DOCUMENT RESUME

ED 064 323 TM 001 488

AUTHOR	Green, Bert F., Jr.
TITLE	The Sensitivity of Guttman Weights.
INSTITUTION	Johns Hopkins Univ., Baltimore, Md. Dept. of Psychology.
SPONS AGENCY	National Science Foundation, Washington, D.C.
PUB DATE	Apr 72
NOTE	10p.; 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	Comparative Analysis; *Reliability; *Scoring Formulas; *Test Results; *Test Wiseness; *Weighted Scores
IDENTIFIERS	*Guttman Weights

ABSTRACT

The use of Guttman weights in scoring tests is discussed. Scores of 2,500 men on one subtest of the CEED-SAT-Verbal Test were examined using cross-validated Guttman weights. Several scores were compared, as follows: Scores obtained from cross-validated Guttman weights; Scores obtained by rounding the Guttman weights to one digit, ranging from 0 to 8; Scores obtained by approximating the Guttman weights by weights of 0, 1, and 2; Scores obtained by weighting a right answer 2, and a wrong answer 1; Scores obtained by weighting a right answer 1, and a wrong answer -- 1/4; Number right; Number wrong; and Number omitted. The reliabilities and intercorrelations of these scores are shown in tabular form. It is concluded that most of the advantages of Guttman weights come from weighting "omits" less than "wrong" answers. Because instructions in common use encourage people to omit items on a test, Guttman weighting is not recommended since with these weights the best strategy is to guess. (DB)

DEPARTMENT OF PSYCHOLOGY

The Johns Hopkins University

Department of Psychology
Ames Hall
The Johns Hopkins University
Baltimore, Maryland 21218

The Sensitivity of Guttman Weights*

Bert F. Green, Jr.

The Johns Hopkins University

Paper presented at the annual meeting of the

American Educational Research Association

Chicago, Illinois, April 7, 1972

*This work was supported by a research grant, GB 15839 from the National Science Foundation.

THE SENSITIVITY OF GUTTMAN WEIGHTS

Bert F. Green, Jr.

The Johns Hopkins University

The method of reciprocal averages (Mitzel & Hoyt, 1954) and Guttman's scaling metric (Guttman, 1941) have been around for a long time, yet they are not much used in practice. It is much quicker for a human scorer to count the number right than to add up a lot of weights. Today's computers and test scoring machinery have removed this onus; it costs the computer less than a millisecond to use weights rather than counts. The renewed interest in differential weighting, e.g. Stanley & Wang (1970) is thus a trend of our mechanized times.

The motivation for differential weighting of item options is clear. A score of 1 or 0 on an aptitude test item is small pay-off for the student's extreme mental anguish in answering the question. Could we not milk some extra information from his answer? If he gets the item right, that is that, but if he chooses one of the distractors, might he not deserve partial credit for choosing a "better" distractor? Skeptics argue there is very little to be gained, especially with a reasonably homogeneous test, and that there is more pay-off in better items than in better ways of scoring present items.

It is well known that differential weighting is at best a second order improvement. The Likert method of attitude scaling is a case in point (See Green, 1954). Respondents rate each statement "strongly agree," "agree" "indifferent" "disagree" "strongly disagree," and get scored 5, 4, 3, 2, or 1 for their effort. Originally Likert proposed choosing item weights for maximum discrimination among the respondents, but found that the simpler 5-4-3-2-1 method yielded scores correlating in the high 90's with the best weighted scores. The added precision of the scoring method was simply lost in the large error variance of individual item responses.

Green (1969) recently discussed the conditions under which differential regression weights are reasonable, arguing that simple one-digit weights are virtually indistinguishable from 3-digit beta weights. Applications of earlier results due to Wilks & Sheffe permit stating a multidimensional interval of indistinguishability of weighting schemes.

The present context is very similar to the regression situation, in that the comparison is between equally weighted distractors (all 0) and differentially weighted distractors. But zero weights are a very special case, and amount to adding or dropping information rather than simply weighting it differently. The usual theorems do not apply to zero weights.

It is clear empirically that reliability can be significantly improved by Guttman weighting; the remaining question concerns how seriously to take the weights. How much is lost by using one-digit weights, or a 5-4-3-2-1 scheme? Even more extreme, what about giving 0 for omit, 1 for any wrong distractor, and 2 for the right answer. (Not, of course, telling the students what we plan to do).

Theoretical analysis indicates that, as in the case of regression weights, we will lose almost nothing by using one-digit weights-- 0-9, and not much more by using 5-4-3-2-1-0. But analysis cannot preceed further without the specific data at hand, so we must turn to the empirical results. Here it would be handy to have an interactive computer system so that we could try different weighting schemes on the basis of results of earlier weights, but we have made do with batch processing.

Empirical Results

One subtest of the CEEB-SAT-Verbal Test was examined in detail. The sample is described more fully by Hendricksen (1971) whose help in obtaining the data is gratefully acknowledged. Our sample consisted of 2,500 men. Cross-validated Guttman weights were used. We compared several scores, on this group, as follows.

1. Scores obtained from cross-validated Guttman weights
2. Scores obtained by rounding the Guttman weights to one digit, ranging from 0 to 8
3. Scores obtained by approximating the Guttman weights by weights of 0, 1, and 2
4. Scores obtained by weighting a right answer 2, and a wrong answer 1
5. Scores obtained by weighting a right answer 1, and a

wrong answer -- 1/4

6. Number right

7. Number wrong

8. Number omitted

The, reliabilities and intercorrelations of these scores are shown in Table 1.

It is clear that the first four scores are virtually identical, correlating at least .98 with each other. Moreover these four scores have nearly the same reliability. The simple weights loose a slight amount (.01) but this is not much, compared with the gain of all these four scores over the formula score (R - 1/4W) or the simple number right.

It is worth noting that the three versions of the Guttman weights, unrounded, slightly rounded and very rounded, all correlate at least .99 whereas the score 2*R+W correlates only .98 with each. So there is definitely some information in the Guttman scores that is not in the 2*R+W scores, but not very much. We must conclude, then, that most of the advantages of Guttman weights comes from weighting omits <u>less</u> than wrong answers.

Table 1a Reliabilities of several scores obtained from the CEEB-SAT-Verbal Test.

	Score	Reliability	Mean	Standard Deviation
1.	Guttman weights	.894	12.82	.448
2.	One-digit Guttman Weights (0-8)	.890	220.37	22.12
3.	Simple Guttman Weights (0-1-2)	.880	58.43	10.13
4.	2*R+W	.879	57.37	9.60
5.	R - 1/4W	.850	17.45	7.87
6.	R	.855	21.20	6.67
7.	W	.805	14.98	5.97
8.	O	.865	3.82	4.40

Table 1b Intercorrelation of Scores

	1	2	3	4	5	6	7	8
1	1.000	.998	.993	.980	.893	.938	-.521	-.715
2		1.000	.992	.978	.895	.938	-.524	-.712
3			1.000	.982	.901	.944	-.531	-.711
4				1.000	.859	.916	-.439	-.793
5					1.000	.992	-.836	-.371
6						1.000	-.763	-.481
7							1.000	-.199
8								1.000

Discussion

The important fact, obtained from an initial scan of the Guttman weights, and borne out by subsequent analysis, is that there is something very special about omits. With the usual instruction about not wasting time by guessing, it turns out that people who omit items do more poorly on the test. The omit option is to be weighted less than the distractors. Most of the gain in the Guttman weights is in the differential treatment of the omits. People who omit items do badly on the test, and by weighting the omit category low, the test reliability is increased.

We need to know more about the properties of the "omit" score. Reilly and Jackson (1972) have presented indirect evidence that suggests that the "omit" score is reliable but invalid for graduate school grades. But the validity of predictors of graduate performance is generally low. It would be helpful to repeat the study using the CEEB-SAT, with undergraduate grades as the criterion. If the omit score is reliable, what does it measure? Incidentally, in our data it is not the case that the people who omit are too slow. At least it is not true that the omits are all bunched at the end. People omit items because they don't know the answer, more often than they omit items because they don't reach them.

At a practical level, it appears that we cannot ethically use Guttman weights, because the instructions in common use encourage omitting items, whereas when Guttman weights are used, the best strategy is to guess. It is an open question whether Guttman weights will be useful when applied to data obtained when everyone guesses if he doesn't know or doesn't have time.

Perhaps we can develop a separate test on which those who omit many items will score well; if so, that test could be used as a suppressor variable.

Presumably the problem would go away with a tailored test given by a computer. An answer would be required for every item, and both speed and correctness would be relevant factors. Most studies of tailored testing (e.g. Lord (1970), Green (1970)), have treated only items scored right and wrong. Work is needed on differential weighting of options, and on the use of second attempts, in this computer-based situation.

Until we can solve these problems, it seems clear that Guttman weighting is not to be recommended.

References

Green, B.F. (1954) Attitude Measurement. Addison Wesley Press, (reprinted from Lindzey, G. (ed.), Handbook of Social Psychology).

Green, B.F. (1968) Psychometric sensitivity & statistical stability. Paper presented at Psychometric Society meeting, April 1968.

Green, B.F. (1970) Comments on Tailored Testing. In Holtzman (ed.), Computer Based Instruction, Testing, & Guidance. New York: Harper & Row.

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

Hendricksen, G.F. (1971) The effect of differential option weighting on multiple choice objective tests. Journal of Educational Measurement.

Lord, F.M. (1970) Source Test Theory for Tailored Testing. In Holtzman (ed.), Computer Based Instruction, Testing, & Guidance. New York: Harper & Row.

Reilly, R.R. & Jackson, R. (1972) Effects of item option weighting on validity and reliability of shortened forms of the GRE aptitude tests. Paper read at AERA meeting, 1972.

Wang, M.D. & Stanley, J.C. (1970) Differential weighting: a review of methods and empirical studies. Review of Educational Research, 40, 663-705.