DOCUMENT RESUME

ED 062 389 TM 001 327

AUTHOR Hendrickson, Gerry F.; Green, Bert F., Jr.

TITLE Comparison of the Factor Structure of Guttman-Weighted vs. Rights-Only-Weighted Tests.

SPONS AGENCY College Entrance Examination Board, New York, N.Y.

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 Analysis of Variance; Correlation; *Factor Analysis; *Internal Scaling; Item Analysis; *Scoring Formulas; *Test Construction; Test Reliability; Test Validity; *Weighted Scores

IDENTIFIERS *Guttman Weighted Test; Rights Only Weighted Test

ABSTRACT

It has been shown that Guttman weighting of test options results in marked increases in the internal consistency of a test. However, the effect of this type of weighting on the structure of the test is not known. Hence, the purpose of this study is to compare the factor structure of Guttman-weighted and rights-only-weighted tests and to relate the change in structure to the change in internal consistency. Test items were intercorrelated and the resultant matrices factored, first weighting with Guttman weights then with rights-only ones. More variance was accounted for by the first several factors of Guttman-weighted tests than by these same factors in rights-only weighted tests. The tests for which there was most increase in variance were those for which there was greatest increase in internal consistency. Finally, the factor structure and conventionally-weighted tests were quite different. The implication of these structural changes on test validity is discussed. (Author)

Comparison of the Factor Structure of Guttman-Weighted vs. Rights-Only-Weighted Tests [1]

Gerry F. Hendrickson[2] and Bert F. Green, Jr.
Johns Hopkins University

[1] The research reported here was supported by the College Entrance Examination Board.

[2] Present address:

Office of Program Planning and Evaluation
U.S. Office of Education
400 Maryland Avenue S.W.
Washington, D.C. 20202

Comparison of the Factor Structure of Guttman-Weighted vs. Rights-Only-Weighted Tests

Abstract

It has been shown that Guttman weighting of test options results in marked increases in the internal consistency of a test. However, the effect of this type of weighting on the structure of the test is not known. Hence, the purpose of this study is to compare the factor structure of Guttman-weighted and rights-only-weighted tests and to relate the change in structure to the change in internal consistency.

Test items were intercorrelated and the resultant matrices factored, first weighting with Guttman weights then with rights-only ones. More variance was accounted for by the first several factors of Guttman-weighted tests than by these same factors in rights-only-weighted tests. The tests for which there was most increase in variance were those for which there was greatest increase in internal consistency. Finally, the factor structure and conventionally-weighted tests was quite different. The implication of these structural changes on test validity is discussed.

Comparison of the Factor Structure of Guttman-Weighted vs. Rights-Only-Weighted Tests

This research is an outgrowth of differential option weighting studies reported by Davis (1959), Davis and Fifer (1959), Sabers and White (1969), and Hendrickson (1971), with the latter study being most directly related. Hendrickson (1971) used an empirical weighting technique described by Guttman (1941) to assign values to the options in each subtest of the Scholastic Aptitude Test (SAT). Guttman's technique is designed to maximize the internal-consistency reliability (coefficient α).

It was found, as expected, that the internal-consistency reliability of each subtest increased substantially as a result of Guttman weighting. However, the intercorrelation of all subtests, except the two verbal subtests, decreased. These results invited the interesting speculation that the underlying structure of a test changed when scored with Guttman weights. In other words, the Guttman-weighted and conventionally-weighted tests were measuring different things. The following line of reasoning led to this hypothesis.

If the internal-consistency coefficient is greater for a Guttman-weighted test than for a conventionally-weighted test, and if the loading pattern of items on factors is similar in the two cases, then the former test is measuring the same function as the latter test, but measuring it more efficiently. Therefore, the former test should correlate better with another test or external criterion variable than the latter. If, however, the loading pattern of items on factors is different, then the two tests are measuring different functions. In this case it is likely that a Guttman-weighted test will correlate less well with another test or outside criterion than a conventionally-weighted test. The fact that the values of the intercorrelation coefficients of the Guttman-weighted subtests of the SAT decreased is, therefore, strong evidence that the structure of the subtests was altered.

The purpose of this study is to compare the factor structure of Guttman-weighted and rights-only-weighted SAT. A comparison of the factor structure will point out the effect of this type of weighting on the underlying structure of the test. Such information should make "blind" empirical weighting somewhat less blind.

PROCEDURE

The sample used in this study consisted of 200 boys and 200 girls, randomly selected from the examinees retained by Educational Testing Service for item-analysis purposes. The test used was form QSA43 of the SAT, administered at the College Entrance Examination Board's regular testing in November, 1968. All four subtests in the scored portion of the SAT were used. These subtests are numbered 1, 2, 4, and 5. Subtest 3 is not part of the scored portion of the SAT; it is used only for equating and pretesting purposes. Hence, it was not included in this study.

The verbal section comprises subtests 1 (40 items) and 2 (50 items); the mathematicalcomprises subtests 4 (35 items) and 5 (25 items).

Guttman weights were calculated for all options in the SAT in the way described by Hendrickson (1971). These weights were determined separately for each subtest, the weight of a particular option being the total score on the other items of the subtest for all examinees who marked the option in question. The weight for "omit" was computed in the same manner as the weights of the other five options in an item. Crossvalidation of the weights, a procedure which usually should be carried out when dealing with the empirically-calculated Guttman weights, was deemed inappropriate in this study.[3]

The comparison of the factor structure of Guttman and conventionally-weights tests involved two investigation--the first design to compare the difference in common variance and in variance per factor and the second designed to compare the loading pattern of items on factors. In the first investigation the basic unit of analysis was the inter-item, product-moment correlation matrix. Two inter-item correlation matrices were calculated for each subtest of the SAT: one consisting of the correlation coefficients of items with Guttman weights (a six-category distribution for each item),

[3] The intent of this study is merely to show how weighting might effect the structure of a test. Maximum structural changes will result when item weights are calculated on the sample of individuals whose tests are then factored. The extent to which these structural changes exist when weights are calculated on another random sample from the population depends on the extent to which the two samples are similar. The blurring effects of sampling error is undesirable in this instance.

and the other consisting of correlation coefficients of items weighted dichotomously (1 or 0). These correlation matrices were factored using the Minres procedure (Harman and Jones, 1966; Harman, 1968, pp. 187-211). Eleven factors were extracted. This number was chosen, somewhat arbitrarily, because in most cases all factors with eigenvalues greater than one were among the eleven. Further, eleven seemed to be a sufficiently large number to allow trends to appear.

The coefficients used in the first analysis, product-moment correlation coefficients, are calculated from the actual item weights from which the total test score is formed. Hence, these coefficients are the appropriate ones to use when comparing common variance and factor variance of unrotated factors. However, a difficulty factor often emerges in analyses using product-moment correlation coefficients of dichotomously scored variables. Thus in the second analysis, designed to compare the loading pattern of items on factors, matrices of tetrachoric correlation coefficients were used for the dichotomously scored variables in order to lessen the effects of a difficulty factor. Matrices of product-moment correlation coefficients were again used for Guttman-weighted tests. Factors were obtained by the Minres procedure, followed by a varimax (Kaiser, 1958) rotation.

RESULTS AND DISCUSSION

The results of the factor analysis in the first investigation are displayed in Table 1. The first row under each subtest contains the variance accounted for by the first eleven factors extracted from the inter-item, product-moment correlation matrix of the Guttman-weighted subtest. The second row contains the same thing for the rights-only-weighted subtest. At the end of each row is the total variance accounted

Insert Table 1 about here

for by the eleven factors. In all cases the amount of total variance was substantially larger for a Guttman-weighted subtest than for the corresponding rights-only-weighted subtest. Also, a larger amount of variance was accounted for per factor for the first several factors in a Guttman-weighted subtest.

The internal-consistency reliability coefficients for Guttman- and rights-only-weighted subtests are listed in the last column of Table 1, along with the percent by which a rights-only-weighted test would have to be lengthened (as calculated by the Spearman-Brown prophecy formula) to account for the increase in internal consistency that came about as a result of Guttman weighting. It is important to note that the tests for which there is greatest increase in variance are the ones for which there is greatest increase in internal-consistency reliability (as measured by effective increase in test length). This finding is thoroughly predictable because an increase in internal consistency without adding more items implies that the mean intraclass coefficient among the existing items has gone up (Stanley, 1957 and 1971). Thus, items must share more variance in common.

Comparison of the varimax factor matrices obtained by rotating the first eleven Minres factors of the tetrachoric correlation matrix shows that the structure of Guttman- and rights-only-weighted subtests are considerably different.[4] In only one case (Guttman-weighted subtest 1) was there a clear-cut difficulty factor. In this case the difficulty factor was the one which accounted for the next-to-the-most variance. In four of the eight factor matrices a speed factor emerged; however, some confounding was present in two cases. Three of the speed factors emerged in the Guttman-weighted subtests, in all cases as the factor which accounted for most variance. Thus Guttman weighting may be capitalizing on speed at the expense of intellectual variables. (In the verbal subtests the most difficult items are often not the last ones, making it possible to distinguish between difficulty and speed factors. In the mathematical subtests, on the other hand, the last items are usually among the most difficult ones; thus, the naming of speed and difficulty factors must be more tentative in these two subtests.) Further attempts to name the factors were fruitless. Suffice it to say that the underlying structures of both the Guttman- and rights-only-weighted SAT are extremely complex.

[4] These eight factor matrices are displayed in a Johns Hopkins technical report which can be requested from the author.

The findings of this study may offer interesting implications for validity. It is likely that the structural changes imposed by differential weighting for internal consistency will probably make a test correlate less well with an outside criterion, such as grade point average. Guttman (1941, pp. 296-297) himself predicted this outcome, and his prediction was born out almost completely in Hendrickson's (1971) study. Hendrickson intercorrelated the four subtests of the SAT (6 combinations) and found that in five of the six cases the subtests correlated less well with each other when they were weighted with Guttman weights than when they were weighted rights-only. The one case in which the two subtests correlated more highly when weighted with Guttman weights involved the two verbal subtests. It is noteworthy that this pair of subtests are more alike in item type, item difficulty, and item content than the subtests of any other pair. A Guttman-weighted test will correlate better with an outside criterion only if the structural changes brought about by the weighting make the test structurally more like the outside criterion. This outcome is unlikely unless the tests are extremely similar to begin with, especially since Guttman weighting appears, in some cases, to capatalize on spurious factors, such as speed. Thus, it seems to be a good bet that, in most cases, even though the homogenity of a Guttman-weighted test will go up, the validity will go down.

This conclusion directly contradicts the prediction expressed in the correction-for-attenuation formula, which says that the correlation coefficient between two tests increases if the reliability of either or both of the tests increases. Some twenty years ago

Loevinger and her colleagues (Loevinger, 1954; Loevinger, Gleser, and DuBois, 1953) noted and commented extensively upon such "attentuation paradoxes." Recently Gleser (1971) shed further light on the subject. The simple way to resolve the apparent "paradox" in this case is to remember that the correction-for-attenuation formula is only appropriate if the test measures the same thing after the reliability is increased as it did before, i.e., if the loading pattern of items on factors has not changed, the loadings are merely higher. The correction-for-attenuation formula is completely inappropriate in cases when structural changes have occurred.

Other researchers have used slightly different versions of the empirical weighting scheme described by Guttman. The essence of this class of empirical weighting techniques is that the weight of an option is based on the scores of people marking that option on a criterion score distribution. For Guttman (1941) and Hendrickson (1971) the criterion was the score distribution of the test in question; their aim was to improve internal consistency. For Davis and Fifer (1959) the criterion was the score distribution of a parallel-form of the test in question; their aim was to improve parallel-forms reliability. For Sabers and White (1969) the criterion was the score distribution on an outside achievement test; their aim was to improve validity. It should be emphasize that tests weighted by any of these methods will almost certainly be different structurally than a rights-only-weighted test.

One sometimes hears the argument that it is desirable to devise ways of improving the reliability of a test. If this could be accomplished, the argument goes, more measuring ability could be packed into less test-taking time, thus leaving time for the measurement of other attributes. This study has shown that the former line of reasoning is not always an appropriate one, especially if it is an internal-consistency coefficient which is increased and if the increasing is done by an empirical weighting procedure. Blind empirical weighting, indeed, might cause the internalcconsistency to increase, but it also might result in structural changes which make the test measure something altogether different.

REFERENCES

Davis, F. B. Estimation and use of scoring weights for each choice in multiple-choice test items. *Educational and Psychological Measurement*, 1959, *19*, 291-298.

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.

Gleser, L. J. The attenuation paradox and internal consistency. Presented at the American Educational Research Association meeting, February 1971.

Guttman, L. The quantification of a class of attributes: a theory and method of scale construction. In Paul Horst, *The prediction of personal adjustment*. N. Y.: Social Science Research Council, 1941.

Harman, H. H. *Modern Factor Analysis*. (2nd ed.) Chicago: The University of Chicago Press, 1968.

Harman, H. H. & Jones, W. H. Factor analysis by minimizing residuals (Minres). *Psychometrika*, 1966, *31*, 351-368.

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

Kaiser, H. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 1958, *23*, 187-200.

Loevinger, J. The attenuation paradox in test theory. *Psychological Bulletin*, 1954, *51*, 493-504.

Loevinger, J., Gleser, G. C., & DuBois, P. H. Maximizing the discriminating power of a multiple-score test. *Psychometrika*, 1953, *19*, 309-317.

Sabers, D. L., & White, B. W. The effect of differential weighting of individual item responses on the predictive validity and reliability of an aptitude test. *Journal of Educational Measurement*, 1969, *6*, 93-96.

Stanley, J. C. K-R20 as the stepped-up mean item intercorrelation. Pp. 78-93 in the *14th Yearbook of NCME*, 1957.

Stanley, J. C. Reliability, Ch. 13 in R. L. Thorndike (Ed.) *Educational Measurement* (2nd ed.) Washington: American Council on Education, 1971.

Table 1

Variance Accounted for by Factors of

Guttman- and Rights-Only-Weighted Subtests

	Factors											Total Variance	Internal Consistency
	1	2	3	4	5	6	7	8	9	10	11		
Subtest 1													
Guttman	7.44	1.59	1.27	1.13	.63	.60	.52	.47	.43	.40	.38	14.86	.8978
Rights-only	5.70	1.35	1.03	.76	.65	.64	.59	.52	.49	.41	.39	12.51	.8546
													49.46%*
Subtest 2													
Guttman	8.04	2.31	1.41	1.04	.91	.74	.72	.70	.58	.55	.52	17.51	.9048
Rights-only	5.99	1.10	1.02	.98	.81	.69	.67	.64	.61	.58	.53	13.61	.8658
													47.32%*
Subtest 4													
Guttman	6.61	2.19	1.65	.95	.68	.64	.58	.56	.43	.36	.32	14.95	.8886
Rights-only	4.70	1.35	.76	.71	.66	.55	.45	.40	.36	.35	.31	10.58	.8270
													66.86%*
Subtest 5													
Guttman	5.06	1.87	.81	.69	.56	.48	.40	.39	.34	.33	.26	11.20	.8532
Rights	4.13	.85	.72	.68	.57	.55	.46	.39	.34	.32	.30	9.30	.8119
													34.65%*

*Effective increase in test length