Item-Option Weighting of Achievement Tests: Comparative Study of Methods

Ronald G. Downey
Temple University

Previous research has studied the effects of different methods of item-option weighting on the reliability and the concurrent and predictive validity of achievement tests. Generally, increases in reliability are found, but with mixed results for validity. This research attempted to interrelate several methods of producing option weights (i.e., Guttman internal and external weights and judges' weights) and examined their effects on reliability and on concurrent, predictive, and face validity. Option weights to maximize reliability produced cross-validated (N = 974) increases in Hoyt reliability over rights-only scoring (.82 versus .58, respectively); decreases in correlations with other achievement tests; few changes in predictive validity; and a loss in face validity (i.e., some correct options had lower weights than incorrect options). Weights to maximize validity did not cross-validate and led to a reduction in reliability and to mixed validity results. Judges' weights produced increases in reliability and mixed results with validity. The size of Guttman weights were shown to interact with item-option and test characteristics. It was concluded that option weighting offered limited, if any, improvement over unit weighting.

Current scoring systems for multiple-choice achievement test items are based on assumptions about the nature of the individual's response(s) to an item. The major assumptions of "all-or-none" knowledge, random incorrect responses, and equal-option distractability have frequently been criticized (Cureton, 1966; Davis, 1967; Lord, 1963; Stanley, 1954; Wiley, 1960).

Earlier research efforts were directed at the development of methods, which were not based upon the above assumptions about the nature of responses (see Stanley & Wang, 1970; Wang & Stanley, 1971) for differential "item-option weighting" for achievement tests. Nedelsky (1959) conducted one of the earliest studies in this area and found that a test utilizing a worst-distractor weighting procedure was more reliable than a rights-only score. Lord (1965) also found partial support for the worst-distractor procedure. Davis and Fifer (1959) used three different option-weighting procedures: the correlations between the item option and total test scores as weights, judges' weights, and weights as suggested by Flanagan (1935). Their findings indicated that the use of option weights generally increased reliability, but not validity as defined by predicting teacher ratings (see also Davis, 1959). Sabers and White (1969) also found results similar to those of Davis and Fifer.

Recently, a number of studies have been conducted using either a variant of a method originally suggested by Guttman (1941) or an elaboration of the Davis and Fifer (1959) method of judges' weights. Hendrickson (1971) conducted a study with the Scholastic Aptitude Test using the weighting method suggested by Gutt-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 4 Fall 1979 pp. 453–461

man and found substantial increases in reliability and lower intercorrelations of the verbal and quantitative subtests. Reilly and Jackson (1973) and Reilly (1975) used similar procedures with the Graduate Record Examination and again found increases in reliability, a tendency for lower intercorrelations between subtests, and lower validity coefficients with undergraduate grade-point averages (backward prediction). Reilly (1975) presented some evidence that weighting of omitted items produced undesirable results. Waters (1976), using empirical weighting procedures similar to Davis and Fifer's (1959), also found increased reliability and decreased intercorrelations with other measures. Hendrickson (1971) has suggested that these results can be explained if it is assumed that the weighted test is more factorially pure, which would lead to increased reliability and less overlap with other measures.

Hambleton, Roberts, and Traub (1970), Patnaik and Traub (1973), and Kansup and Hakstian (1975) all used a variant of option weighting in which weights were derived by expert judges, obtaining similar results (viz., increased internal reliability, but mixed results for predictive validity). Kansup and Hakstian (1975) have made a strong appeal for dropping research on item-option weighting due to the inability to prove its value and the preponderance of evidence against it.

While the above studies of item-option weighting have generally found moderate to substantial increases in reliability, the question of changes in validity has been less clear. Most studies have found that correlations with other similar achievement tests have decreased, which would follow from the concept that the test is becoming more factorially "pure" (see Hendrickson, 1971). Hendrickson referred to this as quasi-validity. There is a need to produce more evidence regarding both concurrent (quasi-) and predictive validity, as well as to compare the two separate lines of research using Guttman and judges' weights.

The present study was designed to investigate the comparative effects of item option weighting procedures on reliability and on concurrent (quasi-), predictive, and face validity. Three different methods of option weighting were used. The method of "reciprocal averages" (Baker & Hoyt, 1972; Lawshe & Harris, 1958) was used to derive Guttman weights for maximizing reliability (internal consistency), and Guttman weights were used to maximize validity. In addition, judges' option weights were developed. These were compared with the conventional rights-only scoring.

Method

Subjects

The sample was composed of 1,950 entering freshman college students at Temple University. The total sample was randomly split into two groups of approximately equal size (976 in the experimental group and 974 in the cross-validation group). All empirical weights were derived with the experimental group, and comparisons were made on the results from the cross-validation group. Due to placement into different courses, some individuals did not have criterion scores (see Table 2). The placement was based upon the test described below.

Procedures

The test used was the English Expression portion of the Cooperative English Test (*Cooperative English Tests*, 1960). Only the Effectiveness section was used, a 30-item test on the ability to determine intended meaning. The concurrent validity measures used were verbal and quantitative scores of the Scholastic Aptitude Test. The English grades for the first two semesters of college English were used as the measures for predictive validity.

Weighting Schemes

In addition to the conventional weights, three other types of weighting schemes were used in this study. Two of the schemes were based on the method proposed by Guttman (1941); the third

was based on the assumption that experts can assign meaningful weights to options, based on the amount of correct (or incorrect) information contained in the option.

Reciprocal averages weights. Guttman proposed a method which weighted the option (or category) by using the mean criterion score of the individuals selecting that option. Guttman assumed that the value he wanted to achieve was one which minimized individual variability over a group of subjects. This minimization was accomplished by maximizing a correlation coefficient represented by the ratio between the variance among subjects and the total variance. Guttman proceeded to show that the set of weights satisfying this requirement are proportional to the mean score of the individuals selecting an option (cf. Guttman, 1941, p. 341). Weights derived from the reciprocal averages procedure were only approximations of the final Guttman weights. Therefore, the procedure was iterated several times using the derived weights to rescore the test, recalculating new weights until the weights were stabilized. Using Lord's (1958) nomenclature, the sources of variance in a test can be described as follows: Let X_{ci} be the scoring weight of option c for item i (m = number of items) and N equal the number of subjects. Let $y_{i_c s}$ be the score obtained by an individual on item i for option c at iteration s, so that $y_{i_c s} = X_{i_c s}$ whenever a person chooses option c. Therefore,

$$y_{a.} = \sum_{i=1}^{m} y_{i_c s} \qquad [1]$$

equals the total score of person a,

$$y_{.1} = \sum_{a=1}^{n} y_{i_c s} \qquad [2]$$

equals the total score of the item and

$$y_{..} = \sum\sum_{ia} y_{i_c s} \qquad [3]$$

equals the grand total. The item-person matrix and the Analysis of Variance table (Table 1) will help explain these sources of variance.

Guttman defined MS_i as equal to zero and achieved this by a priori setting all item sums, $Y_{.i}$, to zero. The solution is to maximize the correlation η_x^2, where

$$\eta_x^2 = MS_P/T \ . \qquad [4]$$

If $MS_P + E$ is substituted for T and the equation reduced, then

$$\eta_x^2 = \frac{1}{1 + E/MS_P} \ . \qquad [5]$$

This formula is equivalent to maximizing the between-person variance and minimizing the error term. This solution is also equivalent to maximizing the Hoyt (1951) internal consistency reliability, which is found by the following formula:

$$r_{tt} = 1 - E/MS_P \ , \qquad [6]$$

where r_{tt} equals reliability. The ratio E/MS_P is common to both solutions. More recent efforts (Bock, 1960; Nishisato, 1976, 1979) have used the term optional scaling to apply to a generalized approach and to provide a more detailed treatment. The above procedure is a special case of Nishisato's (1979) principle of internal consistency.

The procedure used to develop weights was as follows: if $X_{i_c s}$ equals the iterated weights, then

$$\hat{x}_{i_c s} = [\Sigma y_{k(s-1)} - y_{i_c (s-1)} \ LB_{i_c}] / LB_{i_c} \qquad [7]$$

where Y_{ks} (1 by N) is a vector of scores $Y_{a.s}$ for all N examinees,
B_{ic} (N by 1) is a vector with elements 1 if person a chose option c to item i and elements 0 otherwise, and
L (1 by N) is a vector of 1's.

Table 1
Item by Person Matrix and Analysis of Variance of Item by Person Data Matrix

Persons	Items I_1	I_2	I_3	I_i	...	I_m	Sums
P_1	y_{11}	y_{12}	y_{13}	y_{1i}	...	y_{1m}	$y_{1.}$
P_2	y_{21}	y_{22}	y_{23}	y_{2i}	...	y_{2m}	$y_{2.}$
P_a	y_{a1}	y_{a2}	y_{a3}	y_{ai}	...	y_{am}	$y_{a.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
P_N	y_{N1}	y_{N2}	y_{N3}	y_{Ni}	...	y_{Nm}	$y_{N.}$
Sums	$y_{.1}$	$y_{.2}$	$y_{.3}$	$y_{.i}$	...	$y_{.m}$	$y_{..}$

Analysis of Variance

Source	Sum of Squares
Between person (MS_p)	$\frac{1}{m}\sum y_{a.}^2 - \frac{1}{Nm} y_{..}^2$
Between items (MS_i)	$\frac{1}{N}\sum y_{.i}^2 - \frac{1}{Nm} y_{..}^2$
Error (E)	$T - MS_p - MS_i$
Total (T)	$\sum\sum y_{ai}^2 - \frac{1}{Nm} y_{..}^2$

The subtraction of $Y_{ic}(s-1)$ then removed the bias for the item being used. The Guttman procedure to maximize reliability began (Iteration 1) with $Y_{a.s}$ equal to the total conventional score. Each remaining iteration (Iterations 2 through 9) used the following equation to develop a new set of weights:

$$(y_{a.s}) = [B\, x_{(s)}^{T}]\, J \,. \qquad [8]$$

where $X_{(s)}$ is defined as an $I \times K$ matrix of the option weights $X_{i_c s}$ (K is the number of options and I is the number of items),

B (N by K) is the matrix formed from the K column vectors B_{i_c} (see above), and

J (K by 1) is a vector of 1's.

All groups were iterated nine times and weights from the ninth iteration were used for cross-validation.

Validity weights. Guttman did not restrict his method to values determined by internal weighting; as Stanley and Wang (1970) and Nishisato (1979) have suggested, other scores could be used to develop weights. The second Guttman weighting procedure used English 001 grades as the score ($y_{a.s}$) and produced a set of

weights maximizing the differences between subjects receiving different grades. Only one iteration was performed for this procedure. Both Guttman weighting methods treat omitted items as valid options, and therefore weights were derived for them. Preliminary results indicated that the scores were positively skewed, and therefore all weighted Guttman total test scores were normalized

Judges' weights. The third weighting procedure was one suggested by Davis and Fifer (1959). Weights applied were determined by having English teachers rate the various options as to the amount of correct and/or incorrect information displayed by a person choosing this option. Omitted items were scored as zero. Seven instructors in the English department were asked to rate the options. The directions given them for making their judgments were:

> It is generally agreed that when multiple-choice examinations are used, options for a particular question vary in their degree of correctness. You are being asked to rate options on the English Expression portion of the Cooperative English Test as to their degree of correctness. Due to the length of the task, only the first part (30 items, Effectiveness) of the Expression portion will be treated in this manner. This means that since you are rating *each* option (4) of *every* question (30) there will be 120 ratings.
>
> For each option you should rate it in terms of its degree of correctness along the following scale of 1 to 7; Mark a (1) if the option is incorrect; mark a (2) or (3) if the options are partially incorrect; mark a (4) if the option is partially incorrect and partially correct; mark a (5) or (6) if the option is partially correct; and mark a (7) if the option is correct. *In rating the options, you should determine the amount of correct and/or incorrect information a respondent would have to have available in order to mark the option as the right answer.*

The weights applied were the mean of weights assigned by the seven instructors. As a measure of the degree of interjudge agreement on item option weights, Kendall coefficients of concordance were computed for each of the 30 items. Out of 30 coefficients, 29 were significantly different from zero at $p < .01$ and the remaining coefficient was not significantly different from zero. The average (mean) coefficient of concordance was .741, and the average (over items) rank-order correlation between pairs of judges for an item was .698. Since the pooled judgments were used, the coefficients of concordance represent the reliability of the weights; the coefficients were generally moderate to high in value. The Effectiveness test was then scored using these weights.

Analysis

As a check against biased selection procedures, *t*-tests were made on differences between variables for the experimental and cross-validation samples. Hoyt (1951) reliability estimates were derived. Estimates of the predictive validity were the zero-order correlation coefficients between test scores for each type of weighting procedure and English grades. Concurrent validity data were the zero-order correlations between SAT-V (and Q) and test scores for each procedure. Since only comparative results between methods, and not the level of prediction, was of major concern, adjustments for restrictions in range on the English grades were not made. All comparisons between methods were made on the cross-validation sample.

Results

Table 2 presents the means, standard deviations, and number of subjects for the four criterion scores and for the conventional test score for the experimental and cross-validation samples. The *t*-tests between the two samples for each of the variables, also presented in Table 2, did not show any significant differences.

Table 2
Comparisons of Experimental and Cross-Validation Groups: Summary Statistics and t-tests

Variable	Statistic	Group: Experimental	Group: Cross-Validation	t-test
SAT-Verbal				
	$\bar{X}$	525.26	528.82	.765
	S.D.	81.40	80.11	
	N	844[a]	844[a]	
SAT-Quantitative				
	$\bar{X}$	540.38	540.98	.159
	S.D.	78.05	76.63	
	N	844[a]	844[a]	
Grades-1st Semester				
	$\bar{X}$	3.34	3.36	.455
	S.D.	.85	.88	
	N	744[b]	738[b]	
Grades-2nd Semester				
	$\bar{X}$	3.42	3.43	.089
	S.D.	.97	.96	
	N	611[c]	575[c]	
Test Score Conventional				
	$\bar{X}$	20.08	20.07	.090
	S.D.	3.73	3.62	
	N	976	974	

[a]Lower N is due to missing data.
[b]Lower N is due to placement procedures.
[c]Lower N is due to placement procedures and drop outs.

Table 3 summarizes the reliability and validity coefficients for the experimental group for each of the four weighting methods. Table 4 summarizes the results for the cross-validation sample. While reliability and validity are separate concepts, they have been found to interact and they will therefore be discussed jointly (see Lord & Novick, 1968, and Tucker, 1946, for a discussion of "the attenuation paradox"). A further complexity was the face validity of the weights for the correct option. If the procedures produced the highest weight for the correct option, then the item (and extended over items to the test) was considered to have face validity.

Using only the conventional procedure as the comparative baseline, the Guttman internal weighting procedure produced test scores (see Table 4) which were more reliable (.82), tended to have lower correlation with the SAT scores, but had only a moderate to negligible effect upon prediction of English grades. Out of 30 items, 21 received the highest item-option weight (positive) for the correct option.

For the weights derived to maximize predictive validity, it can be seen from Table 4 that reliability was lower (.45 versus .58 for the conventional group); the correlations with SAT were lower; and finally, the predictive validity for first

Table 3
Experimental Group: Summary of Reliability and Validity Coefficients for Each Weighting Method

	Method			
	Conventional	Internal	External	Judges
Reliability[a]	.61	.84	.47	.69
Validity				
SAT-Verbal	.58	.48	.47	.58
SAT-Quantitative	.25	.21	.14	.23
Grades-1st Semester	.18	.14	.40	.18
Grades-2nd Semester	.08	.07	.20	.08

[a]Hoyt Reliabilities

semester grades did not change, but prediction of second semester grades improved. Less than half of the 30 items had the highest weight for the correct option.

Table 4 shows a slightly different pattern for the judges' weights. Judges' weights produced a slightly more reliable test (.66) with little change in the concurrent validity, a moderate increase in the prediction of first-semester English grades, and a moderate decrease for the second semester. It should be noted that the results for judges from the experimental group are independent of the weighting procedure for judges and indicated no changes in predictive validity (see Table 3). The judges produced the highest weight for the correct option for all 30 items.

Discussion and Conclusion

The results from this study are similar to previous findings indicating that an internal weighting procedure can produce a more reliable test with a lower relationship to other similar measures (see Hendrickson, 1971; Reilly & Jackson, 1973; Waters, 1976). But this procedure produced little, if any, improvement in predictive validity and at a much greater administrative cost and lower face validity. The

Table 4
Cross-validation Group: Summary of Reliability and Validity Coefficients for Each Weighting Method

	Method			
	Conventional	Internal	External	Judges
Reliability[a]	.58	.82	.45	.66
Validity				
SAT-Verbal	.62	.48	.48	.62
SAT-Quantitative	.21	.17	.19	.19
Grades-1st Semester	.20	.20	.19	.23
Grades-2nd Semester	.10	.12	.17	.13

[a]Hoyt Reliabilities

weights derived by maximizing validity produced a less reliable test with only a hint that validity would be improved. Weighting for increases in validity had high administrative costs with a loss in face validity. The judges' weights produced the most positive results with moderate increases in reliability and a moderate increase in predictive validity. With the exception that the test would generally have to be scored by computer, the costs for developing the judges' weighting procedure are small.

Several more general points should be made. First, the results provide a particularly vivid illustration of the "attenuation paradox." with increases in reliability not producing increases in validity and increases in validity not being stable and lowering reliability. Second, the empirically derived weights produced undesirable side effects, with incorrect item options having higher weights than correct options. Third, the Guttman procedures for deriving weights had other undesirable side effects including large negative weights, large weights assigned to omissions, and skewed score distributions. Almost all these effects upon the option weights are the direct result of an unanticipated relationship between the option difficulty and the size of the weight. Since the sum of the weights for options in an item was set equal to zero, low-difficulty options will have small weights approaching zero. A corollary of this rule is that difficult options (including omitted items) will tend to have relatively large weights due to the possibility that a highly selected group or individual responded to that option. The findings, therefore, suggest that Guttman option weightings interact with the item and test characteristics.

The results do not support the use of either of the Guttman procedures for option weighting because of the high costs associated with the minimum gains. The judges' weighting procedure showed the most promise for producing a more reliable and valid test. While, as Wang and Stanley (1971) point out, option-weighting techniques for achievement and aptitude tests have been "studied with interest," the evidence to date indicates that option weighting offers only a limited improvement over the conventional method of unit weights for correct options.

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, April 1972.

Bock, R. D. *Methods and applications of optimal scaling* (Research Memorandum No. 25). Chapel Hill: The University of North Carolina, Psychometric Laboratory, 1960.

Cureton, E. E. The correction for guessing. *Journal of Experimental Education*, 1966, *34*, 44–47.

Davis, F. B. Estimation and use of scoring weights for each choice in multiple-choice test items. *Educational and Psychological Measurement*, 1959, *14*, 291–298.

Davis, F. B. A note on the correction for chance success. *Journal of Experimental Education*, 1967, *35*, 42–47.

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, *14*, 159–170.

Cooperative English Tests (1960 Revision). Princeton NJ: Educational Testing Service, Cooperative Test Division, 1960.

Flanagan, J. C. *Factor analysis in the study of personality.* Stanford, CA: Stanford University Press, 1935.

Guttman, L. An outline of the statistical theory of prediction. In P. Horst (Ed.), *The prediction of personal adjustment.* New York: Social Science Research Council, 1941.

Hambleton, R. K., Roberts, D. M., & Traub, R. E. A comparison of the reliability and validity of two methods of assessing partial knowledge on a multiple-choice test. *Journal of Educational Measurement*, 1970, *7*, 75–82.

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

Hoyt, C. J. Test reliability estimated by analysis of variance. In E. F. Lindquist (Ed.), *Principles of educational and psychological measurement.* Washington, DC: American Council on Education, 1951.

Kansup, W., & Hakstian, A. R. A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures. *Journal of Educational Measurement*, 1975, *12*, 219–230.

Lawshe, C. H., & Harris, D. H. The method of reciprocal averages in weighting personnel data. *Educational and Psychological Measurement*, 1958, *18*, 311–336.

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

Lord, F. M. Formula scoring and validity. *Educational and Psychological Measurement*, 1963, *23*, 663–672.

Lord, F. M. *Worst distractor study*. Unpublished manuscript, 1965. (Available from Frederic M. Lord, Educational Testing Service, Princeton, NJ.)

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, MA.: Addison-Wesley, 1968.

Nedelsky, L. Ability to avoid gross error as a measure of achievement. *Educational and Psychological Measurement*, 1959, *19*, 459–472.

Nishisato, S. *Optimal scaling as applied to different forms of data* (Technical Report No. 4). Toronto: The Ontario Institute for Studies in Education, Spring 1976.

Nishisato, S. *An introduction to dual scaling* (Technical Report No. 5). Toronto: The Ontario Institute for Studies in Education, Summer 1979.

Patnaik, D., & Traub, R. E. Differential weighting by judged degree of correctness. *Journal of Educational Measurement*, 1973, *10*, 281–286.

Reilly, R. R. Empirical option weighting with a correction for guessing. *Educational and Psychological Measurement*, 1975, *35*, 613–619.

Reilly, R. R., & Jackson, R. Effects of empirical option weighting on validity and reliability of an academic aptitude test. *Journal of Educational Measurement*, 1973, *10*, 185–194.

Sabers, D. L., & White, G. 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. "Psychological" correction for chance. *Journal of Experimental Education*, 1954, *22*, 297–298.

Stanley, J. C., & Wang, M. E. Weighting test items and test-item options, an overview of the analytical and empirical literature. *Educational and Psychological Measurement*, 1970, *30*, 21–35.

Tucker, L. R. Maximum validity of a test with equivalent items. *Psychometrika*, 1946, *11*, 1–14.

Wang, M. W., & Stanley, J. C. Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 1971, *40*, 663–705.

Waters, B. K. The measurement of partial knowledge; a comparison between two empirical option-weighting methods and rights-only scoring. *The Journal of Educational Research*, 1976, *68*, 256–260.

Wiley, C. F. The three-decision multiple-choice test: A method of increasing the sensitivity of the multiple-choice item. *Psychological Reports*, 1960, *7*, 475–477.

Acknowledgments

This study was conducted as part of a dissertation under the direction of Professor Harold C. Reppert, in partial fulfillment of the requirements for the doctoral degree at Temple University, Philadelphia, PA. Portions of this article were presented at the meeting of the American Psychological Association, Washington, DC, September 1976.

The computer program used to obtain the Guttman solutions was originally developed by J. Hendrickson at Johns Hopkins University; only minor modifications were made in her system.

The author thanks an anonymous reviewer and Dallas Johnson for their help in revising the weighting formuli.

Author's Address

Send requests for reprints or further information to Ronald G. Downey, Center for Student Development, Kansas State University, Manhattan, KS 66506.