DOCUMENT RESUME

ED 073 132 TM 002 376

AUTHOR	Reilly, Richard R.
TITLE	Empirical Option Weighting with a Correction for Guessing.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RB-72-59
PUB DATE	Dec 72
NOTE	21p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Correlation; Data Analysis; Guessing (Tests); Mathematical Applications; Research; *Scoring Formulas; *Test Construction; *Test Reliability; *Test Validity
IDENTIFIERS	Graduate Records Examination

ABSTRACT

Because previous reports have suggested that the lowered validity of tests scored with empirical option weights might be explained by a capitalization of the keying procedures on omitting tendencies, a procedure was devised to key options empirically with a "correction-for-guessing" constraint. Use of the new procedure with Graduate Record Examinations data resulted in smaller increases in reliability than those observed when unconstrained procedures were used, but validities for quantitative subforms were not appreciably lowered. Validities for verbal subforms were lowered slightly, however. (Author)

RESEARCH BULLETIN

RB-72-59

EMPIRICAL OPTION WEIGHTING WITH A CORRECTION FOR GUESSING

Richard R. Reilly

Educational Testing Service
Princeton, New Jersey
December 1972

Empirical Option Weighting with a Correction for Guessing

Richard R. Reilly

Educational Testing Service

Abstract

Because previous reports have suggested that the lowered validity of tests scored with empirical option weights might be explained by a capitalization of the keying procedures on omitting tendencies, a procedure was devised to key options empirically with a "correction-for-guessing" constraint. Use of the new procedure with Graduate Record Examinations (GRE) data resulted in smaller increases in reliability than those observed when unconstrained procedures were used, but validities for quantitative subforms were not appreciably lowered. Validities for verbal subforms were lowered slightly, however.

Empirical Option Weighting with a Correction for Guessing[1]

Richard R. Reilly

Educational Testing Service

Two recent reports (Hendrickson, 1971; Reilly & Jackson, 1972) have suggested that weighting options empirically results in substantial increases in reliability and test homogeneity, but at the expense of lowered test validity. These findings are at variance with those reported in an earlier study by Davis and Fifer (1959) who found similar increases in reliability and slight increases in validity when options were weighted empirically. All three studies employed modifications of a weighting technique originally known as The Method of Reciprocal Averages (Mosier, 1946) which, in effect, maximizes the product-moment correlation between item scores and criterion scores by assigning to each item-option values proportional to the mean criterion score for all individuals choosing that option.

A key difference between the Davis and Fifer study and the first two mentioned was that tests in the first two were administered with formula score instructions while Davis and Fifer instructed examinees to attempt every item. Thus, Hendrickson and Reilly and Jackson had an additional "option," that of omit. Hendrickson, reporting on the weights generally assigned to the omit category comments, "...An interesting finding of this study was that the weight of 'omit' was almost always lower than any of the other distracters in an item..." (Hendrickson, 1971). Reilly and Jackson (1972) take this a step further and suggest that, "...the empirical keying procedures described capitalize on the tendency to omit and...while this tendency is reliable, it is not valid."

Because of these suggestions, it was decided to devise and test a procedure which weighted options subject to the constraint that the weight for omit equal the mean weight for the options. The rationale is similar to that used in the usual formula scoring method in that it assumes that an individual omitting an item should receive the expected weight under conditions of random response to that item.

In order to determine the optimum weights for a single item, subject to the "correction-for-guessing"constraint,the following objective function was set up:

$$F = \sum_{ji}\sum(y_{ij} - w_j)^2 - 2\lambda[(k - 1)w_p - \sum_j \delta_j w_j] \quad ,$$

where

y_{ij} denotes the criterion score of the ith individual making the jth response;

w_j is the weight for the jth response, $j = 1,..,p,.... k$; and

w_p is the weight for the omit category.

δ_j = one for $j \neq p$, and zero otherwise; and

λ is the LaGrange multiplier.

Taking partial derivatives and solving for the weights which minimize the function we find that the solution, which requires a small $(k - 1 \times k - 1)$ matrix inversion, has the following properties (see appendix): (1) The mean item score over all individuals is equal to the mean criterion score; (2) the weights arrived at are proportional to the weights which will maximize the correlation between the item and the criterion subject to the constraint of a fixed item variance (and, of course, the constraint that the omit weight equals

the mean of the option weights); (3) unlike the unconstrained option weights, the weights arrived at will not, in general, yield the maximum possible product-moment correlation; (4) for unconstrained weights it has been pointed out (Stanley & Wang, 1970) that a slope of 1.0 and a zero intercept will describe the regression of the criterion scores on the item scores. The appropriate slope for the regression of criterion scores on item scores yielded by the new method will not, in general, be 1.0, nor will the appropriate intercept, in general, be zero.

Procedure

Two parallel forms each, of the verbal (denoted as V_1 and V_2) and quantitative (Q_1 and Q_2) sections of the GRE, were devised by assigning one-half of the items on each section to each of the two special parallel forms. Forms V_1 and V_2 consisted of 50 items each, while forms Q_1 and Q_2 consisted of 27 items each. It should be noted that the two forms in each set, since they were constructed from operational tests, were not administered under separate time limits. Because of practical limitations the more desirable procedure of administering the two parallel forms under separately timed conditions was not possible.

Data were the same as these used in the Reilly and Jackson (1972) study. A spaced sample (i.e., a sample consisting of every nth answer sheet) of 5,000 answer sheets (sample A) from the December 1970 administration of the GRE was employed for study purposes. A second sample (sample B) consisting of the answer sheets of 4,916 individuals from the same administration was taken for validation purposes. Sample A was divided into two randomized block groups of 2,500 (samples A_1 and A_2) by blocking on total GRE score. The

5,000 answer sheets were ordered in terms of the verbal score plus the quantitative score and then randomly assigned to the two subsamples. This increased the likelihood that the two split samples would be comparable in terms of total score distributions. Each subtest was keyed against the scores on its parallel form in sample A_1. The tests in sample A_2 were then scored using these derived weights and intercorrelations, and alpha coefficients were computed. Thus, all results reported are those obtained with cross-validated weights.

The next step involved scoring the sample B answer sheets and computing the single order and multiple correlations between the empirically keyed tests and undergraduate GPA. Sample B was drawn from a total of 40 different colleges. Within-school samples ranged from a low of 16 to a high of 399. A modification of one of Tucker's (1963) central prediction methods was used to pool data across colleges.[2]

Results and Discussion

The results of the keying on parallel forms reliability and internal consistency are presented in Tables 1 and 2. The proportional increases in

---

Insert Tables 1 and 2 about here

---

effective test lengths are comparable to those reported by Hendrickson (1971) but less than those observed by Reilly and Jackson (1972). The smaller increments observed for the quantitative tests are consistent with previous findings, and may, as Hendrickson (1971) suggests, be related to the common observation that differences in the quality of the distracters are less apparent for general mathematical items than for verbal items.

Reilly and Jackson (1972) observed increases in the correlations between verbal and quantitative tests when empirical weights were used and attributed these increases to the capitalization of the keying procedure on an omitting factor common to both tests. Thus, the results shown in Table 3 are of interest since they indicate that when constrained weights

---

Insert Table 3 about here

---

are used the large increases in verbal-quantitative correlations do not occur. When increases in reliability are taken into account the increases are actually slightly less than expected in two of the four cases shown and slightly greater than expected in the remaining two cases.

In Table 4, the correlations are shown between pairs of parallel subtests,

---

Insert Table 4 about here

---

one scored with empirical weights and the other with formula weights. These [illegible] correlations are, in general, slightly higher than the parallel forms reliability, in contrast to the uniformly lower values obtained when unconstrained weights were used (Reilly & Jackson, 1972).

The validity results are presented in Table 5. While the zero-order

---

Insert Table 5 about here

---

validities for the quantitative forms are almost unchanged, the multiple correlations are slightly lower overall owing primarily to the decreases in the correlations between GPA and the empirically keyed verbal subtests. It is difficult to explain why, even with the modified keying procedure, the

verbal test validities were lowered. Apparently, the empirically keyed verbal tests are measuring some additional factors which, though reliable, may not be valid.

Conclusions

While the results reported here certainly do not indicate that steps should be taken to implement empirical option weighting, the findings are not entirely discouraging either. It has been shown that a test can be made more reliable and more homogeneous through option weighting and, at least for the quantitative forms, without any appreciable lowering of validity.

Further research should be done on several key issues which have emerged in this study. First, the issue of omitting behavior should be looked at more closely. Green (1972) has presented data for the SAT which indicate that "orit" scores are even more reliable than rights-only or formula scores. It may be that an omitting score can be used as a suppressor variable along with the formula score to increase the correlation with the criterion.

Another interesting and potentially useful study would be one which examined the effects of keying options directly on the GPA criterion. Examination of the weights for options may reveal consistent patterns which could be helpful in guiding item writers.

References

Briggs, B. Boldt's special case of central prediction, weighted least squares procedure. Statistical Systems Report. SS12. Princeton, N.J.: Educational Testing Service, 1970.

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.

Green, B. F. The sensitivity of Guttman weights. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April 7, 1972.

Hendrickson, G. F. The effect of differential option weighting on multiple-choice objective tests. Report No. 93. Center for the study of social organization of schools, The Johns Hopkins University, Baltimore, Maryland, 1971.

Mosier, C. I. Machine methods in scaling by reciprocal averages. Proceedings, Research Forum. New York: International Business Machines Corporation, 1946. Pp. 35-39.

Reilly, R. R., & Jackson, R. Effects of empirical option weighting on validity and reliability of shortened forms of the GRE aptitude tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago. Illinois, April 7, 1972.

Stanley, J. C., & Wang, M. D. Weighting test items and test-item options, an overview of the analytical and empirical literature. Educational and Psychological Measurement, 1970, 30, 21-25.

Tucker, L. R. Formal models for a central prediction system. Psychometric Monograph No. 10. Richmond, Va.: William Byrd Press, 1963.

Footnotes

[1]The research reported herein was supported by the Graduate Record Examinations Board.

[2]The method used is a least-squares procedure worked out by Robert F. Boldt and is more fully described in a report by Briggs (1970).

Table 1

Cross-Validated Parallel Forms Reliabilities for

Empirically Keyed and Formula Scored Subtests

	Formula	Empirically Keyed	K[a]
Verbal	.8909	.9242	1.49
Quantitative	.8742	.8892	1.16

[a]K gives the estimated proportional increase in test length which would be necessary to yield the increased R 's shown. Rearranging the Spearman-Brown prophecy formula,

$$K = \frac{R_w(1 - R_F)}{R_F(1 - R_w)} ,$$

where R_F is the R obtained with formula score weights and R_w is the cross-validated R obtained with empirical weights.

Cross-Validated Internal Consistency Coefficients for Formula Scored and Empirically Keyed Tests

	Formula	Empirically Keyed	K^a
V_1	.8745	.9069	1.40
V_2	.8755	.9084	1.41
Q_1	.8515	.8817	1.30
Q_2	.8725	.8852	1.13

[a]K gives the estimated proportional increase in test length which would be necessary to yield the increased α's shown. Rearranging the Spearman-Brown prophecy formula,

$$K = \frac{\alpha_w(1 - \alpha_F)}{\alpha_F(1 - \alpha_w)} ,$$

where α_F is the α obtained with formula score weights and α_w is the cross-validated α obtained with empirical weights.

Table 3

Intercorrelations between Verbal and Quantitative Forms

for Formula Scored and Empirically Keyed Tests

	Formula	Empirically Keyed	Expected[a]
V_1Q_1	.4154	.4577	.4269
V_2Q_1	.4190	.4428	.4550
V_1Q_2	.4079	.4304	.4191
V_2Q_2	.4061	.4138	.4173

[a]The expected values represent the expected correlation which should have resulted from the increased reliability of the empirical key scores. These values were obtained by multiplying the true formula score correlations between V and Q by the geometric mean of the empirical key score reliabilities. Parallel forms reliabilities were used in all cases.

Table 4

Intercorrelations between Empirically Keyed

and Formula Scored Parallel Forms

	Parallel Forms Reliability	Empirically Keyed vs. Formula Scored Parallel Form[a]	
		I	II
Verbal	.8909	.8953	.8914
Quantitative	.8742	.8726	.8848

[a]Column 1 shows the correlation between form V_1 (Q_1) empirically keyed and form V_2 (Q_2) formula scored. Column 2 shows the correlation between V_2 (Q_2) empirically keyed and V_1 (Q_1) formula scored.

Table 5

Validity Coefficients[a] for Selected Pairs of Empirically Weighted and Formula Scored Subtests

	V_1	Q_1	V_1+V_2	V_2	Q_2	V_2+Q_2
Formula Scores	.3167	.1909	.3184	.2939	.2054	.3013
Unconstrained Weights[b]	.2703	.1664	.2666	.2532	.1504	.2550
Constrained Weights	.2998	.1894	.2997	.2828	.2055	.2919

[a]Single order coefficients were estimated as follows:

$$r = \sqrt{\frac{\Sigma n_i r^2_i}{\Sigma n_i}} \quad ;$$

multiple correlation coefficients were obtained using a pooling procedure described by Briggs (1970).

[b]The unconstrained weights were those obtained by keying against parallel forms (Reilly & Jackson, 1972).

APPENDIX

First we solve for weights which minimize the least squares criterion subject to the constraint that the weight for omit equals the mean of the option weights. Let

$$F = \sum_j \sum_i (y_{ij} - w_j)^2 - 2\lambda[(k - 1)w_p - \sum_j \delta_j w_j]$$

be the function to be minimized subject to the restriction that the weight for one of the categories, w_p, equals the mean of the remaining $(k - 1)$ weights, where

w_j is the weight for the jth category;

y_{ij} is the criterion score for the ith person in the jth category;

δ_j is one if $j \neq p$, zero if $j = p$;

λ is the LaGrange multiplier;

k is the total number of categories;

i is 1, n_k; and

j is 1, k.

Take the partial derivative with respect to w_j,

$$\frac{\partial F}{\partial w_j} = 2\sum_i y_{ij} - 2n_j w_j + 2\lambda \quad .$$

Take the partial derivative with respect to w_p,

$$\frac{\partial F}{\partial w_p} = 2\sum_i y_{ip} - 2n_p w_p - 2(k - 1)\lambda \quad .$$

Setting both equations equal to zero and multiplying by $(-\frac{1}{2})$, we have

$$\sum_i y_{ij} - n_j w_j - \lambda = 0 \qquad (1)$$

and

$$\sum_i y_{ip} - n_p w_p + (k - 1)\lambda = 0 \quad . \qquad (2)$$

Taking (1) and summing over j ,

$$\sum_j \delta_j \left(\sum_i y_{ij} - n_j w_j - \lambda\right) = 0 \quad .$$

Rearranging,

$$\sum_j \delta_j n_j w_j = \sum_j \delta_j \sum_i y_{ij} - (k - 1)\lambda \quad . \qquad (3)$$

Rearranging equation (2) similarly and adding to equation (3) we have

$$\sum_j n_j w_j = \sum\sum_{ji} y_{ij} \qquad (4)$$

or,

$$\bar{w} = \bar{y} \quad ,$$

a desirable result since the mean of the scores generated with the new weights will always equal the criterion mean. Rearranging (2) we obtain

$$\lambda = \frac{1}{k - 1}\left(n_p w_p - \sum_i y_{ip}\right) \quad .$$

Substituting this last result in equation (1) we have

$$n_j w_j = \sum_i y_{ij} - \frac{1}{k - 1}\left(n_p w_p - \sum_i y_{ip}\right) \quad . \qquad (5)$$

By the constraint, however,

$$w_p = \frac{\sum_j \delta_j w_j}{k - 1} \quad ,$$

so that

$$n_j w_j = \Sigma y_{ij} - \frac{1}{k-1}\left(n_p \frac{\Sigma \delta_j w_j}{k-1} - \Sigma y_{ip}\right) \qquad (6)$$

and

$$n_j w_j + \frac{n_p}{(k-1)^2} \Sigma \delta_j w_j = \Sigma y_{ij} + \frac{1}{k-1} \Sigma y_{ip} \quad . \qquad (7)$$

Thus, we have $k - 1$ such equations and $k - 1$ unknown weights (the weight w_p is fixed at $\frac{\Sigma \delta_j w_j}{k-1}$) .

Let

$$\frac{n_p}{(k-1)^2} = q \quad ,$$

and construct the $(k - 1) \times (k - 1)$ matrix X with diagonal elements $(n_j + q)$, $j = 1,\ldots p - 1, p + 1,\ldots k$, and off-diagonal elements q . Let W be a column vector of $k - 1$ weights, w_j , $j = 1,\ldots p - 1, p + 1,\ldots k$, and let Y be a $(k - 1) \times 1$ vector with elements

$$\Sigma y_{ij} + \frac{1}{k-1} \Sigma y_{ip} \quad , \quad j = 1,\ldots p - 1, p + 1,\ldots k \quad .$$

The equations can be represented in matrix form as follows:

$$XW = Y$$

and the solution

$$W = X^{-1}Y$$

is readily obtained.

Next we prove that the weights derived in the foregoing proof are proportional to the weights which will maximize the correlation between an item and a criterion subject to both the formula score constraint and the constraint of some fixed variance, B. Let the objective function to be maximized be

$$H = \sum_j n_j w_j \bar{y}_j - \frac{1}{2}\lambda_1\left(\sum_j n_j w_j^2 - NB\right) + \lambda_2\left(\sum_j n_j w_j\right) - \lambda_3\left(\sum_j \delta_j w_j - \{k-1\}w_p\right) ,$$

where

$$\bar{y}_j = \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}$$

and where the LaGrange multipliers represent the following constraint conditions:

- (λ_1) the variance of the weights when taken over all individuals in the sample is equal to some constant B;
- (λ_2) the mean item score will be zero; and
- (λ_3) the pth category weight is the average of the other $k-1$ weights.

Taking the partial derivative with respect to any w_j $(j \neq p)$ and setting the result equal to zero we have

$$n_j \bar{y}_j - \lambda_1 n_j w_j + \lambda_2 n_j - \lambda_3 = 0 \quad . \tag{8}$$

Taking the partial derivative with respect to w_p and setting the result equal to zero we obtain

$$n_p \bar{y}_p - \lambda_1 n_p w_p + \lambda_2 n_p + (k-1)\lambda_3 = 0 \quad . \tag{9}$$

Summing equations over j we obtain

$$N\lambda_2 = \lambda_1 \sum_j n_j w_j - \sum_j n_j \bar{y}_j \quad .$$

But, by constraint,

$$\sum_j n_j w_j = 0 \quad ,$$

and by definition,

$$\sum_j n_j \bar{y}_j = 0 \quad .$$

Thus,

$$\lambda_2 = 0 \quad .$$

Solving for λ_3 in equation (9) and substituting the result in equation (8) we have

$$\lambda_1 \left(n_j w_j + \frac{n_p}{k - 1} w_p \right) = n_j \bar{y}_j + \frac{n_p}{k - 1} y_p \quad . \tag{10}$$

Since by constraint, however,

$$w_p = \frac{\sum_j \delta_j w_j}{k - 1} \quad ,$$

$$\lambda_1 \left(n_j w_j + \frac{n_p}{(k - 1)^2} \sum_j \delta_j w_j \right) = n_j \bar{y}_j + \frac{n_p}{k - 1} \bar{y}_p \quad .$$

Let the X matrix and the W vector be defined as in the previous proof, and let Y be a vector with elements

$$n_j \bar{y}_j + \frac{n_p}{k - 1} \bar{y}_p \quad , \quad j \neq p \quad .$$

Thus,

$$\lambda_1 XW = Y$$

and

$$W = X^{-1}Y\lambda_1^{-1} \quad .$$

We see that the solution is identical to that obtained previously except for a proportionality constant. To find the proportionality constant λ_1, let G be the vector of $k - 1$ weights, with elements g_j where

$$G = X^{-1}Y \quad .$$

By the constraint

$$\sum_j n_j w_j^2 = NB \quad .$$

but since

$$W = \lambda_1^{-1} G \quad ,$$

$$\lambda_1^{-2} \sum_j n_j g_j^2 = NB$$

and

$$\lambda_1 = \sqrt{\frac{\sum n_j g_j^2}{NB}} \quad .$$