Biserial Weights: A New Approach to Test Item Option Weighting

John G. Claudy
American Institutes for Research

Optıon weighting is an alternative to increasing test length as a means of improving the reliability of a test. The effects on test reliability of option weighting procedures, including two procedures not previously reported, were compared in two empirical studies using four independent sets of items. In all cases Guttman weights and Biserial weights (based on the Brogden Biserial or Cleman's Lambda coefficient) were superior to Number Right and Correction for Guessing weights. Overall, Biserial weights appeared to be superior to Guttman weights.

An important concern in the development of any testing program, especially one in which the results will be used to make decisions affecting individuals, is that the reliability of the testing instruments should be as high as possible. Given that the items are properly constructed, such a goal is usually approached by increasing the length of the test; however, this approach cannot be used when the testing time is limited. An alternative approach is that of "option weighting." With the most commonly used scoring procedures, an examinee receives one point for selecting the correct response and zero or $-1/(N-1)$ points (where N equals the number of options) for selecting any of the alternate responses.

With an option weighting procedure, an examinee who selects the correct response to an item receives some positive number of points (perhaps fractional). At the same time, an examinee who selects any of the alternate responses also receives some number of points, depending on the particular alternate response chosen. The number of points received for an alternate response can be either positive or negative. For example, a response which is closely associated with the correct response might receive +.75 points, while a response unrelated to or conflicting with the correct response might receive −.85 points. An examinee's score on the total test is the sum of the points received for the individual item responses chosen. A further and potentially quite important characteristic of tests scored by option weighting procedures is that there are few, if any, tie scores since virtually every response in the test earns a different number of points. Having as small a number of tie scores as possible is important in test applications which focus on selection or assignment and which require accurate ranking of the examinees.

There are two approaches to determining the number of points to be awarded for selecting each of the possible response options: judgmental and empirical. In the judgmental procedure, a group of experts in the content area in which the item is included review the options and decide on the number of points to be awarded for

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 2, No. 1 Winter 1978 pp. 25-30
© Copyright 1978 West Publishing Co.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

each choice. This is the approach used by Davis (1959) and Davis and Fifer (1959). An empirical procedure proposed by Guttman (1941) has become practical with the development of large, high-speed computers. Guttman's procedure was developed within the context of non-cognitive items for which there is no correct response, but it also can be applied to cognitive items.

The empirical approach to option weighting has been successfully employed in several studies, e.g., Echternacht (1973), Hendrickson (1971a, 1971b), Hendrickson and Stanley (1970), Reilly (1972), and Reilly and Jackson (1972). These studies have compared the reliabilities of tests scored using number right and/or formula scores with the reliabilities of the same tests scored using empirically derived Guttman option weights. Results have varied from test to test and group to group. However, in every case the use of option weights has resulted in an increase in test reliability, although not necessarily in validity.

These increases in reliability can be expressed in terms of the effective test length of the option-weight-scored test. The effective length, arrived at by using the Spearman-Brown formula, is an indication of the amount a conventionally scored test would have to be lengthened in order to achieve the increase in reliability resulting from the use of option weights. For example, an effective test length of 1.60 for a test scored using option weights indicates that the test would have to be increased to 1.6 times the original length in order to achieve the same reliability using conventional (number right) scoring. For the previous studies, the effective test lengths have ranged from about 1.20 to 1.80 with no change in either testing time or administrative procedures.

In order to implement empirical option weighting procedures, a criterion measure upon which to base the weights is required. This criterion can be either external or internal, that is, a score based on the test itself. If the test itself is used for the criterion measure, a form of "bootstrapping" procedure can be used. In this procedure several iterations of test scores are used to arrive at the final weights. "Bootstrapping," as used for this study, is described below. For a more complete review of published materials on option weighting, see the above-mentioned papers or summary reviews by Stanley and Wang (1968a, 1968b, 1970) and Wang and Stanley (1970).

Purpose

Certain of these option weighting procedures were investigated in the studies described below as a part of a project carried out at the American Institutes for Research to develop the New Medical College Admissions Test (MCAT) for the Association of American Medical Colleges. In addition to the three scoring procedures described above (number right, correction for guessing, and Guttman weights), two new procedures were also investigated. The five procedures were:

1. *Number Right* scoring, in which the correct option receives 1 point and all other options including "omit" receive 0 points. This is the most frequently used procedure; it provides the standard against which the other proce lures were evaluated.
2. *Correction for Guessing* scoring, in which the correct answer receives 1 point, all other options receive $-1/(N-1)$ points (where N equals the number of options), and "omit" receives 0 points.
3. *Guttman Weights* scoring, in which the weight assigned to each option, including "omit," is the z-score corresponding to the mean score on the other test items for examinees who select the option. Theoretically, the weights can take any value a z-score can assume.
4. *Biserial Weights* scoring, in which the weight assigned to each option, including "omit," is the Brogden Biserial (Brogden, 1944) or Clemans Lambda (Clemans, 1958) correlation coefficient between selecting the option (scored 1 or 0) and the score on the other items in the test. In effect, this is a dis-

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

crimination index for the option. Theoretically the weights can take any value between minus one and plus one inclusive.

5. *Proportion Weights* scoring, in which the weight assigned to each option including "omit" is the proportion of examinees in an upper score subgroup of examinees who select the option. Theoretically, the weights can take any value between zero and one inclusive, and sum to one for each item. Two overlapping subgroups were used in this study: the upper 25% and the upper 50% of examinees on the basis of total score.

Study 1

Method

The test items used in this study were 30 science knowledge items administered to regular MCAT examinees in the Spring of 1975. The items were not homogeneous since their content was drawn from biology, chemistry, and physics. All items were discrete from each other.

The dependent variable studied was the split-half reliability coefficient raised by the Spearman-Brown formula in order to estimate the reliability for the full 30-item test. The test halves were determined by using a 1-2-2-1-1-2-2-1 . . . pattern. Half of each type of item (biology, chemistry, and physics) fell into each test half.

Number Right and Correction for Guessing weights are a priori weights; they are not subject to variation due to the sample of examinees used. Guttman, Biserial, and Proportion weights, on the other hand, are subject to this variation; thus, they must be cross-validated in an independent sample. Two nonoverlapping samples of 300 examinees each were randomly selected from the total available pool: the first for use in calculating the weights and the second for cross-validation. Each weighting procedure was carried through three iterations to allow the weights to stabilize. For the first iteration, the criterion score for each item was based on the Number Right score received by the examinee on the set of items minus the item itself. Using these criterion scores, the first set of option weights (Guttman, Biserial, or Proportion) was calculated; new criterion scores based on option weights were then determined, again using the 29 non-overlapping items. For the second iteration, the option-weight-based criterion scores were used to develop a second set of option weights. This second set was in turn iterated to arrive at a third and final set of option weights for the test items. The reliability coefficients reported are based on scores arrived at using this third set of option weights. Iteration was stopped after three iterations, because by this point, the option weights had stabilized and additional iterations resulted in increases in test reliability which were equal to or less than .001 per iteration.

Results

The results, shown in Table 1, are rather striking in both the calculation and cross-validation samples. Scores based on Biserial weights are the most reliable and these were closely followed by scores based on Guttman weights. Scores based on Proportion weights were also superior to scores based on Number Right scoring, but, especially in the calculation sample, not to the degree shown by the other option weighting procedures. What was rather surprising was the fact that scores based on Correction for Guessing weights were less reliable than scores based on Number Right scoring. Overall, even though the study was only exploratory, it appears that both Guttman and Biserial weights offer promise.

Study 2

Method

Because of the positive results of Study 1, a second study was carried out using 78 science knowledge items administered to regular MCAT examinees in the Fall of 1975. In this second study the Proportion weights scoring procedure was not employed, because it had not been as effective as the other empirical procedures investi-

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 1

Results of Study 1--Five Option Weighting Procedures Applied to a Test of 30 Heterogeneous Science Knowledge Items

	Number Right Scoring	Correction for Guessing Scoring	Guttman Weights Scoring	Biserial Weights Scoring	Proportion Weights Scoring(25%)	Proportion Weights Scoring(50%)
Calculation Sample (N=300)						
Split-half reliability	.74	.73	.86	.86	.80	.81
Effective test length	1.00	.94	2.16	2.20	1.36	1.44
Cross-Validation Sample (N=300)						
Split-half reliability	.74	.72	.80	.81	.78	.79
Effective test length	1.00	.92	1.38	1.55	1.28	1.36

gated in Study 1 and because, in terms of computer costs, it is more expensive to calculate. Otherwise, the procedures used in Study 2 were the same as those used in Study 1.

Actually, Study 2 consisted of three independent substudies, each using a different homogeneous subset of the items. Substudy 2A used 24 biology items, Substudy 2B used 30 chemistry items, and Substudy 2C used 24 physics items. For each study, two nonoverlapping randomly selected samples of 800 cases each were selected from the 21,000 examinees responding to the items. The first sample was used for calculating the weights, and the second sample was used for cross-validation of the Guttman and Biserial weights. The dependent variable employed in Table 2 was again the split-half reliability coefficient raised by the Spearman-Brown formula to estimate the reliability of the full 24- or 30-item test.

Results

As Table 2 shows, a consistent pattern was found in all three parts of Study 2:

1. Guttman and Biserial weights were superior to Number Right and Correction for Guessing weights.
2. Guttman weights were superior to Biserial weights in the calculation sample.
3. Biserial weights were superior to Guttman weights in the cross-validation sample.

Discussion

Upon first examination, it would appear that the reversal of the reliabilities of the Guttman and Biserial weights between the calculation and cross-validation samples is due to the operation of factors analogous to those producing shrinkage in multiple regression. That is, with each iteration more of the error variance is "fit" and the derived weights are less representative of the population. This does seem to be the case with Guttman weights, since with each iteration the calculation sample reliability tended to increase while the cross-validation sample reliability tended to decrease. However, with Biserial

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 2
Results of Study 2--Four Option Weighting Procedures Applied to Three Different Achievement Subtests

Test and Sample	Number Right Scoring	Correction for Guessing Scoring	Guttman Weights Scoring	Biserial Weights Scoring
Biology Knowledge (24 items)				
Calculation Sample				
Split-half reliability	.63	.63	.70	.69
Effective test length	1.00	1.00	1.35	1.28
Cross-Validation Sample				
Split-half reliability	.63	.63	.66	.68
Effective test length	1.00	.97	1.13	1.22
Chemistry Knowledge (30 items)				
Calculation Sample				
Split-half reliability	.68	.68	.75	.74
Effective test length	1.00	1.01	1.40	1.38
Cross-Validation Sample				
Split-half reliability	.71	.70	.74	.74
Effective test length	1.00	.96	1.16	1.19
Physics Knowledge (24 items)				
Calculation Sample				
Split-half reliability	.62	.62	.71	.68
Effective test length	1.00	1.00	1.50	1.35
Cross-Validation Sample				
Split-half reliability	.59	.56	.65	.66
Effective test length	1.00	.90	1.27	1.37

Note. N=800 in both the calculation and cross-validation samples.

weights, the reliabilities in both the calculation and cross-validation samples tended either to increase with each iteration, or to remain stable. The reason for this difference is not immediately evident, even though an examination of successive sets of weights indicates that the important factor may have been the relatively large (high negative) weights assigned to the "omit" category under the Guttman procedure. Regardless of the underlying reason, the empirical results do seem to indicate that Biserial option weights are superior to Guttman option weights.

The positive results of these two somewhat limited studies indicate that a further, more comprehensive study comparing this new option weighting procedure with the Guttman procedure should be undertaken. Such a study should examine several additional aspects of option weighting procedures, such as internal vs. external criteria for the weights, the effect of various methods for limiting the size of the weight assigned to the "omit" category, outcomes with different populations and test content areas, and, perhaps most important, the effect of option weighting on test validity. This last point is particularly important since a number of the previous studies of Guttman weights have found that as the reliability increased, the test validity remained the same or even decreased.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

References

Brogden, H. E. A new coefficient: Application to biserial correlation and to estimation of selection efficiency. *Psychometrika*, 1949, *14*, 169–175.

Clemans, W. V. An index of item-criterion relationship. *Educational and Psychological Measurement*, 1958, *18*, 167–172.

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 each choice. *Educational and Psychological Measurement*, 1959, *19*, 159–170.

Echternacht, G. J. *A comparison of various item option weighting schemes* (Research Bulletin 73–6). Princeton, N.J.: Educational Testing Service, 1973.

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. (a)

Hendrickson, G. F. *The effect of differential option weighting on multiple-choice objective tests* (Report No. 93). Baltimore: The Johns Hopkins University, Center for the Study of Social Organization of Schools, 1971. (b)

Hendrickson, G. F., & Stanley, J. C. *An assessment of the effect of differentially weighting the options of a multiple-choice objective test.* Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, March 6, 1970.

Reilly, R. R. *Empirical option weighting with a correction for guessing* (Research Bulletin 72–59). Princeton, NJ: Educational Testing Service, 1972.

Reilly, R. R., & Jackson, R. *Effects of empirical option weighting on reliability and validity of the GRE* (Research Bulletin 72–38). Princeton, NJ: Educational Testing Service, 1972.

Stanley, J. C., & Wang, M. D. *Weighting test items and test-item options, an overview of the analytical and empirical literature.* Paper presented at the Mathematical Psychological Meetings at Stanford University, Palo Alto, August 28, 1968. (a)

Stanley, J. C., & Wang, M. D. *Differential weighting, a survey of methods and empirical studies.* Baltimore: The Johns Hopkins University, 1968. (b)

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.

Wang, M. D., & Stanley, J. C. Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 1970, *40*, 663–705.

Acknowledgements

This research was carried out as part of a project supported by a contract with the Association of American Medical Colleges, Washington, D.C. The opinions expressed are those of the author and do not necessarily reflect those of the Association of American Medical Colleges.

Author's Address

John G. Claudy, American Institutes for Research, Post Office Box 1113, Palo Alto, CA 94302.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/