DOCUMENT RESUME

ED 073 173 TM 002 423

AUTHOR Echternacht, Gary
TITLE A Note on the Variances of Empirically Derived Option Scoring Weights.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RB-73-5
PUB DATE Jan 73
NOTE 10p.; 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; Technical Reports; *Test Construction; Test Interpretation; *Test Reliability; Test Results; Test Validity; *Weighted Scores

ABSTRACT

Estimates for the variance of empirically determined scoring weights are given. It is shown that test item writers should write distractors that discriminate on the criterion variable when this type of scoring is used. (Author)

RESEARCH BULLETIN

RB-73-5

A NOTE ON THE VARIANCES OF EMPIRICALLY DERIVED OPTION SCORING WEIGHTS

Gary Echternacht

Educational Testing Service
Princeton, New Jersey
January 1973

A NOTE ON THE VARIANCES OF EMPIRICALLY DERIVED OPTION SCORING WEIGHTS

Gary Echternacht

Educational Testing Service

Abstract

Estimates for the variances of empirically determined scoring weights are given. It is also shown that test item writers should write distractors that discriminate on the criterion variable when this type of scoring is used.

A NOTE ON THE VARIANCES OF EMPIRICALLY DERIVED OPTION SCORING WEIGHTS[1]

Gary Echternacht

Educational Testing Service

In recent years, the developers of large-scale testing operations have shown an increasing interest in reducing the length of time examinees are required to spend on a given test. Reducing the test administration time would both reduce the cost of developing the test forms, as fewer items would be required, and allow time for additional tests to be administered. This thinking has characterized many of the test programs administered at Educational Testing Service, and, most likely, at other testing establishments. Researchers have thus sought new scoring methods that would result in increases in reliability due solely to the scoring system used. Thus, test length could be reduced, and a previous standard of reliability maintained.

One such scoring method that has proven successful in reliability studies is that of empirically deriving scoring weights (Davis & Fifer, 1959; Echternacht, 1973; Hendrickson, 1971; Reilly & Jackson, 1972; Strong, 1943). If empirically derived scoring weights were to be adopted by such large-scale testing programs as the College Entrance Examination Board, the Graduate Record Examinations, the Law School Admission Test, and other programs, one problem that would have to be faced is that of determining the variances of the derived weights and the implications these variances have for developing test items. This is necessary

---

[1]This research was supported by the Graduate Record Examinations Board.

on repeated occasions, and the scoring weights would only be developed on the initial administration. Since some examinees would not be included in the initial scoring run, the problem of scoring weight variance exists. Also, by knowing this variance, the minimum number of examinees needed to develop the weights, subject to a specified level of precision, can be determined.

There are a number of methods that can be used for deriving the weights. The method that will be discussed here is that used by Echternacht (1973), which is actually the method used by Reilly and Jackson (1972) with no iterations. Briefly, the method consists of assigning the average criterion score of those selecting a given option. The criterion variable is standardized, so that its mean is zero and variance is one. The criterion that is usually used is the score on the remaining items that make up the test although this is certainly not a necessary criterion.

Consider a population of N people who will take a given test at one point in time. Assume further that a simple random sample of n people from the population take the test for the purpose of determining scoring weights. Although this is not exactly true in an operational setting, it does provide a useful approximation to reality. Consider one item for that test. The scoring weight assigned to the ith option of this item is

$$\bar{y}_{i.} = \sum_{j=1}^{n_i} y_{ij}/n_i$$

where n_i represents the number of people responding with the ith option and y_{ij} represents the criterion score for the jth person choosing the ith option. In weighting options, the omit category is considered another option and a weight is also derived. Since the criterion variable is assumed to be standardized,

$$\sum_{i=1}^{c} n_i \bar{y}_{i.} / n = \bar{y}_{..} = 0$$

where $n = \sum_{i=1}^{c} n_i$, the number of people responding to the item with one of the c possible options. Using the standard result for the variance of a mean obtained by simple random sampling from a finite population, the variance of the ith option weight thus becomes

$$(1/n_i - 1/N_i)\, S_i^2$$

where

$$S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_{i.})^2 .$$

N_i indicates the number of examinees in the population responding with option i. The problem becomes one of estimating S_i^2. This is done by using the unbiased estimate

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 .$$

Such estimates of S_i^2 would presumably be obtained through pretesting of the item.

Suppose the whole population of N examinees is used for the purpose of determining scoring weights, and the method previously described is used.

Now,

$$S^2 = \frac{1}{N-1} \sum_{i}^{c} \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_{..})^2 = 1 \quad , \text{ and } \quad \bar{Y}.. = 0$$

where c indicates the number of response options. From the standard algebraic identify for the analysis of variance, with

$$N = \sum_{i=1}^{c} N_i \quad ,$$

$$(N-1)S^2 = (N-1) = \sum_{i=1}^{c} \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}..)^2$$

$$= \sum_{i=1}^{c} N_i (\bar{Y}_{i.} - \bar{Y}..)^2 + \sum_{i=1}^{c} \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_{i.})^2$$

$$= \sum_{i=1}^{c} N_i \bar{Y}_{i.}^{\,2} + \sum_{i=1}^{c} (N_i - 1) S_i^2 \quad . \qquad (1)$$

If the 1/N is negligible (1) may be written as

$$1 = \sum_{i=1}^{c} W_i \bar{Y}_{i.}^{\,2} + \sum_{i=1}^{c} W_i S_i^2 \quad , \qquad (2)$$

where

$$W_i = N_i/N$$

so that

$$\sum_{i=1}^{c} W_i (1-\bar{Y}_{i.}^2) = \sum_{i=1}^{c} W_i S_i^2$$

and

$$\sum_{i=1}^{c} W_i (1-\bar{Y}_{i.}^2) = \sum_{i=1}^{c} W_i S_i^2 = \bar{S}^2 , \qquad (3)$$

which indicates that the S_i^2 are not independent for all c categories.

In obtaining empirically derived scoring weights, it is, of course, desirable to have the variance of the resulting weights be of a minimum. If a large enough pool of examinees are tested in the initial test administration so that the n_i are all large for each item, the variances will likely be small. This is not always the case, though, and it does not tell the item writer anything about how he should write the items to help insure that a small variance results. The item writer can have some influence over both the n_i and the S_i^2. By increasing the n_i and decreasing the S_i^2 the ith option weight's variance will decrease. But, the n_i and S_i^2 are not independent for a given item. Therefore, it seems reasonable to consider minimizing $\bar{S}^2$ and the implications this minimization has for item writers. One can see that $\bar{S}^2$ can be minimized by making the between options sum of squares, $\sum_{i=1}^{c} N_i \bar{Y}_{i.}^2$, a maximum.

Although it is recognized that the following discussion is somewhat esoteric for the item writer and the conditions presented very unrealistic, the discussion following is an attempt to demonstrate some of the basic principles that should be used in minimizing $\bar{S}^2$.

In maximizing $Q = \sum_{i=1}^{c} N_i \bar{Y}_{i.}^2$, a few things need to be noted.

In the case where $c=2$, it can be easily shown that Q attains a minimum when $\sum_{j=1}^{N_i} Y_{ij} = 0$, or when each category mean equals the overall mean. Also, if $\sum_{j=1}^{N_i} Y_{ij}$ can be considered given and Q a function of only the N_i's , Q is minimized when $N_1 = N/2$. Since we are considering a finite population, a maximum value of Q is obtained when all positive Y_{ij} are found in one category and all negative Y_{ij} in the other. The zero values of Y_{ij} are placed in the category with the largest N_i .

In cases where $c>2$, it can be shown that Q is minimized when $\sum_{j=1}^{N_i} Y_{ij} = 0$ for each i , or if the sums, $\sum_{j=1}^{N_i} Y_{ij}$, are considered fixed, when the N_i are proportional to $\left| \sum_{j=1}^{N_i} Y_{ij} \right|$. Maximum

values can be obtained only when the criterion values can be partitioned into nonoverlapping regions, with each region corresponding to a group of people responding with a particular distractor. In topological terms these regions are termed "connected" regions, and their union consists of the entire criterion variable space. This is also the case where each distractor can be used to place the individual responding with that distractor in a categorization of the criterion.

In practice though, it is impossible for an item writer to write items with the property previously noted. The item writer can structure distractors in such a way that examinees of differing ability levels respond to different distractors. Such a practice would tend to approximate the condition mentioned previously, assuming that ability and the criterion are related, and allow Q to be maximized as much as is practical. The procedure of "facet design" as set forth by Guttman (see Elizur, 1970) is one method that might be used to so structure the distractors. In examining the results of item pretesting, the quantity Q should also be taken into consideration in making the decision of whether or not to include a given item as part of a test that will be scored using empirically derived option weights.

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 comparison on various item option weighting schemes. Research Bulletin 73-6. Princeton, N. J.: Educational Testing Service, 1973.

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

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.

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