DOCUMENT RESUME

ED 069 738 TM 002 193

AUTHOR	Reilly, Richard R.; Jackson, Rex
TITLE	Effects of Empirical Option Weighting on Reliability and Validity of the GRE.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RB-72-38
PUB DATE	Aug 72
NOTE	29p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Correlation; Factor Analysis; *Graduate Study; *Scoring; *Test Reliability; *Test Validity; Verbal Tests; *Weighted Scores
IDENTIFIERS	*Graduate Record Examinations

ABSTRACT

Item options of shortened forms of the Graduate Record Examination Verbal and Quantitative tests were empirically weighted by two variants of a method originally attributed to Guttman. The first method assigned to each option of an item the mean standard score on the remaining items of all subjects choosing that option. The second procedure assigned the mean score on a parallel form of all persons choosing the option. When compared with formula scores, it was found that scores generated with the empirical weights were more reliable but less valid when correlated with undergraduate grade-point average (GPA). Test homogeneity was increased through empirical option weighting, and factor analysis revealed large increases in variance accounted for by the first factor. Examination of the actual weights assigned to each option revealed that the weight for omit in most cases differed considerably from the weight which would be assigned under the usual formula score assumptions. It was suggested that the weighting procedures used tended to capitalize on omitting behavior which, although a highly reliable tendency, may actually be negatively related to the GPA criterion used. (Author)

RB-72-38

RESEARCH BULLETIN

EFFECTS OF EMPIRICAL OPTION WEIGHTING ON RELIABILITY AND VALIDITY OF THE GRE

Richard R. Reilly

and

Rex Jackson

Educational Testing Service

Princeton, New Jersey

August 1972

EFFECTS OF EMPIRICAL OPTION WEIGHTING ON RELIABILITY AND VALIDITY OF THE GRE

Richard R. Reilly and Rex Jackson

Educational Testing Service

Abstract

Item options of shortened forms of the GRE Verbal and Quantitative tests were empirically weighted by two variants of a method originally attributed to Guttman. The first method assigned to each option of an item the mean standard score on the remaining items of all subjects choosing that option. The second procedure assigned the mean score on a parallel form of all persons choosing the option.

When compared with formula scores, it was found that scores generated with the empirical weights were more reliable but less valid when correlated with undergraduate grade-point average (GPA). Test homogeneity was increased through empirical option weighting, and a factor analysis revealed large increases in variance accounted for by the first factor.

Examination of the actual weights assigned to each option revealed that the weight for omit in most cases differed considerably from the weight which would be assigned under the usual formula score assumptions. It was suggested that the weighting procedures used tended to capitalize on omitting behavior which, although a highly reliable tendency, may actually be negatively related to the GPA criterion used.

EFFECTS OF EMPIRICAL OPTION WEIGHTING ON RELIABILITY AND VALIDITY OF THE GRE[1]

Richard R. Reilly and Rex Jackson[2]

Educational Testing Service

Concomitant with the gains in scoring simplicity and objectivity realized through the use of the multiple-choice format are losses resulting from, among other things, the lack of scoring flexibility and the introduction of a significant proportion of chance variance through guessing. A good deal of psychometric literature has centered on attempts to counter the limitations of the multiple-choice format while maintaining its more desirable aspects. Much of this literature is summarized in a recent article by Stanley and Wang (1970).

The present study considered one type of approach to improving the measurement properties of the multiple-choice item, that of assigning choice weights empirically. Stated simply, the approach involves choosing some desirable criterion, administering a set of items to a sample for which this criterion information is available, and assigning weights to responses based on their relationship to the criterion.

Guttman (1941) described a solution to the problem of weighting responses to a set of items so as to maximize internal consistency which requires solving for the first principal component of an $mk \times mk$ matrix, where m is the number of items and k is the (fixed in this case) number of possible responses to each item. Lord (1958) later showed that Guttman's weights were the same as those necessary to maximize coefficient alpha for a set of items. Because of practical considerations a computationally simpler method sometimes referred to as the method of reciprocal averages has been used to give an approximation to the results which would be obtained by solving for the first principal component. Although the relationship of

the method of reciprocal averages to Guttman's method has only recently been made explicit (Baker & Hoyt, 1972), it has considerable intuitive appeal since it maximizes the single order correlation between each item and the total score criterion by assigning values proportional to the mean total scores of individuals choosing each option. Independent proofs of this result have been offered by Stanley and Wang (1970) and, in a somewhat different context, by Beaton (1968). It should be pointed out that for a set of multiple-choice items the procedure described does not yield a completely optimum solution in a least-squares sense, since it considers each item apart from other items and therefore does not take into account item intercorrelations. A completely optimum solution, however, would entail assigning a unique set of weights for each possible response pattern. Assuming every person makes one response to each item on a four-choice, 10-item test, the number of parameters which would have to be fit for an optimum least squares solution would exceed one million. The practical impossibility of even obtaining a solution makes it clear why the reciprocal averages method has been used in most of the relevant previous investigations.

Practical considerations also arise in the choice of a criterion against which to key the options. If enough criterion data are available for all individuals in the keying sample and a cross-validational sample and, further, if the criterion scores were obtained under approximately the same conditions for all individuals, then, clearly, options should be keyed against this criterion. Since this situation is rarely the case, most previous investigators have contented themselves with keying on some intermediate criterion such as total test score or the score on another test. Two previous investigations which have a heavy bearing on the present

study best exemplify this approach. The first one (Davis & Fifer, 1959) keyed the options of two parallel arithmetic reasoning tests, with the criterion for the options of one form being the scores on the other form. For a 45-item test, the investigators reported a cross-validated increase in parallel forms reliability from .68 to .76 without lowering validity.

A second study, by Hendrickson (1971), keyed the options of verbal and quantitative sections of the Scholastic Aptitude Test. Instead of attempting to increase parallel forms reliability, Hendrickson sought to raise the internal consistency of each subtest by first keying the options of each item on the total corrected-for-guessing score for the remaining items. After the initial keying several iterations were performed until coefficient alpha, which served as the index of internal consistency, appeared to stabilize. Hendrickson employed a double cross-validation design and performed all analyses separately for males and females. In all cases, substantial cross-validated increases in coefficient alpha were achieved. Lesser increases were noted in the correlations between the two verbal subtests and, interestingly, decreases were observed in the correlations between the two mathematics subtests. All of the correlations between verbal and mathematics subtests which Hendrickson interpreted as "quasi-validity" coefficients showed decreases. Unfortunately, no other criterion data, such as college grades, were available for assessing any changes in validity which might have occurred.

The present study, which employed specially devised parallel forms of the verbal and quantitative sections of the Graduate Record Examinations (GRE), can be viewed as an extension of the work of Davis and Fifer and of Hendrickson. It was hoped that the study would provide evidence bearing on the general question of how the psychometric properties of verbal and

quantitative academic aptitude tests are affected when options are keyed empirically. More specifically, the following questions were asked:

(1) What happens to the internal consistency of a test keyed to increase parallel forms reliability?

(2) What happens to the parallel forms reliability of a test keyed to increase internal consistency?

(3) Does either type of keying result in an increase in validity over conventional scoring methods either for individual subtests or when verbal and quantitative tests are combined to obtain a multiple correlation?

(4) If the answer to the last question is yes, which of the two methods of keying seems to offer the most promise?

(5) How does the factor structure of subtests keyed for internal consistency and parallel forms reliability compare with the factor structure when conventional scoring methods are used?

Method

Test Forms

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. The specific items assigned to each form are listed in Appendix I. 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. The GRE was designed to be primarily a measure of power rather than speed, however, so that effects due to correlated speed components should have been negligible.

Sample

A spaced sample of 5,000 answer sheets (sample A) from the December 1970 administration of the GRE was taken 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 2500 (samples A_1 and A_2) using total GRE score (V + Q) as the blocking variable. The purpose of blocking was to increase the probability that the total score means and standard deviations for these two groups would be approximately equal. Double cross-validation was carried out for each weighting method. Thus for each type of keying two independent sets of keys were derived. for each subtest (one in sample A_1, and one in sample A_2) and independently cross-validated on the A sample not used to key. Sample B was used for the concurrent validity analysis.

Keying Procedures

(1) Keying for internal consistency. For each subform a procedure designed to increase internal consistency similar to that described by Hendrickson (1971) was employed. Full computational details are provided in Appendix II but the procedure may be briefly described as follows:

(a) First, score the subform using the conventional scoring formula (i.e., rights - 1/4 wrongs).

(b) For each item key each option by assigning the mean standard score on the remaining items for all persons choosing that option.

(c) After all items have been keyed in this manner, compute coefficient alpha.

This procedure can be used iteratively until coefficient alpha appears to stabilize. In the present study, however, increments after the first keying were observed to be negligible and therefore the weights obtained from the initial keying were those used for scoring purposes.

(2) Keying for parallel forms reliability (PF). This procedure is similar to the one employed by Davis and Fifer (1959) and assigned to each option of an item the mean standard score on the corresponding parallel subform of all individuals choosing that option.

Analyses

Each subform in sample A_1 was scored three different ways, once using the conventional correction for guessing formula, once using the weights derived from the internal consistency keying in sample A_2, and once using the weights derived from the parallel forms keying in sample A_2. The same procedure was followed for sample A_2 except that weights derived in sample A_1 were employed for the latter two scorings. For each of the three scoring methods, alpha coefficients were computed for each subform and intercorrelations among subforms were also computed. Thus, two cross-validated alpha coefficients and two parallel forms reliabilities were obtained. In order to investigate changes in the factor structure which might have occurred as a result of empirical keying (against parallel forms), a factor analysis of the items within each test was performed in sample A_2.

Cumulative undergraduate GPAs were obtained for all individuals in sample B, and each subform was scored using the different sets of weights derived from sample A_1. Practical considerations dictated the use of undergraduate rather than graduate GPA. Graduate GPA would not have been available for so large a sample, and in addition tends to be highly restricted in range. On the other hand, both GRE scores and undergraduate grades are generally accepted measures of the same construct, academic ability, so that the validity data reported may be regarded as both construct and concurrent (since the large proportion of GRE candidates take the exam near the end of their undergraduate academic careers) validity. All single order correlation coefficients between each subform and cumulative GPA were computed within college, and multiple correlations between one verbal subform, one quantitative subform, and cumulative GPA were computed within undergraduate institutions. Finally, data across schools were pooled using a central prediction method due originally to Tucker (1963), and overall estimates of the validity of variables singly and in combination were obtained.

Results and Discussion

Table 1 shows the cross-validated internal-consistency coefficients

---

Insert Table 1 about here

---

for each type of weighting system. The k-values shown reflect the proportional increase in test length estimated by the Spearman-Brown formula. The results are quite impressive given the crucial assumption that the same latent trait, or set of latent traits, is being measured by the test. We see in Table 2 that the parallel forms reliability

---

Insert Table 2 about here

---

estimates follow a highly similar pattern with estimates of effective changes in test length ranging from slightly more than one and one-half the original length for one quantitative subform to more than twice the original length for the verbal forms.

These data strongly suggest the answers to our first two questions which were concerned with what happens to internal consistency and parallel forms reliability when options are empirically keyed. It is clear that these measures are increased rather substantially by empirical weighting. It is also worth noting that the two types of keying carried out were, for all practical purposes, identical in their effects and, in fact, cross-validated scores yielded by the two methods were correlated almost perfectly (all correlations were .999 or greater).

The factor analysis of the items for each subform scored with formula weights and empirical weights revealed sharp increases in variance accounted for by the first few empirical weight factors particularly the first (see Table 3). This finding parallels that reported by Hendrickson

---

Insert Table 3 about here

---

and Green (1972) for the SAT. Interpretation of the factors is beyond the scope of the present study and should prove quite difficult in any case as Hendrickson and Green (1972) found. It was observed, however, that individual item loadings (after a varimax rotation), as well as individual item intercorrelations, underwent considerable changes in

many cases after empirical weighting suggesting that changes in the underlying structure of the tests may have occurred.

The real test of this procedure came in the next set of analyses performed. For this purpose the answer sheets of 4,916 college students, who had taken the GRE at the same administration from which the keying samples were selected, were scored with formula-score weights and with empirically derived weights. None of this group was included in the keying sample, and an effort was made to provide a representative range of undergraduate institution attended. A total of 40 institutions provided cumulative undergraduate GPA data for these individuals. Within-school sample sizes ranged from 16 to 399, with a mean within-school sample size of 130. Taking pairs of verbal and quantitative subforms, both single-order and multiple correlations were computed between conventionally scored tests and GPA and between empirically weighted scores and GPA. Both single order and multiple correlations were slightly but consistently higher for the formula scores. The weighted scores produced on the average (unweighted) a multiple correlation which was .05 less than the multiple correlation obtained with formula scores.

A modification of Tucker's (1963) Model III central prediction method was employed to pool data across colleges. Briefly, this method computes a common set of regression weights as well as multiplicative and additive constants for each college which minimize the squared errors of prediction (e.g., see Briggs, 1970). The pooled validity coefficients for each variable and selected pairs of variables are presented in Table 4 along with the arithmetic average and median within-school multiple correlations.

Insert Table 4 about here

Again, the results quite clearly indicate that option weighting lowered test validity. The conclusion that empirical option-weighting did not lead to any increase in validity is clear enough but the reasons for this are not. One would expect the more reliable scores to predict the GPA criterion slightly more accurately.

Several explanations were considered. One possibility is that the weighted score reliabilities which held up so well in the carefully constructed A_1 and A_2 samples broke down in the validation sample (sample B). This was not the case, however. The reliabilities for the weighted scores were consistently and substantially higher than formula scores in the validation sample. A second possibility was that the keying procedure resulted in tests which were more "factor pure" and because of this were less useful for predicting the GPA criterion which is generally assumed to be factorially heterogeneous. The factor analysis results tended to support this notion. If this second explanation were true, however, a lowering of intercorrelations between the verbal and quantitative subtests should have been observed. But this was not the case. The correlation between V and Q, in fact, was increased substantially when empirical weights were applied. This increase is also quite a bit more than one would expect from the increases in reliability (see Table 5).

Insert Table 5 about here

A third possibility is that the empirical weighting was ordering people not only on verbal and quantitative ability but on some other factor which was reliable but not valid. The pattern of intercorrelations between empirically weighted scores and formula scores supports this last explanation.

It can be seen in Table 6, that although the correlation between empirically

---

Insert Table 6 about here

---

weighted parallel forms goes up, the correlation between the empirically weighted form and the formula-scored parallel form goes down. The correlation between V_1 weighted scores and V_2 formula scores, for example, is lower than that between V_1 and V_2, both formula scored. If, as we had assumed, we were merely increasing the reliability with which we estimated true scores, the correlation between V_1 (weighted) and V_2 (formula scored) should have increased and this increase should have been directly related to the increase in reliability.

The GRE like the SAT is a formula-scored test which means that an examinee's score is equal to the number of correct answers minus $\frac{1}{k-1}$ times the number wrong. The effective weight for an omit under this scoring system is the expected score assuming a random response to the choices. In the usual case this is zero. Whether these assumptions are valid or not is a question which cannot be dealt with here. The important point is that the propensity to omit responses (or conversely, to take risks) is a highly reliable behavior (e.g., Green, 1972; Slakter, 1967).

The keying procedures assigned a weight to the omit category which did not, in most cases, meet or even come close to meeting the formula-score condition that the omit category equal the expected score for the item given a random response to the alternatives. If we consider Table 7

Insert Table 7 about here

---

we see that the actual weight assigned to omit usually differs considerably from what would be the expected weight given a random response. For some of the verbal items shown examinees were actually given a bonus for not responding. In other cases they were penalized. For the quantitative tests they always paid a penalty, which was in some cases quite severe.

One explanation of these results is that for a test given with the usual guessing instructions the empirical keying procedures described capitalize on the tendency to omit and that although this tendency is reliable, it is not valid. It seems reasonable to assume that the reliable but nonvalid variance related to omitting lowered the test-GPA correlation. Following this argument, the correlation between V and Q should have been raised because of the correlated omitting patterns.

The problem is further compounded by the possibility that many omits toward the end of the subtests resulted from failure to reach the items, unlike those occurring near the beginning of the test which presumably were due to a "guessing tendency." A recent test analysis of the same GRE form used in this study concluded that there was some speededness present in all sections (Swineford, 1968). Thus, it is possible that the empirically keyed tests also increased the extent to which the tests measured a speed component.

The empirical keying did result in factors which had highest loadings on the last few items and these factors were much more clearly defined than for the formula-scored tests. Since the GRE is given under formula-score conditions, however, it is difficult to say how much of the increased omitting toward the end of the test is due to a speed factor and how much results from the progressive difficulty of the items.

The present findings are not entirely consistent with previous research in this area. Davis and Fifer (1959) obtained substantial increases in reliability and slight increases in validity after empirically weighting options. It should be pointed out, however, that the Davis and Fifer study departed considerably in several respects from the present investigation. First, the tests were arithmetic reasoning measures scored initially with a priori weights for each option. Secondly, a specially tailored criterion was employed for validation purposes. Third, the validation sample was composed of 251 junior high school students in contrast to the almost 5,000 college students employed in the present study. Finally, Davis and Fifer's tests were administered without guessing instructions so that omits never entered into the analyses (i.e., all subjects were told to attempt every item). Although any or all of these factors could have accounted for the differences between the Davis and Fifer results and those presently reported, it is perhaps worth noting that results quite comparable to those found by Davis and Fifer might have been obtained were the validity phase done in only a few schools. In a few schools the results were fairly impressive in favor of empirical option weighting. Overall, however, the formula scores are slightly but decidedly superior.

The present findings also differ in one important respect from the Hendrickson (1971) study. Hendrickson found that the intercorrelation between verbal and quantitative subforms was lowered through empirical option weighting while the present results suggest the opposite. It may be that the different item types in the SAT account, in part, for this result. A less likely possibility is that the additional iterations performed by Hendrickson caused the V and Q intercorrelations to be lowered.

It appears, therefore, that although the reliability of the GRE tests can be increased substantially through empirical option weighting, much of this increase is due to the measurement of a trait which, though reliable, actually tends to suppress the correlations between tests and criterion.

Conclusions

The results reported here do not support the implementation of empirical option weighting. Increases in reliability are meaningless if, as the present data suggest, they result in decreased validity.

Further research and analyses into the reasons for this phenomenon should be undertaken. As Green (1972) suggested in a recent paper, it may be that when the opportunity to omit is taken away from examinees the sharp increases in reliability will disappear.

References

Baker, F. B., & Hoyt, C. J. The relation of the method of reciprocal averages to Guttman's internal consistency scaling model. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois, April 7, 1972.

Beaton, A. E. Criterion scaling of questionnaire items for regression analysis. Research Bulletin 68-17. Princeton, N. J.: Educational Testing Service, 1968.

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.

Guttman, L. The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.), The prediction of personal adjustment. New York: Social Science Research Council, 1941.

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

Hendrickson, G. F., & Green, B. F. Comparison of the factor structure of Guttman-weighted vs. rights only weighted tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois, April 7, 1972.

Lord, F. M. Some relations between Guttman's principal components of scale analysis and other psychometric theory. Psychometrika, 1958, 23, 291-296.

Slakter, M. J. Risk taking on objective examinations. American Educational Research Journal, 1967, 4, 31-42.

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

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

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]Now at the Bureau of Institutional Research, Yale University.

Table 1

Cross-Validated Internal-Consistency Coefficients

for Three Different Sets of Weights

	Formula	Parallel Forms Keyed		Internally Keyed	
Form	α	α	K[a]	α	K
		Sample A_1			
V_1	.8695	.9285	1.95	.9273	1.91
V_2	.8671	.9259	1.92	.9269	1.94
Q_1	.8458	.9105	1.85	.9143	1.95
Q_2	.8715	.9140	1.57	.9113	1.51
		Sample A_2			
V_1	.8745	.9297	1.92	.9292	1.88
V_2	.8755	.9308	1.91	.9312	1.92
Q_1	.8515	.9131	1.83	.9178	1.95
Q_2	.8725	.9164	1.60	.9125	1.52

[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 2

Cross-Validated Parallel Forms Reliabilities

for Three Different Sets of Weights

		Sample A_1			
	Formula	Parallel Forms Keyed		Internally Keyed	
Test	R	R	K[a]	R	K
V	.8780	.9445	2.36	.9427	2.30
Q	.8722	.9276	1.88	.9183	1.65
		Sample A_2			
V	.8909	.9479	2.23	.9497	2.31
Q	.8742	.9170	1.59	.9267	1.82

[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 reliability obtained with formula-score weights and R_W is the cross-validated reliability obtained with empirical weights.

Table 3

Variance Accounted for by Factors with

Eigenvalues Greater Than One

	Factor	1	2	3	4	5	6	7	8	9	10	11	12	Total
V_1	Formula	14.81	3.89	2.79	2.43	2.33	2.28	2.16	2.14	2.11	2.09	2.07	2.01	41.11
	Weighted[a]	22.72	5.66	3.40	2.87	2.69	2.52	2.28	2.19	2.03	-	-	-	46.2
V_2	Formula	15.25	3.93	2.83	2.56	2.38	2.32	2.21	2.14	2.11	2.11	2.05	2.01	41.90
	Weighted	23.12	5.64	3.59	3.00	2.70	2.54	2.28	2.16	2.06	2.02	-	-	49.11
Q_1	Formula	21.33	5.87	3.92	3.81	-	-	-	-	-	-	-	-	34.93
	Weighted	31.18	7.81	4.36	3.93	-	-	-	-	-	-	-	-	47.28
Q_2	Formula	23.52	5.91	4.13	-	-	-	-	-	-	-	-	-	33.56
	Weighted	31.73	8.83	4.69	3.88	-	-	-	-	-	-	-	-	49.13

[a]Weights used were those derived by keying against parallel forms in sample A.

Table 4

Validity Coefficients for Pairs of Weighted and Unweighted Scores

	Median within School Multiple	Average (Unweighted) within School Multiple[a]	Central Prediction Coefficients V	Q	Multi
V_1+Q_2 (Formula)	.3443	.3248	.1629	.1087	.3185
V_1+Q_1 (Keyed against parallel form)	.2768	.2741	.1341	.0974	.2666
V_1+Q_1 (Keyed against total score)	.2818	.2747	.1347	.0944	.2656
V_2+Q_2 (Formula)	.3135	.3105	.1523	.1023	.3013
V_2+Q_2 (Keyed against parallel form)	.2589	.2637	.1259	.0921	.2550
V_2+Q_2 (Keyed against total score)	.2563	.2627	.1255	.0954	.2577

[a]The sum of the within-school validities was divided by 40 with no weighting based on within-school sample size.

Table 5

Intercorrelations between V and Q for Three Different Types of Scoring Systems

Forms	Formula	Parallel Forms Keyed[a]	Internally Keyed[a]
		Sample A_1	
V_1+Q_1	.4509	.5440 (.4823)	.5454 (.4794)
V_2+Q_1	.4531	.5290 (.4847)	.5487 (.4818)
V_1+Q_2	.4253	.5097 (.4549)	.4906 (.4522)
V_2+Q_2	.4286	.4934 (.4584)	.4889 (.4557)
		Sample A_2	
V_1+Q_1	.4154	.5300 (.4416)	.5223 (.4388)
V_2+Q_1	.4190	.5270 (.4443)	.5051 (.4415)
V_1+Q_2	.4079	.4863 (.4436)	.5064 (.4309)
V_2+Q_2	.4061	.4800 (.4317)	.4894 (.4291)

[a]The values in parentheses represent the 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 6

Correlations between Empirically Weighted Scores and Formula Scores for Parallel Forms[a]

	Formula-Score Reliability	r's between Formula and Empirically Weighted Scores[b]	
Sample A		I	II
Verbal	.8780	.8509	.8518
Quantitative	.8722	.8264	.8599
Sample B			
Verbal	.8909	.8492	.8584
Quantitative	.8742	.8333	.8579

[a]Only scores generated with weights derived by keying on parallel forms are shown.

[b]The correlations between V_1 (Q_1) scored with empirical weights and V_2 (Q_2) scored with formula weights are shown in column I. The correlations between V_2 (Q_2) scored with empirical weights and V_1 (Q_1) scored with formula weights are shown in column II.

Table 7

Empirical Option Weights for Selected Items

Form V_1 - Sample A

Item #	Correct Option	Incorrect Option 1	2	3	4	Omit	Expected[a] Omit
1	.144	-1.180	-1.128	- .211	-1.347	- .474	- .744
11	.194	- .971	- .530	- .718	- .317	- .455	- .468
21	.186	- .656	-1.167	- .955	-1.233	- .753	- .773
31	.273	.126	- .965	- .073	- .174	- .964	- .166
41	.199	- .915	- .398	- .631	-1.018	-1.396	- .553
51	.524	- .039	.131	- .166	- .318	- .581	.026

Form Q_1 - Sample A

1	.128	- .734	-1.089	- .631	- .881	-1.925	- .641
6	.141	- .838	.187	- .501	- .924	-1.186	- .387
11	.158	- .518	- .141	- .443	- .516	-1.266	- .292
16	.397	- .488	- .585	- .918	- .951	-1.117	- .509
21	.287	- .616	- .027	-1.178	- .493	- .740	- .405
26	.666	.150	.166	- .295	.010	- .477	- .139

[a]Expected score for the item given a random response to the alternatives.

Appendix I

Items Assigned (Form QGR1) to Subforms

Subform	
V_1	1, 4, 5, 8, 9, 12, 13, 16, 17, 20, 21, 24, 25, 28, 29, 32, 33, 36, 37, 40, 41, 44, 45, 48, 49, 52, 53, 56, 57, 60, 61, 64, 65, 68, 69, 72, 73, 76, 77, 80, 81, 84, 85, 88, 89, 92, 93, 96, 97, 100.
V_2	2, 3, 6, 7, 10, 11, 14, 15, 18, 19, 22, 23, 26, 27, 30, 31, 34, 35, 38, 39, 42, 43, 46, 47, 50, 51, 54, 55, 58, 59, 62, 63, 66, 67, 70, 71, 74, 75, 78, 79, 82, 83, 86, 87, 90, 91, 94, 95, 98, 99.
Q_1	101, 104, 105, 108, 109, 112, 113, 116, 117, 120, 121, 124, 125, 128, 129, 132, 133, 136, 137, 140, 141, 144, 145, 148, 149, 152, 153.
Q_2	102, 103, 106, 107, 110, 111, 114, 115, 118, 119, 122, 123, 126, 127, 130, 131, 134, 135, 138, 139, 142, 143, 146, 147, 150, 151, 154.

Appendix II

Procedure Used to Key Options against the Total Score Criterion

The first procedure used to key the options of an item in this study assigned the mean standardized total score on the $m-1$ remaining items. An option, for purposes of this study, will be defined as any of the following mutually exclusive categories: (1) the correct alternative; (2) each of the $K-2$ incorrect alternatives; (3) omit (i.e., no response to the item). Thus, for a 5-choice item 6 mutually exclusive categories will be keyed.

		Keying Procedure
Let	k	denote item option
	i, j	denote item
	p	denote individual
	X_{ipk}	denote the total score of individual p choosing option k on item i
	S_i^2 , S_{ij}	= item variances and covariances, where $S_{ij} = S_{ji}$
	S_t^2	= total test score variance
	w'_{ik}	= the original a priori weight assigned to option k of item i
	$\bar{a}_i$	= the mean <u>item</u> score for all individuals in the sample

<u>Step 1</u>. All items are scored conventionally (i.e., right = 1, wrong = $1/c$, where c is one less than the number of alternatives, and

omit = 0). Compute all item means, variances, and covariances as well as mean item scores for each option (note that for the initial keying these latter means will be 1 for correct alternatives, -1/c for incorrect alternatives, and 0 for omits). In addition, compute total test score mean and variance.

At this point the internal consistency coefficient may be computed as follows:

$$\alpha = \left(\frac{m}{m-1}\right)\left(1 - \frac{\sum_{i=1}^{m} S_i^2}{S_t^2}\right).$$

Step 2. Find the mean standard score on the m-1 other items for all individuals choosing option k of item i. This becomes the new weight for option k.

$$W_{ik} = \frac{\bar{X}_{..\,k} - W'_{ik} - \bar{X} + \bar{a}_i}{\sqrt{S_t^2 - S_i^2 - 2\sum_j S_{ij}}}.$$

Step 3. Using the W_{ik} derived in Step 2 rescore all tests.

Step 4. Repeat Steps 2 and 3 until either a desired number of iterations are performed or until coefficient alpha stabilizes.

The procedure outlined here keys on standard scores to avoid the differences in mean item weights which might result for very easy vs. very hard items (remembering that the keying is done on the m-1 remaining items). The procedure has two other desirable aspects. First, the necessity of recomputing the entire total score distribution for each of the m-1 item "tests" used to key the items is avoided. Second, the standardization serves to prevent the test scores from becoming unmanageably large.