These results are in contrast to those found by Reilly and Jackson (1972), where a decrease in validity was found. It should be pointed out that they used undergraduate grades as a criterion measure rather than another test as was done in this study, so that the findings of these two studies are not contradictory, but rather illustrate different choices of criteria.

A few further points regarding the conduct of this research should be pointed out so that the conclusions resulting from this study can be taken in proper context. One key area that has been ignored up until this time is the conditions under which the experimental test was taken. The examinees were given instructions for the usual formula scored test sections. It was not possible to use directions specifically designed for a priori option weighting, saying that the respondent could receive some form of "partial credit" for wrong answers--because it was believed that by

introducing such directions, examinees would recognize the section as being experimental and be less likely to respond in earnest. It was also believed that such directions would increase administrative costs. The result of using these directions certainly contributed to the poor showing of a priori option weighting.

Another related point is that by directing the Test Development Division to construct distractors of differential quality may have given empirical keying an edge over the conventional method. Certainly the variances of the option weights were highly stable given the sample size and method of item construction (see Echternacht, 1973).

Cost

The cost of constructing a priori items was significantly higher than that for the traditional items. In general, the cost of constructing the a priori items ran about 60% greater. Thus, for the a priori method to prove cost-effective, an increase in reliability would have to be obtained that would allow the 18-item test to be reduced to an 11-item test. Such an increase in reliability was not noted in this study.

Conclusions

It becomes obvious that, in this case, the a priori option weighting was inferior to that of empirical option weighting. In fact, a priori option weighting did not even measure up to traditional formula scoring with respect to reliability on the items. Thus it appears that by using only empirical option weights, one can cut the cost of developing items (reduce the length of the test) and maintain standards of reliability, at least in the case of the GRE Aptitude Test.

There are still many details that need to be worked out before such a procedure can become operational. For example, how do you explain the scoring to an examinee? What should his strategy be? And also, what can or should be done with an item where a wrong answer receives more weight than the correct answer (one such item appeared in the special section)?

References

Davis, F. B., & Fifer, G. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. Educational and Psychological Measurement, 1959, 19, 159-170.

Echternacht, G. J. A note on the variances of empirically derived option scoring weights. Research Bulletin 73-5. Princeton, N. J.: Educational Testing Service, 1973.

Elizur, D. Adapting to innovation. Jerusalem, Israel: Jerusalem Academic Press, 1970.

Gage, N. L. Logical vs. empirical scoring keys: The case of the MTAI. Journal of Educational Psychology, 1957, 48, 213-216.

Guttman, L. A. A faceted definition of intelligence. In R. Eiferman (Ed.), Studies in psychology, scripta hierosolymitana, Vol. 14. Jerusalem, Israel: The Hebrew University, 1965.

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

Reilly, R. R., & Jackson, R. Effects of empirical option weighting on reliability and validity of the GRE. Research Bulletin 72-38. Princeton, N. J.: Educational Testing Service, 1972.

Stalnaker, J. M. Weighting questions in the essay-type examination. Journal of Educational Psychology, 1938, 29, 481-490.

Strong, E. K., Jr. Vocational interests for men and women. Stanford, Calif.: Stanford University Press, 1943.

Swineford, F. Test analysis, Graduate Record Examinations Aptitude Test. Statistical Report SR-68-36. Princeton, N. J.: Educational Testing Service, 1968.

Weitzman, R. A. Ideal multiple-choice items. Journal of the American Statistical Association, 1970, 65, 71-89.

Wilks, S. S. Weighting systems for linear functions of correlated variables when there is no dependent variable. Psychometrika, 1938, 3, 23-40.

Yee, A. H., & Kriewall, T. A. New logical scoring key for the Minnesota Teacher Attitude Inventory. Journal of Educational Measurement, 1969, 6, 11-14.

Zinger, A. A note on multiple-choice items. Journal of the American Statistical Association, 1972, 67, 340-341.

Table 1

Values of Alpha for Each Scoring Scheme and Sample

	Sample					
Scoring	2	3	4	5	6	7
Number Correct	.800	.820	.835	.828	.824	.823
Formula Score	.788	.806	.823	.814	.809	.810
A priori	.773	.791	.807	.799	.798	.797
Empirical	.830	.847	.859	.856	.851	.850
Z1	.779	.797	.815	.805	.800	.801
Z2	.774	.792	.809	.799	.795	.796

Table 2

Correlations between the Experimental and Regular Quantitative Test Sections for Each Scoring Scheme and Sample

	Sample					
Scoring	2	3	4	5	6	7
Number Correct	.850	.846	.859	.847	.843	.848
Formula Score	.854	.847	.855	.8451	.844	.844
A priori	.840	.841	.849	.838	.833	.838
Empirical	.862	.864	.873	.8591	.849	.858
Z1	.851	.841	.849	.840	.840	.838
Z2	.847	.837	.846	.835	.838	.835