DOCUMENT RESUME

ED 062 402 TM 001 341

AUTHOR	Reilly, Richard R.; Jackson, Rex
TITLE	Effects of Item Option Weighting on Validity and Reliability of Shortened Forms of the GRE Aptitude Tests.
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	Apr 72
NOTE	14p.; Paper presented at the annual meeting of the American Educational Research Association (Chicago, Illinois, April 1972)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Academic Aptitude; Achievement Tests; *Aptitude Tests; Predictive Validity; Scoring; *Standardized Tests; Statistical Analysis; Test Interpretation; Test Reliability; Test Validity; Verbal Tests; *Weighted Scores
IDENTIFIERS	Graduate Record Examinations

ABSTRACT

Evidence on how the psychometric properties of verbal and quantitative academic aptitude tests are affected when item options are weighted using rather simple conceptual procedures is presented. This is discussed in connection with the scoring methods used on the Graduate Record Examinations. (DG)

Effects of Item Option Weighting on Validity and Reliability of Shortened Forms of the GRE Aptitude Tests [1]

Richard R. Reilly and Rex Jackson [2]
Educational Testing Service

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

[2] Now at Yale University

The study we carried out took a look at some of the issues related to empirical option weighting using a large and representative data base. We hoped to obtain some fairly general answers to the following questions:

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

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

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

In part, the study attempted to replicate the findings of Hendriksen (1971) and Davis and Fifer (1959) with a high level aptitude test, the Graduate Record Examinations (GRE). Both these studies produced evidence indicating that by empirically weighting options, reliability can be increased by practically significant amounts.

It was hoped that the study would provide further evidence on how the psychometric properties of verbal and quantitative academic aptitude tests are affected when options are keyed using rather conceptually simple procedures.

Method

Test Forms

The first step was to devise 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,

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. It should be noted that the forms within each set were not administered under separate time limits, since the forms were constructed from operational tests. While the more desirable procedure would have been to administer the two parallel forms under separately timed conditions, this was not possible. The GRE, however, is considered by most definitions to be a power test so that any effects due to correlated speed components should have been negligible.

Sample

Next, a space sample of 5,000 answer sheets from the December 1970 administration of the GRE was taken for study purposes. A second sample (sample C) consisting of the answer sheets of individuals from the same administration was taken for validation purposes. The first sample was divided into two randomized block groups of 2500 (samples A and B) by blocking on the total GRE score (V + Q). This increased the probability that total score means and standard deviations for these two groups were approximately equal.

Keying Procedures

(1) Two different types of keying were carried out. The first, designed to increase internal consistency was similar to that described by Hendriksen (1971). The procedure first scored each sub-form using the conventional scoring formula (i.e., rights - ¼ wrongs) and then for each item keyed each option including the omit category, by assigning the mean standard score on the remaining items for all persons choosing that option.

We departed in one respect from Hendriksen's method in that we did not perform any iterations. The second procedure was 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 sub-form of all individuals choosing that option.

Analyses

The next step was to score each sub-form in Sample A using the weights derived in Sample B and vice-versa. Thus, for each sub-form 3 scores were generated: the conventional formula score, the score using weights derived on a parallel form, and the score derived using weights derived by keying on the m-1 remaining items. For each of the three scoring methods, alpha coefficients were computed for each sub-form and intercorrelations among sub-forms were also computed. Thus, cross-validated alpha coefficients and parallel forms reliabilities were obtained for both Samples A and B.

Table 1 shows the cross validated internal consistency coefficients 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 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 for one quantitative sub-form to more than twice the original length for the verbal forms.

These data enabled us to give some pretty solid answers to our first question which was, what happens to internal consistency and parallel forms reliability when options are empirically keyed? The answer is clearly that these measures are increased rather substantially by empirical weighting. It is also worth noting that the two types of keying we carried out were for all practical purposes identical in their effects and, in fact, cross-validated scores yielded by the two methods were correlated close to 1.0 (all correlations were .999 or greater).

The real test of this procedure came in the next set of analyses we performed. For this purpose the answer sheets of over 4,000 college students who had taken the GRE at the same administration from which we selected our keying samples were scored with formula score weights and with empirically derived weights. None of this group were included in the keying sample, but were selected based on 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 sub-forms we computed both single order and multiple correlations between conventionally scored tests and GPA and between empirically weighted scores and GPA. The results were highly consistent. Both single order and multiple correlations were slightly but consistently higher for the formula scores. The weighted scores produced on the average a multiple R .05 less than the multiple R obtained with formula scores. In only one case was there a substantial difference in favor of the weighted scores (.10). The conclusion that empirical option weighting did not lead to any increase in validity was clear enough but the reasons for this were not. One would have expected the more reliable scores to predict the GPA criterion slightly more accurately.

Several explanations were considered. One possibility was that the weighted score reliabilities which held up so well in our carefully constructed A and B samples broke down in the validity sample. This was not the case, however. The reliabilities for the weighted scores were consistently and substantially higher in the validation sample. A second possibility was that the keying procedure resulted in tests which were "factor pure" and because of this were less useful for predicting the GPA criterion which is generally assumed to be factorially heterogenous. The increased alpha coefficients certainly supported this notion. If this second explanation were true, however, one should observe a lowering of intercorrelations between the verbal and quantitative sub-tests. But this was not the case. The correlation between V and Q in fact was increased substantially when empirical weights were applied. This increase was also quite a bit more than one would expect from the increases in reliability (see Table 3).

This led us to consider a third possibility 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 weighted and unweighted scores supports this last explanation. Considering the verbal sub-forms only we see in Table 4 that although the correlation between weighted parallel forms goes up, the correlation between the weighted form and the unweighted parallel form goes down. The r between PF_1 and F_2, for example, is lower than that between F_1 and F_2. If, as we had assumed, we were merely increasing the reliability with which we measured true scores, the correlation between PF_1 and F_2 should have increased and this increase should have been directly related to the increase in reliability.

The pattern was similar for the quantitative sub-forms.

Our analyses are continuing but at this point we can at least suggest what may be happening. The GRE like the SAT is a formula scored test which means that an examinees' score is equal to the number of correct answers minus $\frac{1}{k}$ times the number wrong. The effective weight for an omit under this scoring system is the mean 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. Slakter, 1967).

The procedure we used to key 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 mean expected score for the item given a random response to the alternatives.

If we consider Table 5 we see that the actual weight assigned (in the O column) differs considerably from what would be the mean expected weight (the O' column). For some of the verbal items shown examinees were actually given a bonus for not responding. In other cases they paid a penalty. For the quantitative tests they always paid a penalty which was in some cases quite severe.

What we are suggesting is that when a test is given with the usual guessing instructions the empirical keying procedures described capitalize on the tendency to omit and that while this tendency is reliable, it is not valid. This would explain the decreases in validity in spite of increases in reliability that we observed and would also explain the increase in the correlation between V and Q.

A new keying procedure which hopefully will offer more promise has been worked out and will be applied shortly. This procedure assigns weights to responses which are optimum in the least squares sense, but subject to the constraint that the weight for omit equals the average of the other 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.

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

Slakter, M.J. Risk taking on objective examinations. American Educational Research Journal, 1967, 4, 31-43.

Table 1

Cross-Validated Internal-Consistency Coefficients

for Three Different Sets of Weights

Sample A

	Formula	Parallel Forms Keyed		Internally Keyed	
Form	α	α	K[1]	α	K
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 B

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

---

[1] 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 α obtained with cross-validated empirical weights.

Table 2

Cross-Validated Parallel Forms Reliabilities

for Three Different Sets of Weights

		Parallel Forms Keyed		Internally Keyed	
Test	Formula	R	K[1]	R	K
		Sample A			
V	.8780	.9445	2.36	.9427	2.30
Q	.8722	.9276	1.88	.9183	1.65
		Sample B			
V	.8909	.9479	2.23	.9497	2.31
Q	.8742	.9170	1.59	.9267	1.82

---

[1] 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 R obtained with cross-validated empirical weights.

Table 3

Intercorrelations Between V and Q

for Three Different Types of Scoring Systems

Sample A

	Formula	Parallel[1] Forms Keyed	Internally Keyed
$V_1 Q_1$	.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 B

$V_1 Q_1$	.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)

[1]The values in parentheses 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

Sample A Correlations Between Formual Scores

and Scores Using Weights derived on Parallel Forms

	F 1	F 2	PF 1	PF 2
F.		.8780	.9161	.8518
F_2			.8509	.9200
PF_1				.9434

Table 5

Empirical Option Weights for Selected Items

Form V_1 Sample A

Item #	R	W_1	W_2	W_3	W_4	O	O′
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

Item #	R	W_1	W_2	W_3	W_4	O	O/
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