DOCUMENT RESUME

ED 209 243		TM 810 250

AUTHOR	Budescu, David V.
TITLE	Differential Weighting of Multiple-Choice Items.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-79-19
PUB DATE	Dec 79
NOTE	24p.
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Multiple Choice Tests; *Predictive Validity; *Scoring Formulas; *Test Items; Test Theory; *Weighted Scores

ABSTRACT

This paper outlines a technique for differentially weighting options of a multiple choice test in a fashion that maximizes the item predictive validity. The rule can be applied with different number of categories and the "optimal" number of categories can be determined by significance tests and/or through the R2 criterion. Our theoretical analysis indicates that more complex scoring rules have: higher item validities, higher item variances, higher score variances, and are also likely to increase the interitem correlations and the test reliability. A plausible explanation for the apparent paradox of lack of improvement in the test validity, based on the relation between interitem correlations and item validities, is offered. (Author)

RESEARCH REPORT

DIFFERENTIAL WEIGHTING OF MULTIPLE-CHOICE ITEMS

David V. Budescu
Research Triangle Institute

Educational Testing Service
Princeton, New Jersey
December 1979

Differential Weighting of Multiple Choice Items

David V. Budescu

Research Triangle Institute

*This work was completed as a part of the activities of the author in the summer program for graduate students at ETS (1979)

**Drs. Isaac Bejar, Tom Donlon and William Angoff have read earlier drafts of this manuscript. Their comments and suggestions are appreciated.

Abstract

This paper outlines a technique for differentially weighting options of a multiple choice test in a fashion that maximizes the item predictive validity. The rule can be applied with different number of categories and the "optimal" number of categories can be determined by significance tests and/or through the R^2 criterion. Our theoretical analysis indicates that more complex scoring rules have: higher item validities, higher item variances, higher score variances, and are also likely to increase the interitem correlations and the test reliability. A plausible explanation for the apparent paradox of lack of improvement in the test validity, based on the relation between interitem correlations and item validities, is offered.

Differential Weighting of Multiple-Choice Items

Background

The question of differential weighting of multiple-choice items has generated a large number of studies in the psychological and educational literature (see Stanley & Wang 1970 and Wang & Stanley 1970 for reviews). The bulk of the literature suggests that assigning different weights to the items does not significantly affect the test characteristics and performance, but the possibility of differentially weighting the options (distracters) of any given item has some attractive aspects. As a result, several studies comparing and evaluating a variety of procedures of Differential Options Weighting (DOW) have been conducted in recent years (e.g. Hendrickson 1971, Ramsay 1968, Reilly & Jackson 197[illegible] Echternacht 1976, Bejar & Weiss 1977, Donlon & Fitzpatrick 1978). These studies suggest that the use of scoring procedures more complex than the regular 0-1 rule, has a beneficial effect on some of the test characteristics. When these weights were applied to real and artifical data, indices of reliability and internal consistency have been improved. With one exception, however (Echternacht 1976), no significant improvement in the predictive validity of the tests has been reported.

This fact is surprising. One would expect that when the information conveyed by each item is more complete and better measured, the predictive validity of both item and test will be increased. In this paper we offer a theoretical analysis of the effects of differential weighting on validity. By validity we refer to the prediction of an external criterion independently measured. By taking this approach we eliminate the item-test regression (often labelled the discriminating power of the item) which is a special case of validity. We will comment on this problem in a separate section. We examine a procedure which has the property of maximizing the item-criterion correlation. Therefore any other nonoptimal DOW, or regular scoring rule, can be evaluated by comparing its prescribed item weights to the optimal weights. Such a rule provides an indication of how well a scoring rule can be expected to improve the prediction of the criterion and provides a meaningful standard of comparison for any other alternative non-optimal procedure. It should be emphasized that optimality here refers to item validity only, and that the rule may have damaging effects (at least theoretically) on other aspects of the items and the test. We will also examine some of the side effects of this technique which will enable us to better assess its overall performance.

Definition of the problem and some notation

Imagine we have a quantitative criterion, X, which we want to predict by a multiple-choice test, Y, containing k items (Y1, Y2... Yk). Without any loss of generality, we can assume that the scores of X are scaled, or grouped, in a finite number of values (C). Therefore any person taking X has a score such that:

$$0 \leq X \leq C \qquad (1)$$

A typical item, Yg, has a options: one correct and (a-1) incorrect. Since not every examinee attempts to answer all items we must define an additional category for omissions. We consider this category to be as important, meaningful and informative as the other a options. If we let r=(a+1), we can represent the responses of all the examinees to a given item in an rxC contingency table. Each row represents one option of the item, each column represents one level of score on the criterion X, and the typical entry in the table, nij, is the number of people with score Xj who selected option i. Following the regular statistical notation we let n.j, ni. and n denote the marginal column, row and total frequencies respectively. At this stage we need to select an index of association to describe the relation between X and Yg as reflected by the contingency table. By direct analogy to the dichotomous scoring rule the multinomial generalizations of the biserial and point biserial correlations suggest themselves as possible candidates. Indeed Donlon & Fitzpatrick (1978) have already proposed using the multiserial correlation (Jaspen 1946) as a generalized discrimination index. For our purposes we prefer the point multiserial coefficient (Das Gupta 1960; Hamdan & Schulman 1975) for several reasons:

(i) Unlike the multiserial, it is a PRE measure (Costner 1965), i.e. R^2_{pms} can be interpreted as the percentage of variance of X accounted for by Yg.

(ii) Unlike the multiserial, its values are bounded, i.e. $-1 \leq R_{pms} \leq 1$.

(iii) Unlike the multiserial, the weights assigned to the different categories of Yg are not determined by any distributional assumption.

(iv) These weights can be selected in a way that maximizes the linear relationship between X and Y_g. These weights (Y_{gi}) are a linear function of the mean criterion score of the examinees selecting option i. In particular, if we let $\bar{X}_i$ be the mean score of the people who selected the i^{th} option (i = 1...r):

$$\bar{X}_i = (\sum_j nij\ X\ j)/ni. \quad , \qquad (2)$$

then the optimal weights are given by (Das Gupta 1960):

$$Y_{gi} = A \bar{X}_i + B \tag{3}$$

If we select A = 1 and B = 0 we can express the point multiserial index in a very convenient form. (Hamdan & Schulman 1975).

$$R^2_{pms} = \frac{[\frac{1}{n} \sum_i ni. \bar{X}_i{}^2 - (\frac{1}{n} \sum_i ni. \bar{X}_i)^2]}{[\frac{1}{n} \sum_j n.j \; X j^2 - (\frac{1}{n} \sum_j n.j \; X j)^2]} \tag{4}$$

This particular weighting has two attractive properties:

(a) As Das Gupta (1960) points out, the squared optimal point multiserial is equal to the square multiserial eta (Wherry & Taylor 1946).

(b) R_{pms} can be expressed as a ratio of two standard deviations (Hamdan & Schulman 1975):

$$R_{pms} = \frac{S(\bar{X}i)}{S(X)} \tag{5}$$

A model for evaluating the effects of DOW

It seems only natural to compare the optmal scoring rule to the regular dichotomous alternative (1 = right, 0 = wrong). This indeed is easily done within the framework of this model. Note that if the number of categories, r, is reduced to 2, then the point multiserial is just the regular point biserial. Furthermore it is well known (Das Gupta 1960) that if r=2, R_{pbs} is invariant to linear transformations of Y_{g1}, and Y_{g2}. In other words the correlation will not be changed if we replace the 0 - 1 weights by the optimal weights. The implication is obvious--one can compare the effectiveness of the two scoring rule by the percentage of variance accounted for, when 2 or r categories are used and, if some distributional assumptions are made, test whether the difference is significant. But note that scoring by 2 or r categories are only end points on a continuum of different optimal scoring rules. We can define a hierarchy of models (all of them optimal) which vary in terms of their complexity and of the number of categories used by the scoring procedure. Consider the following models:

(i) r categories - all r options

(ii) (p+1) = (r-q+1) options - q categories are combined into one while p are left unchanged.

(iii) 3 options - right, wrong, omit.

(iv) 2 options - right, wrong + omit.

Models (i), (iii), and (iv) are natural and well known. We need to say a word about (ii). It defines a class of models in which two or more options are combined on the basis of empirical or theoretical justifications. If one option is selected with very low probability it may be reasonable to combine it with the "omit" option. If there is some natural relation between some of the distracters it may seem natural to combine them according to this characteristic (see Echternacht 1976 for such items), etc. The most important point is that the responses can be scored in a variety of ways, using different number of categories, and for each model optimal weights can be easily derived by the same rule (3). One could compare all these models and select the best one - i.e. the one which predicts the highest proportion of variance in X relative to the number of parameters fitted (the number of categories).

We now examine the effect of combining q categories into one, while keeping the first p unchanged, on the correlation. Define the new category Yc, and also define:

$$nc. = \sum_{i=p+1}^{r} ni. \tag{6}$$

$$\bar{X}c. = (\sum_{i=p+1}^{r} ni.\ \bar{X}i)/nc. \tag{7}$$

These manipulations do not affect the denominator and the second term in the numerator of (4). The first term in the numerator can be rewritten as:

$$\frac{1}{n} \{\sum_{i=1}^{P} ni.\ \bar{X}i^2 + nc.\ \bar{X}c^2] \tag{8}$$

and if we let $R'_{pms(q)}$ be the new point multiserial correlation it can be easily shown that:

$$R^2_{pms} - R^2_{pms(q)} = \frac{\sum\sum_{i\neq j=p+1}^{r} ni.\ nj.\ (\bar{X}i - \bar{X}j)^2}{n\ nc.\ S^2(X)} . \tag{9}$$

If we only combine two categories (say k and l), this is reduced to:

$$R^2_{pms} - R^2_{pms(2)} = \frac{n_{k.}\ n_{1.}\ (\bar{X}_1 - \bar{X}_k)^2}{n\ (n_{k.} + n_{1.})\ S^2(X)} . \tag{10}$$

Eq. (9) is always positive, which implies that if one reduces the number of categories the correlation with the criterion must always decrease. The reduction in percentage of variance accounted for is a monotonically decreasing function of the sample size, the variance of the criterion and the size of the new category; it is a monotonically increasing function of the weighted sum of squared pairwise differences between the means of the q categories combined. These relations suggest that using simpler scoring rules (combining categories) may have only a negligible effect on the item validity when the means of the combined groups are relatvely homogeneous and the sample size and criterion variance are large. On the other hand, if the sample size and variance of X are small and if the means are relatively heterogeneous, the more complex rule can significantly increase the correlation. Finally, for a given criterion (with a fixed variance) administered to a fixed sample (fixed n), the best way to simplify the scoring rule is to combine the categories with the most similar means.

If we are interested in testing hypotheses about R^2_{pms} we must assume that the criterion conditional distribution at the i^th level of $Y(i = 1...r)$ is $N(\mu_i, \sigma^2)$ (Hamdan & Schulman 1975). In this model we can test independence ($\rho_{pms} = 0$) for any scoring rule with s categories ($s = 2... r$), by:

$$F = (n-s)\, R^2_{pms}/(s-1)(1-R^2_{pms}) \qquad (11)$$

This statistic has an F distribution with (s-1) and (n-s) d.f. under the null hypothesis. To test equality of two models with s1 and s2 categories [(Ho: $\rho_{pms(s1)} = \rho_{pms(s2)}$] we can use the statistic:

$$F = \frac{[R^2_{pms(s1)} - R^2_{pms(s2)}](n-s1)}{[1 - R^2_{pms(s1)}](s1-s2)}, \qquad (12)$$

which is distributed as an F with (s1-s2) and (n-s1) d.f. under Ho. Generally, for each item, Yg, a series of tests similar to those performed in a standard regression analysis can be used in order to assess the best scoring rule and its effectiveness (Cramer 1972).

The effect of different scoring rules on other item and test characteristics

(a) Item variance

Once the weights to be attached to the r options are determined we can calculate the item's variance by the regular formula:

$$S_g^2 = [\frac{1}{n}\sum_{i=1}^{r} ni.\ Y_{gi}^2 - (\frac{1}{n}\sum_{i=1}^{r} ni.\ Y_{gi})^2] \quad , \qquad (13)$$

which is just a re-expression of the numerator of (4). Therefore, after combining q categories into one, the reduction in the item's variance is:

$$S_g^2 - S_{g(q)}^2 = \frac{\sum\sum_{i\neq j=p+1}^{r} ni.\ nj.\ (\bar{Y}_{gi} - \bar{Y}_{gj})^2}{n\ nc.} \qquad (14)$$

The sum of squared pairwise differences in (14) is just a linear function of the variance of the means in the combined categories around X̄c. Therefore, (14) indicates that when a simpler scoring rule is employed the variance of the responses in each item is invariably reduced, and this reduction is proportional to the variance of the optimal weights and inversely related to sample size. Minimal reduction in variance for any given item will be obtained when we combine categories with homogeneous means.

(b) Interitem correlation

Consider two arbitrary items in the test, Yg and Yh, scored on all categories. Given their correlations with the criterion, Rxg and Rxh (we drop the pms notation for simplicity), their intercorrelation is restricted by (Glass & Gollins 1970):

$$Rgh = \{Rxg\ Rxh \pm \sqrt{(1-R^2xh)(1-R^2xg)}\} \qquad (15)$$

We consider the effect of combining q categories on this interval. Let 11 and 11(q) denote the lower limits of the interval when the items are scored with r and (p+1) categories, respectively. The difference between these lower bounds is:

$$\begin{aligned} 11 - 11(q) = {} & [Rxg\ Rxh - Rxg(q)\ Rxh(q)] + \\ & [\sqrt{(1-R^2xg(q))(1-R^2xh(q))} - \sqrt{(1-R^2xh)(1-R^2xg)}] \ . \end{aligned} \qquad (16)$$

Since it was shown that $R^2_{xs} > R^2_{xs}(q)$ (s = 1... r), and consequently that $(1-R_{xs})<(R^2_{xs}(q))$, it follows that eq. (16) is always positive--combining categories reduces the lower bound for inter-item correlation.

Let 1c and 1c(q) represent the length of the interval, or in other words the range of values that Rgh can take, when r or (p+1) categories are used. It can be [illegible] shown that:

$$1c(q) - 1c = 2 [\sqrt{(1-R^2xg(q))(1-R^2xh(q))} - \sqrt{(1-R^2xh)(1-R^2xg)}] \ . \qquad (17)$$

The range of possible values of Rgh is increased or, in other words, the restrictions imposed on the internal relations between items through their correlations with the external criterion X are relaxed. The lower bound and the length of the interval determine its upper limit (ul).

Combining the information from (16) and (17) it follows that:

$$Ul(q) \geq Ul \quad \text{if} \quad [lc(q) - lc] \geq [ll - ll(q)] \ . \tag{18}$$

It becomes clear that the upper limit of the interval can increase, decrease or remain unchanged depending on the nature and magnitude of the changes in the item-criterion correlations. This is a particularly interesting result because it demonstrates one possible explanation for the lack of improvement in validity of a test. Although DOW improves the individual items validities, it can also simultaneously increase the interitem correlations and the overal system validity can remain practically unchanged.

If we assume that the values of Rgh are symmetrically distributed within the interval, its expected value is at the central point (see Mulaik 1976 for an elaborate proof for the special case Rxh = Rxg = R). If we use an r categories scoring system:

$$E\ [Rgh|Rxb;\ Rxg] = Rxh\ Rxg\ , \tag{19}$$

and for (r-q+1) categories:

$$E\ [Rgh(q)|Rxh(q);\ Rxg(q)] = Rxh(q)\ Rxg(q)\ . \tag{20}$$

Therefore we can write

$$E\ [Rgh|Rxh;\ Rxg - Rgh(q)|Rxh(q);\ Rxg(q)] = [Rxh\ Rxg - Rxh(q)\ Rxg(q)]\ . \tag{21}$$

The expected value of the correlation between a pair of items decreases after combining q categories. Consider the explanation for the lack of improvement in validity offered in the previous paragraph. The last result demonstrates that such a situation is not only possible but also very likely to occur.

(c) Variance of scores

The score of each individual on scale Y is defined as the sum of the scores on the k items composing the scale. Therefore the variance, $S^2(Y)$, is given by:

$$S^2(Y) = \sum_{i=1}^{k} Si^2 + \sum\sum_{i \neq j=1}^{k} Sij \qquad (22)$$

where Sij is the covariance of items i and j. We have already shown that simplified scoring rules have the effect of invariably reducing the item variances, standard deviations and the lower bounds of their intercorrelation and, conditional upon the symmetric distribution of Rij given Rix and Rjx, the expected interitem correlations. These facts, combined together, indicate that the variance of the Y scores is very likely to decrease when categories are combined. In fact, a sufficient condition for this to happen is that $Rij(q) \leq Rij$ $(i,j = 1\ldots k)$.

A special case, which will be discussed later, is one where the combination of the categories has an uniform effect on all the items, i.e. each item variance is reduced by the same proportion. Note that this does not imply that the item variances are equal when r or p+1 categories are used, but rather indicates the fact that there is a functional relation between the number of categories and the item variances and that combining q categoris has a relatively homogeneous effect on all q variances. Formally, let $S_i^2(q) = d^2 S_i^2$ $(i = 1\ldots k, o \leq d \leq 1)$, and in this case:

$$S^2(Y) - S^2(Y)(q) = (1-d) \left[\sum_{i=1}^{k} Si^2 + \sum_{i\neq j=1}^{k}\sum Si\ Sj\ (Rij - Rij(q))\right] . \qquad (23)$$

(d) Reliability

A popular method of calculating reliability is to obtain the ratio of the mean interitem covariance and the mean item variances, Ryy, and to use it as an estimator of the reliability of a single item in the Spearman-Brown prophecy formula (Stanley 1971). If the score is based on r categories:

$$Ryy = \frac{k \sum_{i\neq j=1}^{k}\sum Si\ Sj\ Rij}{(k-1)^2 \sum_{l=1}^{k} S^2 l} , \qquad (24)$$

and if only (p+1) are used:

$$Ryy(q) = \frac{k \sum_{i\neq j=1}^{k}\sum Si(q)\ Sj(q)\ Rij(q)}{(k-1)^2 \sum_{l=1}^{k} S^2 l(q)} . \qquad (25)$$

The difference between the two estimates can be written as:

$$Ryy - Ryy(q) = \frac{k \left[\sum_{l=1}^{k} \sum_{i\neq j=1}^{k}\sum S^2 l(q) SiSjRij - S^2 lSi(q)Sj(q)Rij(q)\right]}{(k-1)^2 \left[\sum_{l=1}^{k} S^2 l \sum_{l=1}^{k} S^2 l(q)\right]} . \qquad (26)$$

It appears that the effect of the scoring rule on the reliabilities depends on the pattern of variances, covariances and their respective reductions. To simplify formula (25) we assume that when categories are combined the variance of each item is reduced by an amount proportional to its initial magnitude, i.e. Si(q) = d Si (i = 1...k, o ≤ d ≤ 1). In this case:

$$Ryy - Ryy(q) \doteq \frac{k \left[\sum_{1}^{k} \sum_{i \neq j}^{k} \sum S^2 l \; Si \, Sj \; (Rij - Rij(q))\right]}{(k-1)^2 \quad \left[\sum_{l=1}^{k} S^2 l\right]^2} \tag{27}$$

The amount of reduction in the test reliability is independent of the constant d, and it is proportional to the weighted sum of reductions in item intercorrelations which were discussed in a previous section. The direct relation between the reliability of a test and the mean item intercorrelation was demonstrated empirically in a recent paper by Bejar and Weiss (1977).

(e) Test validity

We now combine some of the results from the previous sections in order to examine the behavior of the validity of Y (Rxy). Gulliksen (1950 p. 382) gives the formula for the total test validity as a function of the item validities and the test variance:

$$Rxy = \left[\sum_{g=1}^{k} Rxg \; Sg\right]/S(Y) \; . \tag{28}$$

After combining q categories the validity becomes:

$$Rxy(q) = \left[\sum_{g=1}^{k} Rxg(q) \; Sg(q)\right]/S(Y)(q) \; . \tag{29}$$

and the reduction in the percentage of variance of the criterion explained by the predictor is:

$$R^2xy - R^2xy(q) = \frac{[S^2(Y)(q) \sum_g^k S^2g\, R^2xg - S^2(Y) \sum_g^k S^2g(q)\, R^2xg(q)]}{[S^2(Y)\, S^2(Y)(q)]} \qquad (30)$$

Using again the assumption of uniform reduction is variance across items ($Si(q) = dS_i$) we can rewrite the last equation as a function of variances and correlations:

$$R^2xy - R^2xy(q) = \frac{[\sum_g^k \sum_l^k S^2l\, S^2g(R^2xg - R^2xg(q)) + \sum_g^k \sum\sum_{i\neq j}^k S^2g\, SiSj(R^2xg\, Rij(q) - R^2xg(q)\, Rij)]}{[\sum_l^k S^2l + \sum\sum_{i\neq j}^k SiSj\, Rij][\sum_l^k S^2l + \sum\sum_{i\neq j}^k SiSj\, Rij(q)]} \qquad (31)$$

Note that the second term in the numerator involves the item-test as well as the interitem correlations. It is therefore very difficult to evaluate the impact of the new scoring rule on the validity. While the first term in the numerator is always positive, the second can also assume negative values. In fact, if we assume that all correlations with the criterion are reduced by an amount proportional to their initial value ($Rij(q) = d^1 Rij$, $i\neq j = 1\ldots k$, $0 \leq d^1 \leq 1$), the second term vanishes. Equation (31) provides further support to the explanation offered in the previous section to the lack of improvement in validity. It is clear that the overall improvement in validity depends on the effect of the scoring procedure on both the item correlations and interitem correlations. We can expect a significant gain in the percentage of variance predicted in tests in which we can significantly improve the item validities and reduce the interitem correlations (or at least not increase them). This is more likely to happen if the initial item validities are low.

Final Remarks

In the introduction we have emphasized that the function being optimized is the item-criterion correlation, and that an external and independently measured criterion is necessary. We are not aware of any empirical or theoretical study in which the procedure examined here was used, although French (1952) has pointed out some of its desirable properties. However, several studies (e.g. Hendrickson 1971, Echternacht 1976) have used a similar technique. The main difference between their approach and the present one is that, instead of an external criterion, they use the score on the remaining (k-1) items of the test and therefore, instead of optimizing external validity, they optimize internal consistency. A problem in this approach is that the two variables being correlated are not experimentally independent--the weights for item Yg depend on the scores on the other (k-1) items, and these scores depend on the optimal weights. One solution to this problem is to use an iterative procedure in which the weights and the criterion are recalculated until the increase in reliability does not exceed a fixed prespecified value. Typically the convergence was found to be very quick and the improvement in reliability only marginal. What are the implications of these findings to the procedure outlined here? It is hard to judge but there are good reasons to believe that using an external criterion to determine the weights should yield better results. In the iterative procedure the initial weights are either (0,1) or $(-\frac{1}{(a-1)}, 1)$. Note that these are the most non-optimal weights, since it was pointed out that the improvement in item validity is proportional to squared differences between the means. The internal consistency procedure is likely to improve its performance if different starting values are used. Possible candidates for this role seem to be (a) option-test point biserial or biserial correlations, (b) theoretically determined a priori weights, or (c) weights proportional to the means calculated from a second independent sample. Empirical work comparing these different starting points for the iterative algorithm and the procedure outlined above is needed.

We have outlined a technique for differentially weighting options of a multiple choice test in a fashion that maximizes the item predictive validity. The rule can be applied with different number of categories and the "optimal" number of categories can be determined by significance tests and/or through the R^2 criterion. Our theoretical analysis indicates that more complex scoring rules have: higher item validities, higher item variances, higher score variances, and are also likely to increase the inter-item correlations and the test reliability. A plausible explanation for the apparent paradox of lack of improvement in the test validity, based on the relation between interitem correlations and item validities, was offered.

The mechanism suggested as the cause of this phenomenon was developed within the framework of the particular optimization procedure examined in this study. Yet, similar explanations could be offered for other DOW procedures since all of them are developed at the item level and do not account for the interitem relations.

Overall, it appears that the key to the success of any DOW procedure is in the nature of the test's items. A scoring rule is likely to be successful if a test contains items with distracters which can differentiate between various levels of partial information, i.e. distracters that have differential appeal for different ability levels. If the distracters are relatively homogeneous this procedure (or any other DOW technique) is not likely to be successful. Therefore we speculate that DOW have a higher probability of success in achievement and criterion referenced tests, and in tests in which the distracters are systematically designed to reflect different levels of partial information. (e.g. Echternacht 1976). More theoretical and empirical work on this question is necessary.

References

Bejar, I.I. and Weiss, D.J. A comparison of empirical differential option weighting scoring procedures as a function of inter-item correlaton. Educational and Psychological Measurement, 1977, 37, 335-340.

Costner, H.L. Criteria for measures of association. American Sociological Review, 1965, 30, 341-352.

Cramer, E.M. Significance tests and tests of models in multiple regression. The American Statistician, 1972, 26, 26-30.

Das Gupta, S. Point biserial correlation coefficient and its generalization. Psychometrika, 1960, 25, 393-408.

Donlon, T.F. and Fitzpatrick, A.R. The statistical structure of multiple choice items. Paper presented at the NERA meeting in Ellenville, N.Y., 1978.

Echternacht, G. Reliability and validity of item option weighting schemes. Educational and Psychological Measurement, 1976, 36, 301-309.

French, J.W. A technique for criterion-keying and selecting test items. Psychometrika, 1952, 17, 101-106.

Glass, G.V. and Collins, J.R. Geometric proof of the restriction on the possible values of r_{xy} when r_{xz} and r_{yz} are fixed. Educational and Psychological Measurement, 1970, 30, 37-39.

Gulliksen, H. Theory of Mental Tests. N.Y., Wiley, 1950.

Hamdan, M.A. and Schulman, R.S. On point multiserial correlation. Australian Journal of Statistics, 1975, 17, 84-86.

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

Jaspen, W. Serial correlation. Psychometrika, 1946, 11, 23-30.

Mulaik, S.A. Comments on "The measurement of factorial indeterminacy", Psychometrika, 1976, 41, 249-262.

Ramsay, J.O. A scoring system for multiple choice test items. The British Journal of Mathematical and Statistical Psychology, 1968, 21, 247-250.

Reilly, R.R. and Jackson, R. Effects of empirical option weighting on reliability and validity of the GRE. RB 72-38. Princeton, N.J.: ETS 1972.

Stanley, J.C. Reliability. In R.L. Thorndike (Ed.) Educational Measurement, Washington, D.C.: American Council on Education. 1971.

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

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

Wherry, R.J. and Taylor, E.K. The relation of multiserial eta to other measures of correlation. Psychometrika, 1946, 11, 155-162.