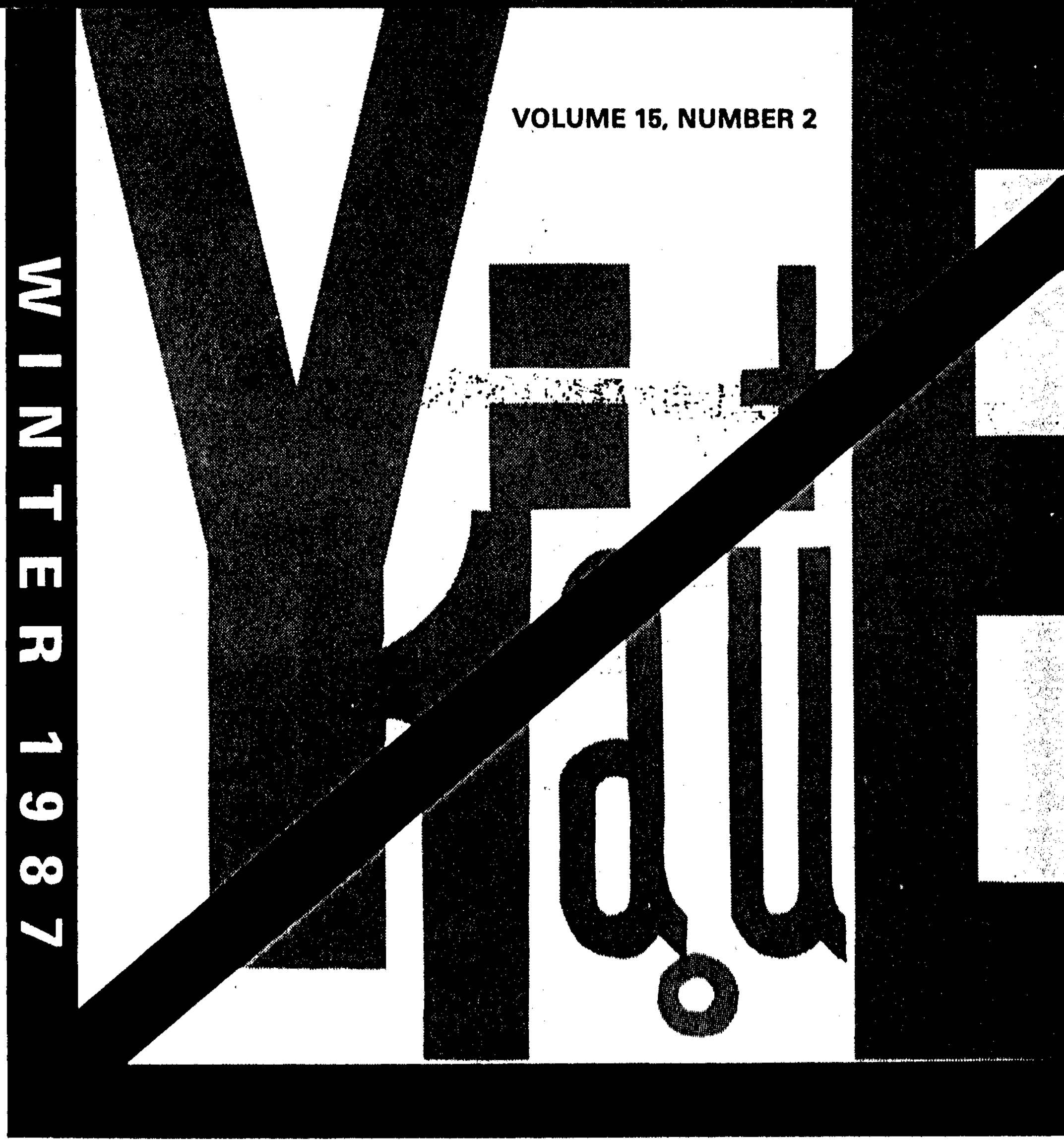



# MULTIPLE LINEAR REGRESSION VIEWPOINTS

A publication of the Special Interest Group on Multiple Linear Regression of the American Educational Research Association

MLRV Abstracts appear in
microform and are available from
University Microfilms International
MLRV is listed in EBSCO Librarians Handbook.

**ISSN 0195-7171**

Library of Congress Catalog Card #80-648729

# MULTIPLE LINEAR REGRESSION VIEWPOINTS

## TABLE OF CONTENTS

MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 15, NUMBER 2, WINTER 1987

# Using Diagnostics for Identification of Biased Test Items

Donald T. Searls
University of Northern Colorado
Edgar Ortiz
Citicorp

## ABSTRACT

This paper demonstrates how recent developments in the analysis of regression models may prove useful in the identification of atypical and potentially biased test items. Regression diagnostics studied are based on analysis of the sensitivity of leverage points, studentized residuals, and ratios of covariances due to the sequential deletion of each test item from the analysis. These procedures appear to offer a substantial refinement over existing approaches.

## IDENTIFICATION OF INFLUENTIAL ITEMS : THEORETICAL RATIONALE

Many statistical procedures have been proposed for detecting biased items. Although they differ in their conceptualization of bias, they nevertheless exhibit a commonality in their purpuse which is to identify those items which hamper the performance of one group relative to another.

Irrespective of the approach, the proposed statistical procedures for identifying biased items rely directly or indirectly on variants of the concept of statistical distance. A major limitation with all of these approaches is that no distribution theory is available to determine objectively when one atypical score is statistically different from others. This shortcoming is particularly evident in Angoff's delta-plot method and extensions of this procedure (Angoff and Ford, 1973. Rudner, et al., 1980).

A lack of distribution theory is also evident in the chi-square methods of Scheuneman (1979) and Camilli (1979). These procedures aim at detecting biased items by performing tests of randomness on the distribution of responses into ability intervals. However, setting of cut-off levels to establish the various ability intervals is done after examining the data. Such a posteriori determination of cutoff points to define ability intervals in effect violates the assumption of random assignment, since factors other than chance are influencing the results. Consequently, rather than detecting biased items, results so derived may identify instead an item's sensitivity to clustering into the ex post facto determined ability classes.

Statistical procedures for detecting biased items based on latent trait models have also been proposed. (Lord and Novick, 1968; Hambleton and Cook, 1977). In these methods, item characteristic curves are fitted to the observed performance scores of different groups. If the fitted curves are not the same for the groups being compared, the item is said to be biased. A major shortcoming of this approach is the lack of specification of the underlying theoretical distributon of the observed delta-values that characterize the differences in performance between the groups being compared. Although some progress has been reported (Lord, 1977), the validity of tests of significance to identify biased items based on the assumptions of latent trait models is as yet an issue that remains unresolved (Lord, 1977; p. 25). A comparative analysis of the performance of latent trait models to identify biased items (Rudner, et al. 1980), does not deal with the subject of statistical significance of the various indices of bias reported in that study.

A comprehensive review of the various statistical techniques proposed for detecting item bias is given in Petersen (1977), Merz (1978) and Sheppard et al. (1980). Statistical analyses, however, do not detect biased items. They only identify those items in which the achievement scores of the groups being compared deviate from the pattern established by other items that make up a test. These items, in turn, may reveal specific content characteristics that either increase or decrease the a priori probability of a correct response in one group of examinees but not in the other.

The statistical procedures to be exemplified in this investigation offer an objective set of statistical criteria to examine individual items for potential bias. These methods are based on generalizations of regression models as developed by Belsley, Kuh and Welsch (1980). The identification of potentially biased items, based on regression diagnostics offers a substantial refinement over existing approaches in that :

a) Distribution theory is used to determine cutoff levels and identify atypical items objectively.

b) Statistical methods are available that measure the sensitivity of parameter estimates to perturbartions in the data, e.g. the effects of the deletion of each item on the estimates of the regression coefficients.

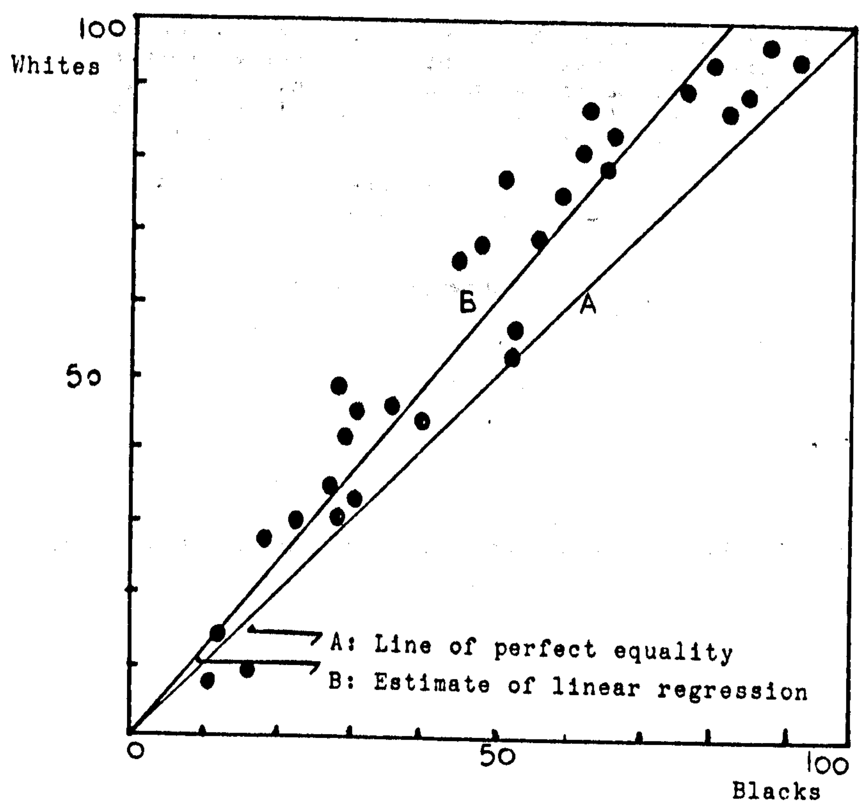
c) These methods offer measures of statistical distance independent of sample size.

Analysis of data based on these procedures can yield important information concerning atypical items which cannot be readily obtained by means of delta-plot, chi-square and latent trait models.

The data to be analyzed comprise the proportion of white and black students who attempted and responded correctly (p-values) to an assessment booklet consisting of 30 items. A scatter plot of the p-values is given in figure 1. Points on line A correspond to items in which the performance of both groups was equal. Points lying above and below this line correspond to items in which the groups being compared performed differently. Points above this line correspond to items in which the group represented by the vertical axis, performed better than the group represented by the horizontal axis. Similarly, points lying below this line correspond to items in which the group represented by the horizontal

FIGURE 1

PLOT OF ACHIEVEMENT SCORES OF WHITE AND BLACK EXAMINEES



axis performed better than the group represented by the vertical axis.

An estimate of the regression line is given by line B (slope=1.19, p=.0001). From graph 1, the consistent scatter of points above line A indicates that white examinees have performed consistently above the performance level set by black examinees. The dispersion pattern of p-values around this line suggests a strong curvature at both extrema, i.e., in the range of the easiest and most difficult exercises. In order to correct for these bottom and ceiling effects, the the p-values were transformed to logits. The logistic transformation is widely used in the analysis of proportional data. Reexpressing quantal response data in logits provides a straightforward procedure to correct for interaction often found in exercise data in the easy and difficult range.

The techniques to be exemplified in this investigation, aim at identifying potentially biased items, by measuring the sensitivity of regression models to the deletion of individual items from the bulk of the data. These diagnostic methods will be applied to parameter estimates in regression models relating the performance of white and black examinees with p-values transformed into logits. Items whose deletion from the body of the data, cause atypical perturbations on parameter estimates are suspect.

For example, given a simple bivariate regression model, the magnitude of the perturbation on the estimated regression coefficients due to deletion of the ith item, can identify atypical items which warrant further examination for potential bias. This procedure is akin to estimating N regression models, where each model corresponds to the 'not i observation'. Within the context of our investigation, items whose deletion cause large and atypical perturbations on estimates of the regression parameters are therefore suspect. From a practical viewpoint this procedure is equivalent to a pseudo-experiment in which it is asked, how would white and black examinees have performed if the ith item had been deleted from the assessment booklet? With these regression diagnostics, items having large deviations from the performance pattern observed in the remaining items can be readily identified.

## RESULTS

### DETECTION OF POTENTIALLY BIASED ITEMS BASED ON REGRESSION DIAGNOSTIC PROCEDURES

The regression diagnostics to be exemplified for use in the detection of potentially biased items are based on analysis of the sensitivity of leverage points, studentized residuals, and ratios of covariances due to the sequential deletion of each item from the model. Two regression models are examined. In model 1, the achievement scores of white examinees are predicted based on the performance of black examinees. Similarly, in model 2, the achievement scores of black examinees are predicted based on the performance of white examinees. The proposed diagnostics attempt to detect biased items by identifying those items that in either model 1 or model 2 elicit performance scores significantly different from the pattern of variability established in the remaining items that make up the achievement booklet. These diagnostic statistics follow from the usual linear model

$$Y = X\beta + e \qquad e \sim N(0,\sigma^2) \qquad (1)$$

where Y is a (n x 1) vector of observations on the dependent variable, X is a (n x p) matrix of observations on the explanatory variables, B is a (p x 1) vector of unknown regression parameters, and e is a (n x 1) vector of random errors. From (1), the least squares estimate of the vector of regression coefficients is

$$B = (X'X)^{-1}X'Y \qquad (2)$$

The least squares projection matrix, often called the hat matrix, is of fundamental importance in the identification of items that elicit atypical performance scores between the groups being compared. The hat matrix is defined as

$$H = X(X^1X)^{-1}X^1 \qquad (3)$$

The diagonal elements of H, denoted h , measure the influence or leverage of the response variable y on its corresponding fitted value.

Results derived by Belsley, et. al., (1980), and Hoaglin and Welsch (1978) provide a statistical criterion to set cutoff levels to identify observations whose pattern of influence is atypical. Their results indicate that values of h larger than 2*(p/n) need further examination due to their unusually large influence on the hat matrix, H. Observations that exceed this cutoff level are often termed 'leverage points' in the statistical literature.

Values of the diagonal elements of the H matrix are recorded in column 1 of tables 1 and 2 respectively. An examination of these values indicates that the cutoff level of .133 is exceeded by items 1 and 14 in model 1, and items 13 and 14 in model 2. The quantitative influence of these items on other aspects of the regression analysis is examined further in the following sections of this investigation.

A common practice in the item bias literature has been that of identifying as biased those items with large residual values in fitted linear models. This approach fails to take into account the fact that the variances of the residuals are not constant, but a function of the X matrix. Therefore, results so derived may lead to unwarranted conclusions concerning their potential bias. To avoid the problems associated with the non-constancy of the variances of the residuals, atypical items can be identified by scaling the residuals by their respective variances. For these purposes the residuals can be modified in ways that enhance our ability to detect those items which elicit the statistically most dissimilar performance. This transformation of the residuals is illustrated next. From (1) a least squares fit produces residuals given by

$$e = (I - X(X^1X)^{-1}X^1) \tag{6}$$

and mean square residuals

$$s = \frac{e^1e}{n-p} \tag{7}$$

The variance-covariance matrix of estimates of the residuals is

$$Var(e) = \sigma^2(I-H) \tag{8}$$

where H is the least squares projection matrix defined in (3). Standardizing the residuals by estimating $\sigma^2$ by the residual mean square based on regression estimates without the ith observation yields the ratio of 'studentized residuals',

$$e(i) = \frac{e(i)}{s(i)\sqrt{1-h_i}} \tag{9}$$

These residuals are distributed as a t-distribution with n-p-1 degrees of freedom. Therefore, if the Gaussian assumption holds, the significance of any one of these studentized residuals can be readily assbssed from tabulated values of the t-distribution with n-p-1 degrees of freedom.

Estimates of the studentized residuals are listed in column 3 of tables 1 and 2. The magnitude of the studentized residual for items 1 and 26 consistently exceeds the critical value of 1.70 ( t, 27 df alpha= .05). In this particular

TABLE 1

White Regression Model

Model 1

| Item No. | Hat Matrix | Raw Resid. | Stdzed. Resid. | Covar. Ratio | DFFITS | DFBETAS Const. | DFBETAS Slop |
|---|---|---|---|---|---|---|---|
| 1 | 0.20* | -1.07 | -3.24* | 0.70* | -1.65* | -0.71* | -1.5 |
| 2 | 0.03 | - .56 | -1.43 | 0.96 | -0.27 | -0.26 | -0.0 |
| 3 | 0.03 | - .11 | -0.29 | 1.11 | -0.05 | -0.05 | 0.0 |
| 4 | 0.09 | - .47 | -1.22 | 1.06 | -0.40 | -0.24 | -0.3 |
| 5 | 0.03 | - .05 | -0.14 | 1.11 | -0.02 | -0.02 | -0.0 |
| 6 | 0.04 | 0.10 | 0.26 | 1.11 | 0.05 | 0.05 | 0.0 |
| 7 | 0.04 | 0.25 | 0.62 | 1.09 | 0.14 | 0.11 | -0.0 |
| 8 | 0.09 | - .61 | -1.63 | 0.98 | -0.54 | -0.29 | 0.4 |
| 9 | 0.04 | .08 | 0.19 | 1.12 | 0.04 | 0.03 | -0.0 |
| 10 | 0.05 | 0.43 | 1.10 | 1.03 | 0.25 | 0.20 | -0.1 |
| 11 | 0.04 | 0.31 | 0.77 | 1.07 | 0.16 | 0.14 | 0.0 |
| 12 | 0.03 | 0.28 | 0.70 | 1.07 | 0.13 | 0.13 | -0.0 |
| 13 | 0.12 | 0.26 | 0.68 | 1.19 | 0.26 | 0.14 | 0.2 |
| 14 | 0.14* | -0.44 | -1.20 | 1.13 | -0.49 | -0.22 | 0.4 |
| 15 | 0.04 | 0.32 | 0.81 | 1.06 | 0.16 | 0.15 | 0.0 |
| 16 | 0.05 | - .12 | -0.31 | 1.12 | -0.07 | -0.05 | 0.0 |
| 17 | 0.03 | 0.29 | 0.73 | 1.06 | 0.13 | 0.13 | -0.0 |
| 18 | 0.03 | 0.17 | 0.43 | 1.10 | 0.08 | 0.08 | 0.0 |
| 19 | 0.08 | 0.21 | 0.55 | 1.14 | 0.16 | 0.10 | -0.1 |
| 20 | 0.03 | - .35 | -0.88 | 1.05 | -0.16 | -0.16 | -0.0 |
| 21 | 0.12 | - .18 | -0.47 | 1.20 | -0.17 | -0.08 | 0.1 |
| 22 | 0.06 | 0.24 | 0.62 | 1.12 | 0.17 | 0.12 | 0.1 |
| 23 | 0.03 | - .00 | -0.00 | 1.11 | -0.00 | -0.00 | 0.0 |
| 24 | 0.06 | .02 | 0.07 | 1.15 | 0.01 | 0.01 | -0.0 |
| 25 | 0.03 | 0.59 | 1.52 | 0.94 | 0.28 | 0.28 | 0.0 |
| 26 | 0.04 | 0.75 | 1.97* | 0.85 | 0.41 | 0.37* | 0.1 |
| 27 | 0.10 | -0.43 | -1.12 | 1.09 | -0.38 | -0.22 | -0.3 |
| 28 | 0.05 | - .26 | -0.67 | 1.09 | -0.15 | -0.12 | 0.0 |
| 29 | 0.08 | 0.48 | 1.24 | 1.04 | 0.37 | 0.24 | 0.2 |
| 30 | 0.05 | - .19 | -0.47 | 1.11 | -0.11 | -0.08 | 0.0 |

TABLE 2

Black Regression Model

Model 2

| Item No. | Hat Matrix | Raw Resid. | Stdzed. Resid. | Covar. Ratio | DFFITS | DFBETAS Const. | DFBETAS Slope |
|---|---|---|---|---|---|---|---|
| 1 | 0.10 | 1.00 | 3.90* | 0.49* | 1.35* | 0.31 | 1.12* |
| 2 | 0.03 | 0.46 | 1.42 | 0.96 | 0.27 | 0.27 | -0.06 |
| 3 | 0.04 | .05 | 0.16 | 1.11 | 0.03 | 0.03 | -0.01 |
| 4 | 0.06 | 0.49 | 1.56 | 0.96 | 0.42 | 0.17 | 0.29 |
| 5 | 0.03 | .06 | 0.20 | 1.11 | 0.03 | 0.03 | 0.00 |
| 6 | 0.04 | - .03 | -0.11 | 1.12 | -0.02 | -0.01 | -0.01 |
| 7 | 0.04 | - .25 | -0.78 | 1.07 | -0.16 | -0.16 | 0.07 |
| 8 | 0.13 | 0.36 | 1.17 | 1.12 | 0.47 | 0.36 | -0.40 |
| 9 | 0.04 | - .12 | -0.36 | 1.11 | -0.07 | -0.07 | 0.03 |
| 10 | 0.03 | - .41 | -1.26 | 0.99 | -0.25 | -0.25 | 0.09 |
| 11 | 0.05 | - .19 | -0.58 | 1.10 | -0.14 | -0.07 | -0.08 |
| 12 | 0.03 | - .22 | -0.68 | 1.07 | -0.12 | -0.11 | -0.02 |
| 13 | 0.14 | - .05 | -0.17 | 1.25* | -0.07 | -0.01 | -0.06 |
| 14 | 0.17 | 0.19 | 0.62 | 1.26* | 0.29 | 0.20 | -0.26 |
| 15 | 0.04 | - .21 | -0.65 | 1.09 | -0.14 | -0.08 | -0.08 |
| 16 | 0.05 | .03 | 0.10 | 1.13 | 0.02 | 0.02 | -0.01 |
| 17 | 0.03 | - .24 | -0.73 | 1.07 | -0.12 | -0.12 | -0.01 |
| 18 | 0.04 | - .10 | -0.31 | 1.11 | -0.06 | -0.04 | -0.02 |
| 19 | 0.06 | - .28 | -0.86 | 1.09 | -0.23 | -0.21 | 0.16 |
| 20 | 0.03 | 0.29 | 0.88 | 1.05 | 0.16 | 0.16 | -0.02 |
| 21 | 0.12 | .00 | 0.01 | 1.23* | 0.00 | 0.00 | -0.00 |
| 22 | 0.07 | - .10 | -0.31 | 1.15 | -0.09 | -0.03 | -0.07 |
| 23 | 0.03 | - .02 | -0.06 | 1.11 | -0.01 | -0.01 | 0.00 |
| 24 | 0.06 | - .11 | -0.34 | 1.13 | -0.09 | -0.08 | 0.06 |
| 25 | 0.04 | - .46 | -1.44 | 0.96 | -0.30 | -0.20 | -0.13 |
| 26 | 0.06 | - .55 | -1.76* | 0.92 | -0.46 | -0.19 | -0.33 |
| 27 | 0.07 | 0.47 | 1.49 | 0.99 | 0.42 | 0.15 | 0.31 |
| 28 | 0.05 | 0.15 | 0.45 | 1.12 | 0.11 | 0.10 | -0.07 |
| 29 | 0.10 | - .27 | -0.85 | 1.14 | -0.30 | -0.06 | -0.25 |
| 30 | 0.05 | .08 | 0.25 | 1.13 | 0.06 | 0.06 | -0.04 |

case there is substantial agreement between those items with relatively large residuals, and those with relatively large studentized residuals. The magnitude of the studentized residuals associated with items 1 and 26 indicate that the performance of white and black examinees in these two items is significantly different from the performance pattern established in other items. And as such, these items warrant further examination for potential bias. The studentized residuals e(i) offer a substantial improvement over the usual analysis of raw residuals, both because they have equal variances and because an underlying distribution theory exists to identify atypical values.

Another important group of diagnostic methods measure the impact of the deletion of the ith observation on the stability of several statistical ratios, and estimated regression coefficients. Statistical procedures that have been developed to estimate the impact of the deletion of the ith observation on these statistics, are examined next. An important diagnostic statistic is the covariance ratio. This ratio is formed by comparing the covariance of the regression model whith the ith observation deleted, and the covariance of the complete regression model. By repeating this procedure for each observation in the sample, a set of N values that corresponds to estimates of the covariance ratios is obtained. Atypical items can be identified by measuring the impact of their deletion on the estimates of the covariance ratios. Covariance ratios based on the 'not ith' observation which deviate from one, indicate that this particular observation is exerting an atypical influence, and needs therefore further examination. From (1) the variance-covariance matrix of the regression coefficients is:

$$Var(b) = \sigma^2 (X^1X)^{-1} \tag{11}$$

Similarly, the variance-covariance matrix of the regression coefficients due to the 'not ith' observation is,

$$Var(b(i)) = \sigma^2 (X^1(i)X(i))^{-1} \tag{12}$$

Several statistics have been proposed for comparing these variance-covariance matrices. A suggested approach is based on analysis of the ratio of determinants of both matrices. If the effect of the deletion of the ith observation from the model is minor, the ratio of the computed values of both determinants would be close to one. On the other hand, if the value of the ith observation is atypical, its deletion from the model, would result in a value of this ratio far from one.

A limitation in using this ratio is the fact that the estimator of $\sigma$ given by S is also affected by the deletion of the ith observation. However, Belsley, Kuh and Welsch (1980) show that by forming the determinantal ratio of both matrices, i.e., with all, and with the 'not ith' observation, a test statistic results

$$COVRATIO = \frac{s(i)^{2p}}{s^{2p}} \left\{ \frac{|(X^1(i)X(i))^{-1}|}{|(X^1X)^{-1}|} \right\} \tag{13}$$

Values of this ratio outside the interval $1 \pm 3p/n$ identify items whose deletion cause atypical perturbations on the estimates of the covariance-ratio. In summary, values of this determinantal ratio greater than one, imply that the deletion of the ith item impairs estimation efficiency. Conversely, determinantal ratios less than one imply that the deletion of ith item enhances estimation efficiency.

Values of the covariance ratio are recorded in column 4 of tables 1 and 2. Examination of these estimates indicates that the deletion of item 1 causes an unusually large perturbation on this statistic. Its computed value of .70 lies outside the interval ( .80 - 1.20 ). This result is consistent with previous findings which identify item 1 as eliciting a pattern of influence statisticallly different from the remaining items. A similar analysis of estimates of this ratio listed in table 2 ( model 2), identifies four items whose deletions cause unusually large perturbations and lie outside the interval ( .80 - 1.20 ). These items are: item 1, 13, 14, and 21. All but item 21 have been previously identified as items whose pattern of influence needs further examination.

Another important regression diagnostic is derived from Analysing the effect of the deletion of the ith observation on the predictive performance of a regression model. This effect can be conveniently summarized by the DFFIT coefficient. Following results of Belsley et. al., (1980), this statistic can be estimated by

$$DFFIT_i \equiv \hat{Y}_i - \hat{Y}_i(i) = x_i\left[\hat{\beta} - \hat{\beta}(i)\right] = h_i e_i / 1 - h_i \quad (14)$$

For purposes of scaling, this quantity is divided by an estimate of $\sigma\sqrt{h_i}$. This adjustment yields the statistic

$$DFFITS_i = \frac{\sqrt{h_i}\, e_i}{s(i)(1-h_i)} \quad (15)$$

where $\sigma$ has been estimated by $S(i)$. Estimates of this coefficient are recorded in column 5 of tables 1 and 2. Values of this statistic larger than $2 * \sqrt{(p/n)}$ ex ert atypical effects on the predictive performance of the model. The DFFIT statistic is useful in the following context. Outliers often pull the estimated regression plane towards themselves. This often results in residual values smaller than their true value. The DFFIT statistic avoids this problem by re-estimating each residual with regression estimates that do not use that observation. The DFFIT statistic offers a very sensitive regression diagnostic for detecting potentially biased items, by identifying unusual patterns of influence on the predictive ability of the model.

Another important regression diagnostic applied to detect potentially biased items is based on analysis of the magnitude of the changes on the regression coefficients caused by the deletion on each item. In the simple bivariate model, for example, items whose deletion effect large perturbation on the intercept and slope estimates can be readily identified. Their large effects on the regression coefficients may indicate particular characteristics of an item that is lacking in others. These characteristics may, in turn, either increase or decrease the a priori probability of a correct response in one group of examinees but not in another. The identification of items whose deletion cause large perturbations on estimates of the regression coefficients is therefore of great value in helping to detect potentially biased items. Atypical perturbations in estimates of regression coefficients that may ensue as a result of their deletion can greatly facilitate the identification of atypical items. If we let b(i) be the vector of regression coefficients in a model that does not use the ith observation, the change or sensitivity of these coefficients can be estimated by

$$DFBETAS_{ij} = \left[ \hat{\beta}_j - \hat{\beta}_j(i) \right] / \left[ s(i) \sqrt{(X^1X)^{-1}_{ij}} \right] \quad (16)$$

Belsley et. al., (1980) suggest several statistical criteria to set cutoff levels to identify atypical coefficient changes. A proposed cutoff is $2 / \sqrt{n}$ . This cutoff measures the change in the estimates of the regression coefficients in units measured in standard deviations. In our analysis, items whose deletion cause a change of a least .365 standard deviations are deemed influential and warrant further examination for potential bias. Items whose DFBETAS exceed this cutoff are noted in columns 6 and 7 of tables 1 and 2 respectively.

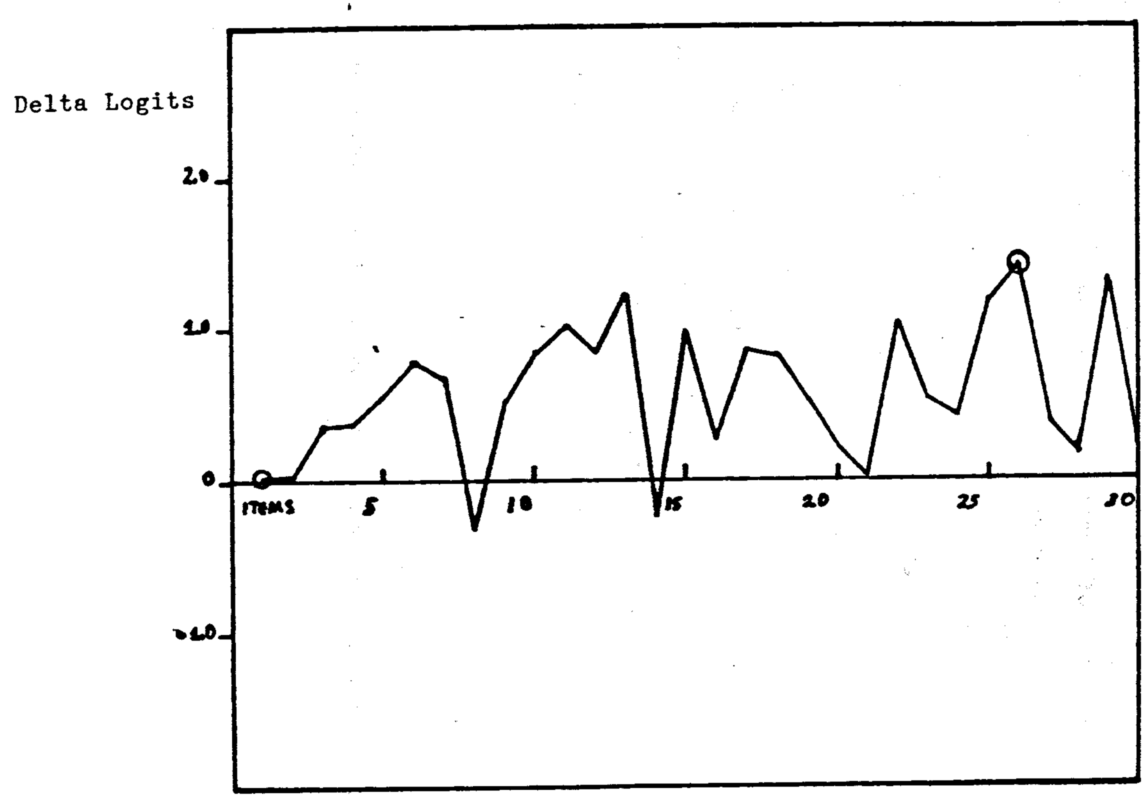
Further statistical analysis was carried out on the differences of logits of individual item p-values. These differences or delta values are defined as

$$DELTA = LOGIT(P_w) - LOGIT(P_b) \tag{17}$$

A plot of these values against national P-values is given in figure 2. Under the assumption of equal performance, a fitted line through these values is expected to have a zero slope and zero intercept term. The observed dispersion of these DELTA values above zero indicates that a higher proportion of white examinees relative to black examinees has responded correctly to those exercises. The magnitude of these DELTA values is not,however,constant. From figure 2, a gradual increase in their magnitude is apparent. This trend suggests that the difference in performance between white and black examinees is not as marked among difficult items, as it is among relatively easier items. This performance

FIGURE 2

PLOT OF DELTA VALUES OF LOGITIZED P-VALUES



differential suggests that some items are equally difficult for both white and black examinees. However, as the level of difficulty decreases, a higher proportion of white examinees relative to black examinees succeeds in given a correct answer. A least squares fit to the dispersion of DELTA values produces a significant slope estimate (.01, p=.001). The estimate of the intercept term is not statistically different from zero (-.07,p=.63). From this gradual pattern in the magnitude of DELTA values, items that elicit atypical performance patterns can then be identified and contrasted with previous results.

Results of analysis of the regression diagnostics is listed in table 3. Examination of the magnitude of raw and studentized residuals identifies items 1 and 26 as eliciting residual values statistically different from the dispersion pattern established by the remaining items. This result is consistent with previous results, which identify the same items as atypical. Analysis of estimates of the covariance ratio identify items 1, 14 and 21 as exceeding the interval (.80 - 1.20). The extremely low value of this ratio due to the deletion of item 1 indicates that this item is highly atypical. This result contrasts well with our previous findings based on predictive models of white and black per-

TABLE 3

Delta Logits Regression Model

Model 3

| Item No. | Hat Matrix | Raw Resid. | Stdzed. Resid. | Covar. Ratio | DFFITS | DFBETAS Const. | DFBETAS Slop |
|---|---|---|---|---|---|---|---|
| 1 | 0.09 | - .95 | -3.44* | 0.57* | -1.09* | 0.52* | -0.8 |
| 2 | 0.03 | - .49 | -1.51 | 0.94 | -0.28 | -0.14 | 0.0 |
| 3 | 0.04 | - .05 | -0.15 | 1.12 | -0.03 | -0.02 | 0.0 |
| 4 | 0.07 | - .53 | -1.68 | 0.95 | -0.49 | 0.20 | -0.3 |
| 5 | 0.03 | - .11 | -0.32 | 1.10 | -0.06 | -0.00 | -0.0 |
| 6 | 0.05 | - .00 | -0.02 | 1.13 | -0.00 | 0.00 | -0.0 |
| 7 | 0.04 | 0.27 | 0.82 | 1.07 | 0.17 | 0.14 | -0.0 |
| 8 | 0.13 | - .38 | -1.21 | 1.11 | -0.47 | -0.47 | 0.4 |
| 9 | 0.04 | 0.14 | 0.42 | 1.11 | 0.09 | 0.08 | -0.0 |
| 10 | 0.04 | 0.42 | 1.29 | 0.99 | 0.26 | 0.20 | -0.1 |
| 11 | 0.06 | 0.16 | 0.48 | 1.12 | 0.12 | -0.03 | 0.0 |
| 12 | 0.03 | 0.19 | 0.57 | 1.08 | 0.11 | 0.01 | 0.0 |
| 13 | 0.10 | 0.21 | 0.65 | 1.15 | 0.21 | -0.11 | 0.1 |
| 14 | 0.15 | - .25 | -0.80 | 1.21* | -0.34 | -0.33 | 0. |
| 15 | 0.05 | 0.18 | 0.53 | 1.11 | 0.12 | -0.03 | 0. |
| 16 | 0.06 | - .00 | -0.00 | 1.14 | -0.00 | -0.00 | 0. |
| 17 | 0.03 | 0.21 | 0.64 | 1.08 | 0.12 | 0.02 | 0. |
| 18 | 0.04 | .05 | 0.17 | 1.12 | 0.03 | -0.00 | 0. |
| 19 | 0.08 | 0.32 | 0.99 | 1.08 | 0.29 | 0.28 | -0. |
| 20 | 0.03 | - .33 | -0.99 | 1.03 | -0.18 | -0.07 | -0. |
| 21 | 0.12 | - .00 | -0.01 | 1.23* | -0.00 | -0.00 | 0. |
| 22 | 0.07 | 0.10 | 0.31 | 1.15 | 0.09 | -0.04 | 0. |
| 23 | 0.03 | .01 | 0.03 | 1.11 | 0.00 | 0.00 | -0. |
| 24 | 0.07 | 0.15 | 0.46 | 1.14 | 0.13 | 0.12 | -0. |
| 25 | 0.04 | 0.43 | 1.32 | 0.99 | 0.28 | -0.03 | 0. |
| 26 | 0.06 | 0.55 | 1.73* | 0.93 | 0.46 | -0.16 | 0. |
| 27 | 0.08 | - .49 | -1.53 | 0.98 | -0.45 | 0.19 | -0. |
| 28 | 0.06 | - .12 | -0.37 | 1.13 | -0.09 | -0.09 | 0. |
| 29 | 0.09 | 0.33 | 1.03 | 1.09 | 0.32 | -0.15 | 0. |
| 30 | 0.06 | -0.05 | -0.16 | 1.14 | -0.04 | -0.04 | 0. |

formance. Similarly, analysis of the significance of the DFFITS and DFBETAS statistics consistently identifies item 1 as eliciting perturbations statistically different from those caused due to the deletion of the remaining items.

## CONCLUSIONS

Results of applying the regression diagnostics proposed in this investigation consistently identify items 1 and 26 as eliciting response patterns statistically different from those observed in the remaining items. Although the preceding results do not imply that these items are biased, the magnitude of the perturbation on several statistics due to their deletion suggests that these items deem further examination.

Given the preceding, the performance of these two groups in these two items was further analyzed. Results of analysis of item 1 indicates that the performance of white and black examinees in this particular item was almost identical, with observed p-values of 93.6 and 93.5 respectively. This is a very atypical performance that substantially deviates from the pattern established by these groups of examinees in the remaining items.

By contradistinction, analysis of item 26 indicates that the observed performance gap is highly atypical. The observed p-values of 87.7 and 63.1 for white and black examinees respectively, substantially deviate from the distribution of performance values observed in the remaining items. Although the preceding results do not imply that these items are biased, the highly atypical performance levels they elicit among these examinees needs serious further examination. Item 26 in particular elicits an inordinately large performance gap that far exceeds the performance differential observed in the remaining items between black and white examinees.

The preceding results indicate how the recent developments in the analysis of regression models may prove useful in the identification of atypical and potentially biased items. Moreover, it is contended that the application of statistical criteria to set cutoff levels and identify atypical observations offers a substantial refinement over existing approaches, namely, delta plot, chi-square and latent trait methods.

FIGURE 3

STUDENTIZED RESIDUALS

WHITE REGRESSION MODEL

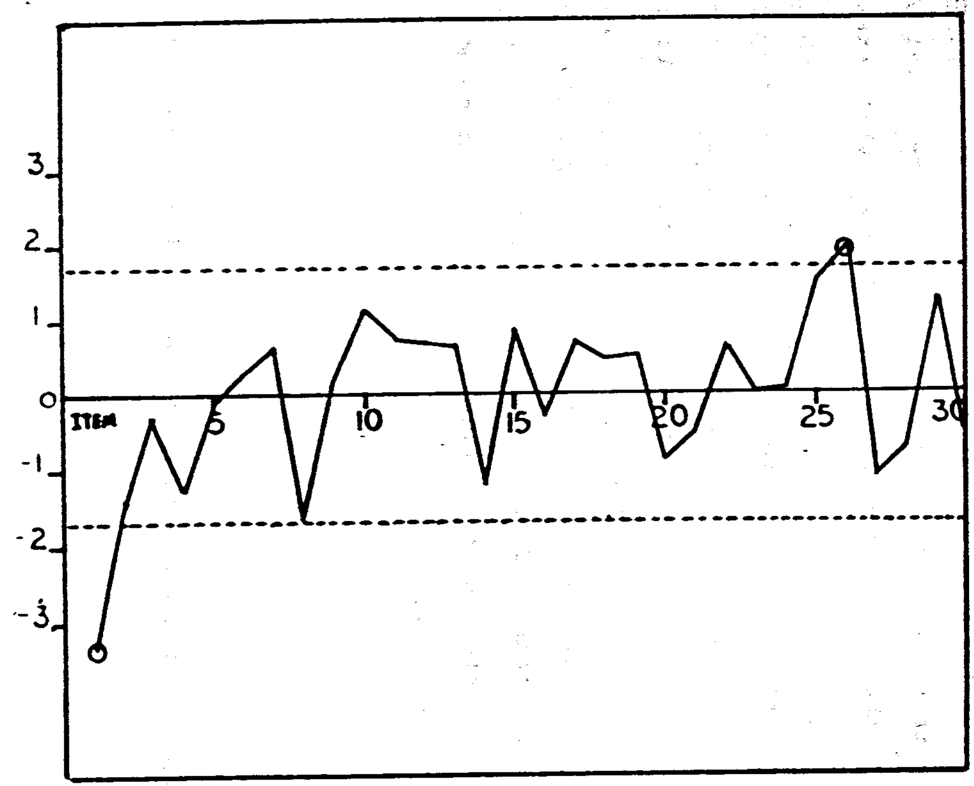


FIGURE 4

STUDENTIZED RESIDUALS

BLACK REGRESSION MODEL

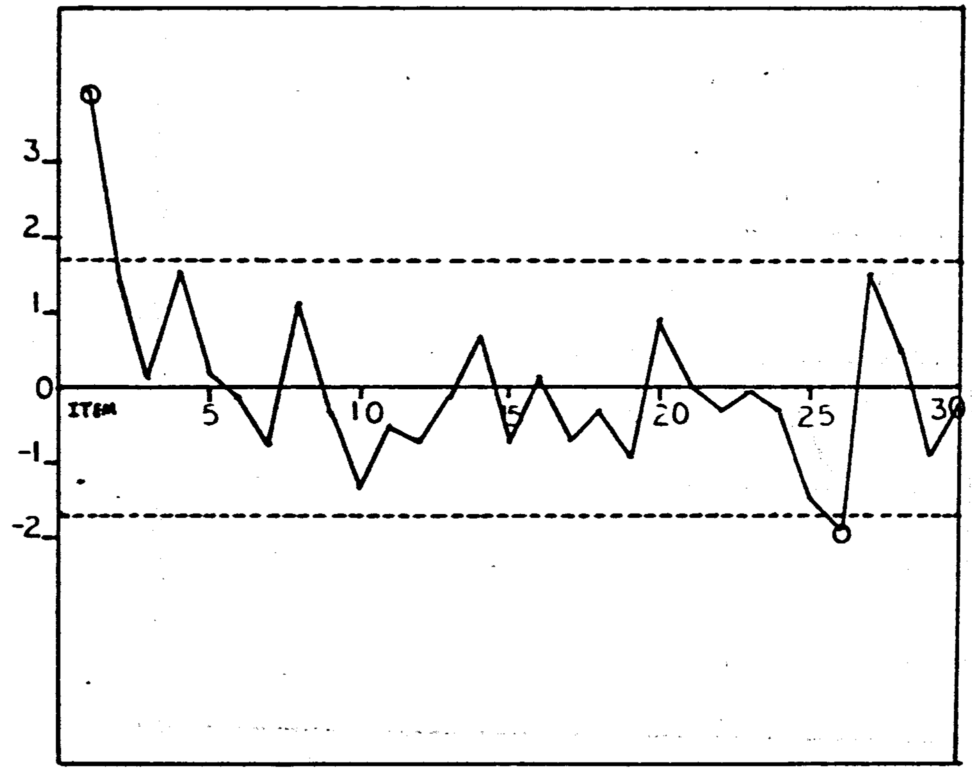
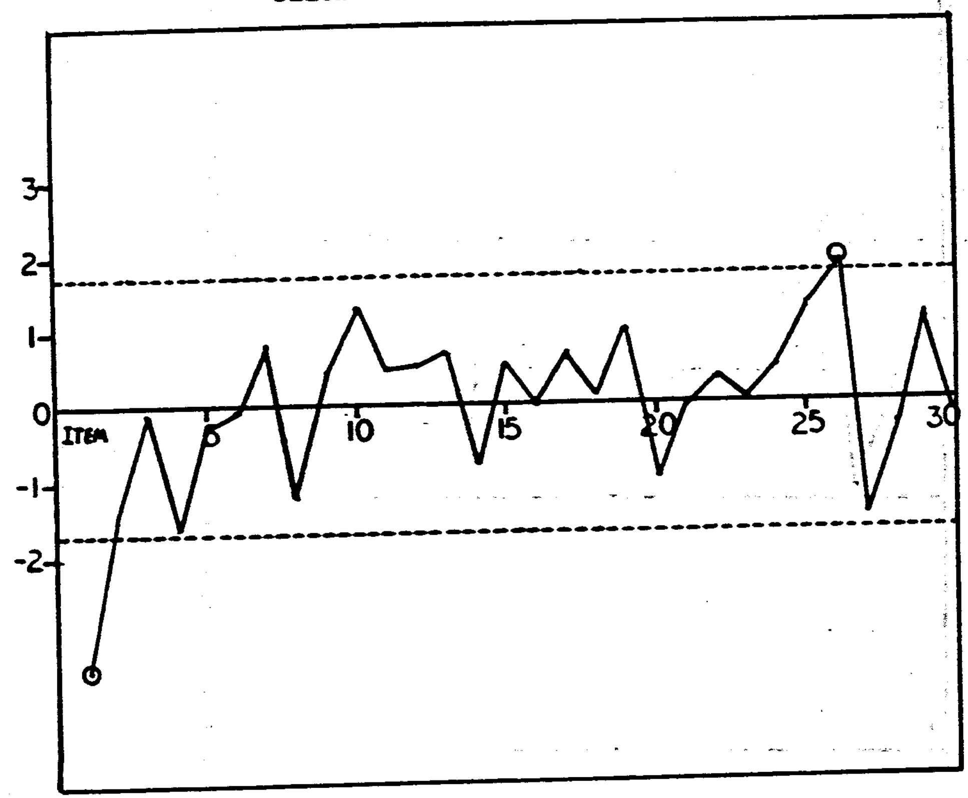


FIGURE 5

STUDENTIZED RESIDUALS

DELTA LOGITS REGRESSION MODEL



## REFERENCES


Angoff, W. H., and Ford, S. F. Item-Race Interaction on a Test of Scholastic Aptitude, Journal of Educational Measurement, 1973, Vol. 10, pp. 95-105.

Belsley, D. A., et al. Regression Diagnostics, John Wiley and Sons: New York, 1980.

Camilli, G. A Critique of the Chi-Square Method for Assessing Item Bias. Unpublished paper, Laboratory of Educational Research, University of Colorado, Boulder, 1979.

Hambleton, R. K. and Cook, L. L. Latent Trait Models and Their Use in the Analysis of Educational Test Data, Journal of Educational Measurement, 1977, Vol. 14, pp. 75-96.

Hoaglin D. C. and R. E. Welsch. The Hot Matrix on Regression and ANOVA, The American Statistician, 1978, 32, pp. 17-22.

Merz, W. R., et al. An Empirical Investigation of Six Methods for Examining Test Item Bias. Report submitted to the National Institute of Education, Grant 6-78-0067, California State University, Sacramento, California, 1978.

Lord, F. M. A Study of Item Bias Using Item Characteristic Curve Theory. In N. H. Poartinga (ed.) Basic Problems in Cross-cultural Psychology, Amsterdam: Swits and Vitlinger, 1977.

Lord, F. M. and Novick, M. R. Statistical Theories of Mental Test Scores, Addison-Wesley: Reading, Massachusetts, 1918.

Petersen, N. S. Bias in the Selection Rule: Bias in the Test. Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands, 1977.

Rudner, L. M., Getson, P. R. and Knight, D.L. A Monte Carlo Comparison of Seven Biased Item Detection Techniques, Journal of Educational Measurement, 1980, Vol. 17, pp. 1-10.

Scheuneman, J. A Method for Assessing Bias in Test Items, Journal of Educational Measurement, 1979, Vol. 16, pp. 143-152.

Sheppard, L. A., et al. Comparison of Six Procedures for Detecting Test Item Bias Using Both Internal and External Ability Criteria. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, April 1980.

# The Use of Nonsence Coding with ANOVA Situations

John D. Williams
The University of North Dakota

Summary: Nonsense coding systems can be constructed that retain outcomes regarding $R^2$ values, F values and multiple comparison tests. Nonsense coding highlights the flexibility of coding ANOVA problems to be analyzed by multiple linear regression procedures; however, no additional analytic power appears to be gained from their use.

## Characteristic Coding Compared to Nonsense Coding

Most coding systems for accomplishing ANOVA solutions by multiple linear regression use some variant of characteristic coding (binary coding/dummy coding) with the use of 1's or 0's, depending upon group membership, or contrast coding, which uses 1's, 0's and -1's (see Williams, 1974a). The use of orthogonal contrasts deviates from this usage, including orthogonal polynomials, but none of these systems allow arbitrariness in their coding process.

On the other hand, Cohen and Cohen (1975) assert that regression solutions can be accomplished through the use of "nonsense" coding, though they neither give directions nor examples of this process. Thus, an example of nonsense coding is provided here. The data are taken from Williams (1974b, p. 43, problem 5.3). See Table 1.

Table 1

Sample Data for ANOVA Problem

| Group One | Group Two | Group Three | Group Four | Group Five |
|---|---|---|---|---|
| 19 | 20 | 13 | 12 | 22 |
| 18 | 19 | 12 | 8 | 20 |
| 15 | 16 | 10 | | 19 |
| 13 | 16 | 10 | | 19 |
| 8 | 14 | 10 | | 15 |
| 5 | 14 | | | |
| | 13 | | | |

The data in Table 1 are clearly from unequal sized groups; the intent is to show outcomes that have generality beyond equal cell sized situations. First, to accomplish a characteristic coding of this data:

$Y$ = the criterion score;

$X_1$ = 1 if a member of Group One, 0 otherwise;

$X_2$ = 1 if a member of Group Two, 0 otherwise;

$X_3$ = 1 if a member of Group Three, 0 otherwise;

$X_4$ = 1 if a member of Group Four, 0 otherwise; and

$X_5$ = 1 if a member of Group Five, 0 otherwise.

Table 2 shows these values for the data in Table 1.

Table 2

Characteristic Coding (1 or 0) for Data in Table 1

| Y | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|
| 19 | 1 | 0 | 0 | 0 | 0 |
| 18 | 1 | 0 | 0 | 0 | 0 |
| 15 | 1 | 0 | 0 | 0 | 0 |
| 13 | 1 | 0 | 0 | 0 | 0 |
| 8 | 1 | 0 | 0 | 0 | 0 |
| 5 | 1 | 0 | 0 | 0 | 0 |
| 20 | 0 | 1 | 0 | 0 | 0 |
| 19 | 0 | 1 | 0 | 0 | 0 |
| 16 | 0 | 1 | 0 | 0 | 0 |
| 16 | 0 | 1 | 0 | 0 | 0 |
| 14 | 0 | 1 | 0 | 0 | 0 |
| 14 | 0 | 1 | 0 | 0 | 0 |
| 13 | 0 | 1 | 0 | 0 | 0 |
| 13 | 0 | 0 | 1 | 0 | 0 |
| 12 | 0 | 0 | 1 | 0 | 0 |
| 10 | 0 | 0 | 1 | 0 | 0 |
| 10 | 0 | 0 | 1 | 0 | 0 |
| 10 | 0 | 0 | 1 | 0 | 0 |
| 12 | 0 | 0 | 0 | 1 | 0 |
| 8 | 0 | 0 | 0 | 1 | 0 |
| 22 | 0 | 0 | 0 | 0 | 1 |
| 20 | 0 | 0 | 0 | 0 | 1 |
| 19 | 0 | 0 | 0 | 0 | 1 |
| 19 | 0 | 0 | 0 | 0 | 1 |
| 15 | 0 | 0 | 0 | 0 | 1 |

Next five different linear models can be defined to complete an analysis by multiple linear regression:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + e_1; \quad (1)$$

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_5X_5 + e_1; \quad (2)$$

$$Y = b_0 + b_1X_1 + b_2X_2 + b_4X_4 + b_5X_5 + e_1; \quad (3)$$

$$Y = b_0 + b_1X_1 + b_3X_3 + b_4X_4 + b_5X_5 + e_1; \quad (4)$$

and $$Y = b_0 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e ; \quad (5)$$

Equations 1 thru 5 are reparameterizations of each other and are reparameterizations of

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e_1. \quad (6)$$

The use of equations 1 thru 5 require the use of a unit vector for solution (commonly a part of typical multiple use multiple lienar regression programs) and represent solutions that allow psuedo-Dunnett formulations that permit construction of all simple comparisons of means (see Williams, 1976). Also, the $b_i$'s are unique to each equation. Each of the formulations yields $R^2 = .49362$, $F = 4.874$ with df = 4,20 and $p < .05$. A part of the printout is shown in Table 3 for equation 1.

Table 3

Portions of Printout for Multiple Linear Regression for the Sample Data in Table 1 Using 1 or 0 Coding

| Variable | Mean | Correlation | Regression Coefficient | Standard Error of Regression Coefficient | Comput[illegible] t Valu[illegible] |
|---|---|---|---|---|---|
| $X_1$ | .240 | -.181 | -6.000 | 2.089 | -2.872 |
| $X_2$ | .280 | .230 | -3.000 | 2.020 | -1.485 |
| $X_3$ | .200 | -.392 | -8.000 | 2.182 | -3.667 |
| $X_4$ | .080 | -.299 | -9.000 | 2.886 | -3.118 |
| Criterion | 14.400 | | | | |
| Intercept | 19.000 | | | | |

In Table 3, means refer to the proportion in a group for characteristic (1 or 0) coded data. The regression coefficient is the difference between the mean of the particular coded group and the "left-out" group (Group Five). If the regression coefficient is divided by its own standard error, the computed t value is found which can be compared to a table for an appropriate multiple comparison method (e.g., Tukey's test). The correlations in Table 3 represent point-biserial correlations of the group membership variables with the criterion. The criterion is the overall mean for the Y scores, and the intercept ($b_0$) is the mean of the "left-out" group (Group Five). A reformulation of equation 1 makes these relationships more obvious:

$$Y = \bar{Y}_5 + (\bar{Y}_1-\bar{Y}_5)X_1 + (\bar{Y}_2-\bar{Y}_5)X_2 + (\bar{Y}_3-\bar{Y}_5)X_3 + (\bar{Y}_4-\bar{Y}_5)X_4 + e_1. \quad (7)$$

The set of all simple multiple comparisons, omitting signs and lower diagonal entries is shown in Table 4.

Table 4

Means and Computed t Values for all Simple Comparisons Using Characteristic (1 or 0) Coding

| Group | One | Two | Three | Four | Five |
|---|---|---|---|---|---|
| Mean | 13.00 | 16.00 | 11.00 | 10.00 | 19.00 |
| One | | 1.563 | .957 | 1.065 | 2.872 |
| Two | | | 2.475 | 2.169 | 1.485 |
| Three | | | | .346 | 3.667 |
| Four | | | | | 3.118 |

Using Tukey's Test ($p < .05$) a t value of 2.992 is required for significance.

## Using Nonsense Coding

Nonsense coding consistent with the characteristic coding process can be accomplished in the following manner:

Let Y = the criterion score

$X_1$ = a if a member of Group One, b otherwise ($a \neq b$);

$X_2$ = c if a member of Group Two, d otherwise ($c \neq d$);

$X_3$ = e if a member of Group Three, f otherwise ($e \neq f$);

$X_4$ = g if a member of Group Four, h otherwise ($g \neq h$); and

$X_5$ = i if a member of Group Five, j otherwise ($i \neq j$).

It can be noted that the solution given earlier is the special case using this notation where a = c = e = g = i = 1 and b = d = f = h = j = 0. As an example of choosing values for a thru j, let a = 7, b = 3, c = 2, d = 9, e = 4, f = 1, g = 8, h = 5, i = 6, and j = 2. Using these values, similar equations were constructed to equations 1 thru 5 and multiple regressions were completed. For the data set itself, see Table 5.

Table 5

Characteristic Coding Using a Nonsense Coding Process for Data in Table 1

| Y | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|
| 19 | 7 | 9 | 1 | 5 | 2 |
| 18 | 7 | 9 | 1 | 5 | 2 |
| 15 | 7 | 9 | 1 | 5 | 2 |
| 13 | 7 | 9 | 1 | 5 | 2 |
| 8 | 7 | 9 | 1 | 5 | 2 |
| 5 | 7 | 9 | 1 | 5 | 2 |
| 20 | 3 | 2 | 1 | 5 | 2 |
| 19 | 3 | 2 | 1 | 5 | 2 |
| 16 | 3 | 2 | 1 | 5 | 2 |
| 16 | 3 | 2 | 1 | 5 | 2 |
| 14 | 3 | 2 | 1 | 5 | 2 |
| 14 | 3 | 2 | 1 | 5 | 2 |
| 13 | 3 | 2 | 1 | 5 | 2 |
| 13 | 3 | 9 | 4 | 5 | 2 |
| 12 | 3 | 9 | 4 | 5 | 2 |
| 10 | 3 | 9 | 4 | 5 | 2 |
| 10 | 3 | 9 | 4 | 5 | 2 |
| 10 | 3 | 9 | 4 | 5 | 2 |
| 12 | 3 | 9 | 1 | 8 | 2 |
| 8 | 3 | 9 | 1 | 8 | 2 |
| 22 | 3 | 9 | 1 | 5 | 6 |
| 20 | 3 | 9 | 1 | 5 | 6 |
| 19 | 3 | 9 | 1 | 5 | 6 |
| 19 | 3 | 9 | 1 | 5 | 6 |
| 15 | 3 | 9 | 1 | 5 | 6 |

Using formulations like equations 1 thru 5, each equation yields $R^2 = .49362$, $F = 4.874$, with $df = 4,20$ and $p < .05$, identically the same as before.

The appearance of other portions of the printout is somewhat changed; a portion of the printout corresponding to equation 1 is shown in Table 6 and can be compared to Table 3.

Table 6

Portions of Printout for Multiple Linear Regression Using Nonsense Coding for the Sample Data in Table 1

| Variable | Mean | Correlation | Regression Coefficient | Standard Error of Estimate | Computed t Value |
|---|---|---|---|---|---|
| $X_1$ | 3.960 | -.181 | -1.500 | .522 | -2.872 |
| $X_2$ | 7.040 | -.230 | .429 | .289 | 1.485 |
| $X_3$ | 1.600 | -.392 | -2.667 | .727 | -3.667 |
| $X_4$ | 5.240 | -.299 | -3.000 | .962 | -3.118 |
| Criterion | 14.400 | | | | |
| Intercept | 37.309 | | | | |

It is by no means obvious what the meaning of the mean, regression coefficient or standard error of estimate are from a cursory glance at the output. However, the correlation coefficients remain point-biserial correlation coefficients of each group membership variable with the criterion even though they are not 1's and 0's. Also, except for sign, the computed t values are identical with those found earlier. Thus, even though much of the output is unfamiliar, important aspects are identical to those found earlier. Actually, the means represent simply the mean values of a variable assigned by our coding scheme; for example, the coding in Group One on $X_1$ is 3 and .24 of the scores are from this group. The remaining .76 are from other groups and are coded 7. Then .24(3) + .76(7) = 3.96, the mean of $X_1$. The regression coefficients are part of the least squares process that achieve the same expected values as was found previously, that is, the mean for the group. A rather intractable equation, siimilar to equation 7, relates the means for the nonsense coding situation:

$$Y = \bar{Y}_5 - \{[b(\bar{Y}_1 - \bar{Y}_5)/ (a - b)] + [d(\bar{Y}_2 - \bar{Y}_5)/ (c - d)] + [f(\bar{Y}_3 - \bar{Y}_5)/(e - f)]\ [h(\bar{Y}_4 - \bar{Y}_5)/(g - h)]\} + [(\bar{Y}_1 - \bar{Y}_5)/(a - b)]X_1 + [(\bar{Y}_2 - \bar{Y}_5)/(c - d)]X_2 + [\bar{Y}_3 - \bar{Y}_5)/ (e - f)]X_3 + [(\bar{Y}_4 - \bar{Y}_5)/(g - h)]X_4 + e_1. \quad (8)$$

The relationship of the regression coefficients to the standard errors of estimate remains proportional so that the computed t values remain identical to those found for the characteristic coding solution.

## Contrast Coding with Nonsense Coding

Some researchers prefer to use contrast coding (see Williams, 1974a) to characteristic coding systems, particularly if they are interested in a traditional analysis of variance solution.* A typical contrast coding systems using either a 1 or -1 or 0 is as follows:

*Because the computed t values are directly interpretable as multiple comparisons (see equation 7) characteristic coding solutions would seem to be preferable for testing most hypotheses of interest making the characteristic coding solution not only simpler to achieve but more useful as well.

$X_1$ = 1 if a member of Group 1, -1 if a member of Group 5, 0 otherwise;

$X_2$ = 1 if a member of Group 2, -1 if a member of Group 5, 0 otherwise;

$X_3$ = 1 if a member of Group 3, -1 if a member of Group 5, 0 otherwise; and

$X_4$ = 1 if a member of Group 4, -1 if a member of Group 5, 0 otherwise.

A nonsense contrast coding can be accomplished as follows:

$X_1$ = a if a member of Group 1, -a if a member of Group 5, b otherwise ($a \neq b$);

$X_2$ = c if a member of Group 2, -c if a member of Group 5, d otherwise ($c \neq d$);

$X_3$ = e if a member of Group 3, -e if a member of Group 5, f otherwise ($e \neq f$); and

$X_4$ = g if a member of Group 4, -g if a member of Group 5, h otherwise ($g \neq h$).

If these two separate formations are used in a multiple linear regression solution, $R^2 = .49632$, $F = 4.874$, with df = 4,20 and $p < .05$ for both solutions, the same as found previously. Here, the computed t values contrast the group mean to the overall mean. Results for computed t values and correlation coefficients are the same for the usual contrast coding solution (using 1, 0 and -1) and the nonsense contrast coding solution (through different than those found for the characteristic coding scheme), although the means, regression coefficients and standard error of the regression coefficients differ from each other, as before. An equation similar to equation 8 (but even more intractable) can be developed for the nonsense contrast coding scheme.

## What is the Advantages/Disadvantages of Using Nonsense Coding

Perhaps the major advantage of nonsense coding is that it should allow users of regression a larger understanding of the coding process, and the "robustness" of the coding procedures. On occasion, a particular nonsense coding scheme may make a "bit of sense" in that application. On the other hand, simple binary (1 or 0) coding is much easier to learn and to interpret the outcomes. Perhaps then the major use of nonsense coding is to instill in regression users a sense of versatility in the regression methodology.

## References

Cohen, J., & Cohen, P. (1975). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Williams, J. D. (1974a). A note on contrast coding vs. dummy coding. Multiple Linear Regression Viewpoints, 4, 1-5.

Williams, J. D. (1974b). Regression analysis in educational research. New York: MSS Publishing Company.

Williams, J. D. (1976). Multiple comparisons by multiple linear regression. Multiple Linear Regression Viewpoints, 7(1).

IPLE LINEAR REGRESSION VIEWPOINTS
OLUME 15, NUMBER 2, WINTER 1987

# A General Model for Estimating and Correcting the Effects of Nonindependence in Meta-Analysis

Michael J. Strube
Washington University

Abstract

This paper describes a general meta-analysis model that can be used to represent the four types of meta-analysis commonly conducted. The model explicitly allows for nonindependence among study outcomes, providing exact statistical solutions when the nonindependence can be estimated. Also discussed are the directional biases that result if nonindependence is ignored.

## A General Model for Estimating and Correcting the Effects of Nonindependence in Meta-Analysis

Over the past several years there has been a dramatic increase in the use of meta-analytic procedures. At the same time there has been relatively little attention given to some of the problems that are encountered when traditional statistical procedures are applied to the nontraditional data bases that meta-analysts encounter (for exceptions, see Rosenthal & Rubin, 1986; Strube, 1985a; Tracz & Elmore, 1985; Tracz, Newman, & McNeil, 1986). One of the more prevalent and serious problems encountered in a meta-analysis occurs when studies give rise to multiple outcomes. In such cases, the assumption of independence is violated with potentially serious inferential consequences. To date, there has been no clear exposition of the nature or direction of bias that exists when nonindependence is ignored. The purpose of this paper is thus twofold. First, I will present a general model of nonindependence that encompasses the four major types of meta-analysis that are conducted. This model also provides an exact solution for the correction of nonindependence. Second, I will indicate the inferential consequences of ignoring nonindependence.

### A General Model for Meta-Analysis

There are four basic types of meta-analysis that are typically conducted. First, the meta-analyst may examine study outcome defined in terms of an effect size estimate (e.g., $\Delta$ $\underline{d}$, g, or $\underline{r}$) or in terms of an estimate of statistical significance (e.g., $\underline{p}$ or $\underline{Z}$). Second, within these two outcome classes, the meta-analyst can perform two basic tasks (Rosenthal, 1983) by either combining study outcomes or contrasting study outcomes. The former task represents an interest in the overall outcome whereas the latter task corresponds to a search for moderators of study outcome.

What often goes unnoticed is that the various specific statistical procedures described in the literature for carrying out these four types of meta-analysis all represent special cases of a more general approach. In particular, all can be represented as special cases of the following formula:

$$Z = \frac{\Sigma \lambda_i \psi_i}{(\Sigma\ \lambda_i^2 \sigma_i^2 + 2\Sigma \lambda_i \lambda_j \sigma_i \sigma_j \rho_{ij})^{1/2}} \qquad (i \neq j)$$

This formula represents a weighted linear combination of elements, $\psi$, divided by the standard deviation of that linear combination. When the linear combination is tested against the null mean of zero, the ratio will be approximately normally distributed for modest sample sizes. There are several things to note about the formula. First, the elements to be combined or contrasted can be either effect sizes or an index of statistical significance. Second, if $\psi = Z$, and all Z are independent, then the formula provides the familiar Stouffer solution for combined probabilities (see Strube, 1985a). Third, if $\psi$ are to be combined, then all $\lambda = 1$. Finally, if $\psi$ are to be contrasted, then $\Sigma\lambda$ must equal zero (as in ANOVA or regression). As can be seen, all four types of meta-analysis can be represented.

What makes this approach additionally useful is that it provides a means of accounting for nonindependence. As the formula and the variance-covariance matrix in Figure 1 indicate, nonindependence serves to alter the size of the standard deviation of the linear combination. Under the assumption of independence, all covariance terms are zero, and the estimate of the standard deviation of the linear combination is based solely on the main diagonal of the variance-covariance matrix (formulae for estimating the variances of several common effect sizes can be found in Hodges & Olkin, 1985; Rosenthal, 1984). Thus it is the off-diagonal elements that are of particular interest when there is nonindependence.

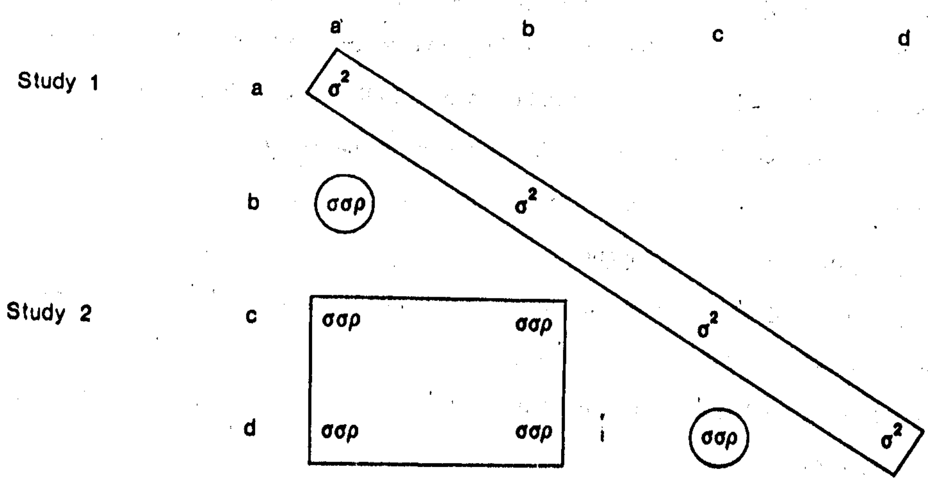



Figure 1. Variance-covariance matrix for two studies, each with two outcomes.

Nonindependence will arise in a meta-analysis whenever the same study (or subject, for N = 1 research, see Strube, 1985b) provides more than one effect size or significance level to be combined or compared. In that case, one must attempt to estimate the magnitude of the off-diagonal elements of the variance-covariance matrix (see Strube, 1985a). Actually, we need not estimate all of the off-diagonal elements. It is probably safe to assume that effect sizes and significance levels from different studie: are independent, and thus the corresponding covariances are zero. Thus, in Figure 1, the covariances in the lower left box can be assumed to be zero. Only the circled covariances need to be estimated. If reasonable estimates for these covariances can be obtained, then an exact combination or contrast is possible.

Consequences of Nonindependence

Given current reporting practices, it may be difficult to estimate the needed covariances. It is still important to recognize the type of influence that nonindependence has so that, even if it cannot be adjusted statistically, it can serve to temper one's conclusions.

Figure 2 displays four basic types of questions that could be asked in a meta-analysis, as represented by the weights ( $\lambda$ ) that would be used in our formula. We also have listed 3 studies each of which gave rise to 2 outcomes measures that we will assun are positively correlated. In the first case, all outcomes are added (a combined result i desired), that is, all $\lambda$ are positive and thus the influence of nonindependence is to infla the denominator of the formula. Accordingly, failing to adjust for nonindependence wi inflate the likelihood of a Type I error. In the second case, two studies are compared. Because the comparison is across correlated units, the influence of nonindependence is inflate the denominator of the formula (i.e., cross-product of $\lambda$s is positive). Again, failing to take nonindependence into account will inflate the Type I error rate. The th case represents a contrast where the two different outcomes within studies are

| | | Type of Contrast | | | |
|---|---|---|---|---|---|
| | | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ |
| Study 1 | A | 1 | 1 | 1 | 1 |
| | B | 1 | 1 | -1 | -1 |
| Study 2 | A | 1 | -1 | 1 | -1 |
| | B | 1 | -1 | -1 | 1 |
| Study 3 | A | 1 | 0 | 1 | 0 |
| | B | 1 | 0 | -1 | 0 |
| Type of Error Increased | | Type I | Type I | Type II | Type II |

Figure 2. Four common meta-analytic contrasts and their associated inferential errors when nonindependence is ignored.

compared. Because the comparison is within studies, the influence of the nonindependence is to decrease the denominator of the formula (all $\lambda_i$ $\lambda_j$ are negative). In this case, failing to adjust for nonindependence will inflate the Type II error rate. The final case represents a pattern of contrasts corresponding to an interaction. Here interest is in whether the difference between the two outcome measures depends on the study. Here too, the effect of unadjusted nonindependence is inflate the Type II error rate.

Thus it can be seen that the effect of nonindependence on the outcome of a met. analysis depends on the type of question being asked.

Summary

In sum, the meta-analyst must be aware of the influence of nonindependence. Where possible, the effect of nonindependence should be adjusted statistically. If this i not possible, the meta-analyst must quality conclusions, taking into account the known directional effects of nonindependence on the likelihood of making Type I and Type II errors. If nonindependence is ignored, meta-analysts may introduce stubborn and erroneous conclusions into the literature.

## References


Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. New York: Academic Press.

Rosenthal, R. (1984). Meta-analytic procedures for social research. Beverly Hills, CA: Sage.

Rosenthal, R. & Rubin, D. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. Psychological Bulletin, 99, 400-406.

Strube, M. J. (1985a). Combining and comparing significance levels from nonindependent hypothesis tests. Psychological Bulletin, 97, 334-341.

Strube, M. J. (1985b, November). The effect of nonindependence on meta-anlaysis in single subject research. In D. P. Hartmann (Chair), Meta-analysis and N = 1 methodology, Symposium presented at the 19th meeting of the Association for the Advancement of Behavior Therapy, Houston.

Tracz, S. M. & Elmore, P. B. (1985, August). The issue of nonindependence in correlational meta-analyses. Paper presented at the annual meeting of the American Statistical Association, Las Vegas, NV.

Tracz, S. M., Newman, I., & McNeil, K. (1986, April). Tests of dependence in meta-analysis using multiple linear regression. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

# The Use of Judgement Analysis and A Modified Canonical JAN in Evaluation Methodology

Samuel R. Houston
University of Georgia

ABSTRACT

Judgment Analysis is presented as a technique for capturing and clustering unidimensional policies among a group of judges or evaluators. JAN utilizes a multiple linear regression model to represent each policy and then cluster evaluators together who are expressing similar policies. JAN is extended to a multidimensional situation in which a modified and simplified Canonical JAN (C-JAN) procedure for capturing policies on more than two criteria is described. Both unidimensional and multidimensional JAN procedures should be of general interest to the evaluation methodologist.

Teacher effectiveness is an area of great concern and the focus of much research in the educational community. The idea of teacher evaluation by students has been popular at the University of Northern Colorado campus for many years. The primary purpose of this paper is to present Judgment Analysis (JAN) as a technique for both capturing and clustering policies about what constitutes teacher effectiveness for individuals serving as evaluators.

Management personnel and evaluators often base decisions upon complex arrays of information. If these administrators could state explicitly how they used this information, these decision makers--and others--could replicate their judgments in subsequent sitations in which the same types of information are available.

By way of an example, consider a situation in which an organization is in the process of recruiting personnel for particular jobs at a specific point in time. The evaluation of prospective applicants for each position is often determined by the judgment of one or more administrators, judges or decision (policy) makers. Frequently the actual rating for each applicant is obtained by combining several different types of informatin into a weighted composite to produce a numerical indicator of the decision maker's judgment or value rating. One method of weighting is to have the decision maker provide the numerical weights to be used with the different types of information (variables) to form composite explicit-weighting evaluations. While explicit-weighting procedures are satisfactory in some situations, it is usually quite difficult to choose the proper multiplier values to form the composite evaluation of the applicant for the position in question that adequtely indicate the value of a person on a job. The problem of determining the appropriate numerical weights to be used can be illustrated in the following example. In Table 1 are presented three test scores in statistics for two students. The instructor desires that each test be weighted equally in the determination of the course grade. Both students obtained the same point total of 120 points. Yet, if the instructor wants each test to carry the same weight, he must <u>not</u> add the three scores together! While each test had the same mean score, the variances for the three tests are quite different. This variation actually influences any explicit-weighting approach which might be applied. As a result of these differences, <u>different</u> weights must be applied to each test score if each test is to carry the same weight in the evaluation process.

The difficulties encountered with explicit-weighting strategies in general have led to a second method--policy-capturing--which involves implicit determination of the numerical weights to be applied.

## 1. <u>JUDGMENT ANALYSIS</u>

A technique for determing implicitly the set of numerical weights to be applied in a decision-making situation was developed by J. H. Ward, Jr.

Table 1

ASSIGNING WEIGHTS TO THREE TESTS IN STATISTICS[1]

| | Test Points | | | |
|---|---|---|---|---|
| | Test 1 | Test 2 | Test 3 | Total Points |
| Student: | | | | |
| Mary | 30 | 40 | 50 | 120 |
| Joe | 50 | 40 | 30 | 120 |
| | Z-Score | | | |
| | Test 1 | Test 2 | Test 3 | Average Z-Score |
| Student: | | | | |
| Mary | 0.00 | 1.25 | 1.67 | 0.97 |
| Joe | 5.00 | 1.25 | 0.00 | 2.08 |
| | Percentile Rank | | | |
| | Test 1 | Test 2 | Test 3 | Average Rank |
| Student: | | | | |
| Mary | 50 | 89 | 95 | 83 |
| Joe | 99 | 89 | 50 | 98 |

[1]Assume Test 1 Scores $\sim N(30, 16)$, Test 2 Scores $\sim N(30, 64)$ and Test 3 Scores $\sim N(30, 144)$.

[2]Determined for the Z-Scores.

It is called Judgment Analysis (JAN) and it involves a hierarchial grouping of data using an iterative procedure (Ward 1961, 1963; Ward and Hook 1963). While this was a cluster analysis technique, Bottenberg and Christal (1968) used this idea of hierarchial grouping to combine regression equations, using minimal loss of predictive efficiency as the grouping criterion.

Originally, JAN was developed to solve problems faced by the Personnel Department of the Air Force (Christal 1968a; Bottenberg and Christal 1968).

## 2. POLICY-SPECIFYING AND POLICY-DEVELOPMENT WEIGHTS IN JAN

### Weights

Policy-capturing requires a set of judgments (Y values) associated with n decision situations to obtain the implicit weights. However, in the policy-specifying process, the weights are determined without empirically obtained judgments (Y values) by stating desired properties of and relations among the predicted values in sufficient detail that the numerical weights become known.

Specifically let

$b_j$ = the unknown weights to be determined by policy-specifying (corresponding to $a_j$ in policy-capturing above). $j = 1, \ldots, k$

$b_o$ = an unknown constant (corresponding to $a_o$)

$x_j$ = variables corresponding to the predictor vectors above. These are not vectors of data but are variables which when given a set of weights $b_j$ and $b_o$ and a set of values for $x_j$ will yield a composite value y.

Then we have the starting function

$$y = b_o + b_1x_1 + b_2x_2 + \ldots + b_jx_j + \ldots + b_kx_k$$

Prior to the policy-specifying process, the range of values for $x_1, x_2, \ldots x_k$ are known but the $b_j$ and $b_o$ values are not known. Policy-specifying proceeds by stating restrictive relations among the predicted values for various values of $x_j$. These policy statements result in restrictions on the values of $b_j$ and $b_o$ so that the numerical values of the weights can be determined. Specification is completed when k + 1 independent restrictions are imposed. Once the values of $b_j$ and $b_o$ are known, then predicted values, y, can be calculated for any values $x_j$.

Policy-capturing and policy-specifying can be combined to form a general process of policy-development. A particular decision maker may start by specifying several properties about relations among the predicted values. Whereas policy-specifying resulted in k + 1 restrictions on the k + 1 weights, $b_j$ and $b_o$, the expressin of desired properties may result in only r k + 1 restructions on the $b_j$ and $b_o$ values.

Then imposing these r restrictions on the starting model results in a restricted model

$$y_r = c_o + c_1z_1 + c_2z_2 + \ldots + c_jz_j + \ldots + c_{k-r}z_{k-r}$$

where

$z_i$ = new variables resulting from imposing the r restrictions.

Each $z_i$ variable is a linear combination of the $x_i$ variables. Now since there are still k + 1 - r unknown weights $c_j$ and $c_o$ to be computed it would be possible to use policy-capturing to find the $c_j$ values. The decision maker could provide, for each of the n[n (k + 1 - r)] decision situations, $y_i$ (i = 1,...,n) values associated with various profiles of information about the different situations. Then the least squares values of $c_j$ can be computed for the model.

$$Y = c_o U + c_1 Z^{(1)} + c_2 Z^{(2)} + \ldots + c_j Z^{(j)} + \ldots + c_{k-r} Z^{(k-r)} + E^{(2)}$$

where

Y = a vector of judged values of dimension n.

$Z^{(j)}$ = the $j^{th}$ predictor vector, of dimension n formed as linear combinations of the predictor vectors $X^{(j)}$ generated from information associated with the decision situations.

Having computed the least squares values for $c_j$ and $c_o$ the weighting system now produces values that both reflect the policy restrictions imposed by the policy-specifying process and the best fit to the empirical judgments.

## 3. GENERAL APPLICATIONS OF JAN

JAN has been used in several studies conducted by the U.S. Air Force for job evaluations and to stimulate officer promotion boards with a high degree of efficiency. Equations have also been designed to simulate career counselors in making initial assignments of airmen graduating from basic training (Dudycha, 1970).

The JAN technique has been applied in a prediction study of success in graduate education. In a study by Houston (1967) two variations of JAN were investigated--Normative JAN and Ipsative JAN. The purpose of the Normative JAN study was to determine the extent to which a policy regarding graduate admission standards existed among selected graduate faculty members at Colorado State College (now University of Northern Colorado). Basically, three sets of independent profile variables were used: (1) biographical dat (2) test data, and (3) major subject field data. Results from the Normative Jan study indicated essentially one policy was present in the group of judge

The Ipsative JAN study used for its dependent variable the rankings submitted by the judges who were requested to rank, without access to the three sets of independent profile variables used in the Normative JAN study, the doctoral graduates on a basis of personal knowledge. It was the intent this phase that the ratings or rankings be loaded with personality factors readily available in the Normative JAN study. Results of this phase were tatistically significant, though weak from the predictive standpoint. The ractical significance of the Ipsative JAN study was in the suggestion of new irections for subsequent research.

Williams, Gab, and Linden (1969) replicated Houston's Normative study at the University of North Dakota and sought to determine the policy of a university doctoral admissions board. Twelve members of the graduate faculty evaluated each graduate student's profile and place it into one of seven criterion categories (Q-sort). Each rater's policy was assessed or captured and the raters were grouped into appropriate clusters by the JAN process. The investigators found that at least two separate judgmental systems were present.

A further illustration of the versatility of the technique is provided in a study by Stock (1969) who sought to determine if systematic differences existed in the placement policies for special education students among special education personnel (teachers, administrators, and the members of the special education screening committee) responsible for placing the students in the public schools of Cheyenne, Wyoming. Colvert (1970) used JAN techniques in the identification and analysis of the consultant ratings of elementary student teachers at the University of Northern Colorado. Using JAN procedures, Chang (1970) designed a study to determine whether individuals serving in different official capacities in the State of Colorado had differing attitudes toward selection criteria for awarding college financial grants. Keelan et al. (1973) captured the leadership policies of selected fireman in the State of Colorado with the use of JAN.

The question of what is pornographic was investigated by J. Houston and S. Houston (1974) who used JAN as a methodology by testing this technique with three groups concerned wit this issue. These groups included doctoral students majoring in Psychology, Counseling and Guidance at the University of Northern Colorado, lawyers and police officers from the city of Greeley, Colorado. The JAN technique proved to be surprisingly effective in capturing and clustering the policies (specific and complex) of the judges from the three groups identified. As expected, many policies were present.

The problem of evaluating curriculum packages was explored by Torgunrud (1971) in a doctoral dissertation completed at the University of California at Lost Angeles under the direction of Dean John I. Goodlad. Torgunrud identified from the educational literature the following independent variables as important dimensions of any curriculum package or set of materials which are under consideration for possible adoption. These include: (1) valid and significant content, (2) significant elements of organization, (3) sequence providing a cumulative effect, (4) integration providing horizontal relationships, (5) value position clearly stated, (6) specificity providing direction, (7) flexibility providing alternatives, (8) accommodation for student participation, and (11) provision for measurement of achievement. After defining the variables, Torgunrud generated a sample of 100 profiles, each described on the 11 variables, by using techniques described by Naylor and Wheery (1965) for simulating stimuli with specified factor structure.

In another evaluation at the University of California at Los Angeles, Duff (1969) utilized JAN techniques to capture both the teacher-hiring policies (Ex Ante) of selected administrators and the administrators' evaluation policies (Ex Post) of teachers' on-the-job performance after their first year of paid teaching experience. Both types of policies (hiring and job performance) were analyzed for elements of predictive validity by the investigator.

The effectiveness of JAN in capturing and clustering raters' policies was investigated by Dudycha (1970) in a Monte Carlo evaluation of JAN as a methdology. Dudycha's outcomes show that the grouping process begins to break down when there are fewer than 200 stimuli being evaluated or 100 if ten or more stimulus dimensions are used. Consequently, the researcher using JAN must be concerned with the number of stimulus dimensions used in a relationships to the stimuli being evaluated. It is the present recommendation of the writer that a minimum of 100 stimuli be available for each judge on a maximum of 10 stimulus dimensions.

Other examples using Ipsative JAN are Christal (1968b) in which the researchers had to use their own knowledge to discover the variables being used by the single judge, and Holmes and Zedeck (1973) in which the judges were asked to judge paintings and also to relate qualities which the paintings exhibited. These qualities were then used to develop characteristics used as the predictors in the linear mathematical policy model. A Normative study using these characteristics followed.

The type of JAN used in a study can be further specified. Type A JAN would be used if the judges were dealing with the same subjects or profiles. Type B JAN designates a situation in which the judges each are making judgments on a different set of subjects or profiles.

Traditionally, JAN problems have involved predictors having a continuous distribution and have had dependent variables which were either ranked or categorical. It was demonstrated by Houston and Bolding (1974) that JAN is a special case of the general linear model. Because of this, any type of variable which could be used in a linear model could be used in JAN. Sets of non-redundant, dummy variables, for instance, can be used for the categories (Suits 1957). An example of this can be found in Christal (1968b) in which some of the variables were categorical.

Certain issues associated with the use of JAN have been debated (Houston 1974b). It has been suggested that a distribution be specified a priori for the judges to use. A second issue raised by statisticians was how many predictors (independent variables) should be used. Statistical studies have shown that ten should be the minimum. Practical considerations have suggested between five and seven. A third issue was the number of Ss to be given to each judge. Statistical studies employing Monte Carlo techniques have shown that a minimum of 200 should be used. Practical considerations indicate that between 30 and 60 profiles should be used in a policy-capturing situation. Another issue debated is whether a test of significance or a practical test should be used. Regression is a large sample procedure. Tests of significance useful in JAN (t and F) are designed to be powerful when samples are small with increasing power as the sample size increases. With a large sample size even the smallest decrease in predictability can be significant. Ward and Hook (1963) recommended looking for a break in the pattern of $R^2$ (RSQ) value decreases between stages in the analysis. Houston and Gilpin (1971) suggested a modification of this technique. They recommended establishing *a priori* the maximum decrease in predictability which the researcher would allow before considering the decrease to be meaningful. They suggested a .05 level as a general "rule of thumb".

JAN has been widely used as a policy-capturing procedure in the military. Some examples of military policy-capturing applications have been described in the following publications: Black (1973); Christal (1968a, 1968b); Gott (1974); Gooch (1972); Jones, Mannis, Martin, Summers, and Wagner (1976); Koplyay (1970); Koplyay, Albert, and Black (1976); Mullins and Usdin (1970); Ward and Davis (1963).

## 4. STUDENT POLICIES OF TEACHER EFFECTIVENESS

The student judgmental policies of teacher effectiveness were analyzed in study completed by Houston and Gilpin (1971).

Procedures. The primary problem of the investigation was to analyze the results of a teacher description study and to identify judgmental policies of selected subsets of students at the University of Northern Colorado. The subjects for which profile and judgment scores were generated were faculty members of the University of Northern Colorado.

The judges. Students rated the teachers using the criteria represented in Instrument One. For purposes of this study, the students were grouped into selected subsets. The first grouping was made by schools or colleges within the university and resulted in seven subsets or groups of students. The researcher treated each of the individual groups as a judge in the first JAN investigation. The second grouping of students was determined by grade level and allowed for five subsets of students ranging from freshman through graduate level. Each of these distinct groups was treated as an individual judge in the second JAN analysis. Therefore, in the JAN analyses, a slight innovation was used. In the usual JAN a judge is an individual; however, in this study the individuals were grouped into subsets and each subset, consisting of numerous individuals, was considered a judge.

The instrument. The student raters were requested to rank teachers on the first 9 items and to provide biographical information asked for in item 10 of the following instrument:

## Teacher Description Instrument (Instrument One)

Please rate only this teacher in this particular course in accordance with this rating scale. 1) Poor 2) Fair 3) Average 4) Good 5) Excellent

| | | | | | | |
|---|---|---|---|---|---|---|
| 1. | Teacher's interest and enthusiasm for course | 1 | 2 | 3 | 4 | 5 |
| 2. | Ability to adequately answer questions | 1 | 2 | 3 | 4 | 5 |
| 3 | Ability to communicate the subject matter effectively | 1 | 2 | 3 | 4 | 5 |
| 4. | Ability to interest and motivate students | 1 | 2 | 3 | 4 | 5 |
| 5. | Fairness in testing and grading | 1 | 2 | 3 | 4 | 5 |
| 6. | Personal interest and adaptation to student's needs | 1 | 2 | 3 | 4 | 5 |
| 7. | Course objectives are clearly stated | 1 | 2 | 3 | 4 | 5 |
| 8. | Course objectives are met | 1 | 2 | 3 | 4 | 5 |
| 9. | Everything considered, including strengths and weaknesses, I would rate the instructor | 1 | 2 | 3 | 4 | 5 |
| 10. | 1) Freshman 2) Sophomore 3) Junior 4) Senior 5) Grad | | | | | |

The first eight items of Instrument One were considered independent variables while item nine was treated as the dependent variable in multiple linear regression analyses. Responses to the first eight variables were also used as profile scores, and responses to item nine as judgments in the two JAN analyses.

<u>JAN techniques.</u> The JAN technique starts with the assumption that each judge has an individual policy. It gives and $R^2$ (multiple R coefficient squared) for each individual judge and an overall $R^2$ for the initial stage consisting of all the judges, and each one treated as an individual system. Two policies are selected and combined on the basis of having the most homogeneous prediction equations, therefore resulting in the least possible loss in predictive efficiency. This selection reduces the number of original policies by one and gives a new $R^2$ for this stage. The loss in predictive efficiency can be measured by finding the drop in $R^2$ between the two stages. The grouping procedure continues, reducing the number of policies by one at each stage, until finally all of the judges have been clustered into a single group.

Investigators examined the collective drop in $R^2$ from that of the original stage in each of the two JAN analyses. A determination of whether one or more policies were present among the judges was made on the basis of the sequential drop in $R^2$. A slippage greater than .05 was considered *a priori* to represent too great a loss in predictability.

### Findings

The first JAN analysis considered the students grouped into the seven schools and/or colleges of the University of Northern Colorado. Each group was treated in the analysis as an individual judge. A listing and abbreviation of the variables for this study are found in Table 2.

Stages of the JAN procedure for judges by school and/or colleges. The $R^2$s for each of the seven initial systems are reported in Table 3. Note that the magnitudes of $R^2$ are restricted in range. The highest value is .8309 for judge four and lowest is .7443 for judge seven. These high values of $R^2$ for all judges indicated that the judges were consistent in their individual decision-making policies.

Table 4 reports the seven stages of the JAN clustering procedure for the seven judges and the corresponding $R^2$ for each stage. In stage 2, judges two and three have been combined to form one group while all other judges are treated individually. The drop in $R^2$ between stages 1 and 2 is only .004. Continuing this clustering procedure, stage 3 combined judges five and six resulting in a model consisting of five policies or systems. The resulting drop in $R^2$ from stage 1 is .0009.

Stage 7 combined all seven judges into one cluster and resulted in a collective drop in $R^2$ of only .0248. The *a priori* criterion for permissible slippage in $R^2$ was .05. Since the collective drop of .0248 is well within this tolerance level, stage 7 was accepted as the appropriate grouping of judges. Therefore, the investigators concluded that only one policy was present among the seven judges.

Policy of the seven judges. Interpretation of the JAN procedure determined that only one policy existed among the seven judges representing the schools and/or colleges. Regression analysis was then employed in an effort to explain that policy.

The investigators were interested in determining the unique contribution of proper subsets of the predictor variables, 1 through 8, to the prediction of the criterion, GenP. The contribution of a set of variables to prediction may be measured by the difference between the $R^2$ for the full model (FM) and the $R^2$ for a restricted model (RM). The RM differs from the FM in that the proper subset of variables, for which the unique contribution to predictability is desired, have been deleted. The difference between the two $R^2$s may be tested for statistical significance through use of an F test or else an *a priori* acceptable drop can be established. The investigators chose the latter alternative and set a drop tolerance of .05. That is, if $R^2_F - R^2_F$ .05, the investigators concluded that the subset under consideration was making a unique contribution to prediction of the criterion.

A subjective hierarchy of the variables is presented in Table 5. This grouping was used in the regression analysis of the different policies.

Figure 1 presents a schematic to guide the sequence of tests from the FM through the various restricted models. The accompanying $R^2$ for each of these models is found in the appropriate block. For example, the information in block 1 indicates that the independent variables 1 through 8 were used as the predictors in the FM and that the $R^2$ for this model was .8123.

Block 2 displays FM - (5,6,7,8), indicating that variables (5,6,7,8) have been deleted from the full model. This also implies that variables 1, 2, 3, and 4 are used as the predictor variables in the RM. By dropping out variables (5,6,7,8), the unique contribution to prediction of these variables can be determined. The measure of this unique contribution was found by the difference between the $R^2$ = .8123 for the FM and the $R^2$ = .7742 for this RM. The difference .8123 - .7742 = .0381 was less than .05 and therefore indicated that these variables were making little or no contribution to prediction that could not be explained by the other four predictor variables. Since the drop in $R^2$ for this set was not significant, no further tests of subsets of these variables were necessary. The broken line in the chart indicates that further testing of subsets of variables (5,6,7,8) was terminated.

The expression in block 3, FM - (1,2,3,4), indicates that variables (1,2,3,4) were eliminated from the FM. These predictors were grouped on the subjective basis that they were related and measured a general hypothetical category called methodology. The drop .8123 - .6673 = .1450 was greater than .05 and therefore resulted in too great a loss in predictive efficiency. Therefore, further analysis of subsets of these variables was undertaken. However the $R^2$ for the model FM - (1,4) was .7768. Since the drop of .0335 was less than .05, variables (1,4) made no significant contribution to prediction of the criterion. An examination of the subset represented by the model FM - (2,3) showed that the drop in $R^2$ was equal to .0376. Again the drop was less than .05, and it was concluded that variables (2,3) made an insufficient unique contribution to the prediction of the criterion. Multicollinearity of the variables (1,2,3,4) accounted for the fact that no significant drop in $R^2$ was detected when further analysis of the branchings from this set were examined. That is, the variables in this set are highly intercorrelated, and when two of them are eliminated, the presence of the other two in the FM hold up the value of $R^2$. The broken line again indicates that further examination of subsets of these variables was not needed.

In summary, the eight predictor variables were very efficient in predicting the criterion since the $R^2$ was reported to be .8123. The model FM - (5,6,7,8) also had high prediction efficiency with an $R^2$ = .7742. Therefore, all of the judges who were clustered into the only policy-making system were attending to variables 1, 2, 3 and 4 when they were rating teachers in the general overall category.

As reported, the grouping of subsets of the eight predictor variables was a completely subjective determination. The investigators were interested in analyzing Table 6, the intercorrelations of predictors and the validities, to determine if a different hierarchy of variables would result. Perhaps a smaller subset of variables making a unique contribution to prediction could be found if the subsets were grouped differently.

The validities were comparatively high, ranging from .604 to a high of .804. The investigators grouped the predictors into a hierarchy base upon the correlations. This grouping is presented in Table 7.

The schematic sequence of tests is presented in Figure 2. The branching eading from block 2 was terminated in view of the resulting $R^2$ = .7848 for he model FM - (1,5,7,8). This represented a drop of only .0275, well within he .05 level. Of considerable interest was the alternate branching leading o and from block 3. The model FM - (2,3,4,6) yielded a significant drop in $^2$ of .8123 - .6758 = .1365. This prompted further investigation of subsets f this model. The model FM - (2,6) accounted for a drop of only .8123 - .7939 = .0184, and hence further investigation of subsequent branching was nded. However, the model FM - (3,4,6) was of extreme interest in view of the significant drop in $R^2$ of .8123 - .7248 = .0875. Consequently further ranching from this model was investigated. The model FM - (3,4) was also found to make a unique contribution since the drop of .8123 - .7558 = .0565. Further analysis of the unique contribution of variables 3 and 4, treated individually, resulted in nonsignificant findings. The reason for this finding was that variables 3 and 4 were highly related $r_{3,4}$ = .75.

The regression analysis based on correlations (Table 7) allowed for a more refined interpretation than did the analysis based on subjectivity. The hierarchy suggested by the correlations led not only to a set of three variables (3,4,6) making a unique contribution, but also to a set of only two predictors (3,4) making a unique contribution to prediction.

An interesting question arose at this juncture. The two sets of variables (3,4,6) and (3,4) both make unique contributions, but what about their absolute or total prediction? This information is not available from the sequence of tests in Figure 2. The researchers investigated the predictive efficiency of the FM models consisting of the set of variables (3,4,6) and (3,4). The $R^2$ for the FM consisting of variables (3,4,6) was equal to .7678. The difference was .8123 - .7678 = .0445 which, by virtue of the .05 convention used in this study, implied that this RM predicted as well as did the FM. However, the RM consisting of variables 3 and 4 had an $R^2$ = .7340 which obviously was not as efficient as was the FM.

<u>JAN by grade level.</u> The second JAN analysis grouped students according to grade level. Each of the five levels was considered as a judge. Table 8 shows the $R^2$s associated with the prediction equation for each of the five judges. The $R^2$s ranged in value from .7988 for freshmen to .8344 for seniors. The high $R^2$s indicated efficient prediction for each of the respective regression or decision-making equations.

The five stages of the JAN grouping technique are presented in Table 9. As conjectured from observation of the preliminary statistics, the collective drop in $R^2$ from the original stage to stage 5 was somewhat less than the .05 limit.

Stage 2 combined the freshmen and sophomores, leaving the juniors, seniors and graduates as the three single-member systems. This combination resulted in an $R^2$ slippage of only .002. Stage 3 clustered the juniors and seniors leaving the graduate students as the only singleton set. The collective drop in $R^2$ at this stage was a nearly indiscernible .0005. Stage 4 combined the sets containing two judges each into a cluster of four, again leaving judge five as the only single-member system. At this stage the overall drop in $R^2$ was an inconsequential .0015. Stage 5 grouped all of the judges into one decision-making system and resulted in a total $R^2$ slippage of only .003. Certainly this drop in $R^2$ was well within the tolerance range of .05. These data suggest that only one juogmental policy was existent among the five judges.

TABLE 2
List of Variables and Abbreviations

| Number | Variable | Abbr. |
|---|---|---|
| 1. | Teacher's interest and enthusiasm for course | IEth |
| 2. | Ability to adequately answer questions | AnsQ |
| 3. | Ability to communicate subject matter effectively | CSub |
| 4. | Ability to interest and motivate students | MoSt |
| 5. | Fairness in testing and grading | TeCr |
| 6. | Personal interest and adaptation to student's needs | SNds |
| 7. | Course objectives are clearly stated | CObS |
| 8. | Course objectives are met | CObM |
| 9. | General rating (criterion) | GenR |

TABLE 3
$R^2$ Values for All Judges from Regression Models

| | Judge | $R^2$ |
|---|---|---|
| 1. | School of the Arts | .7869 |
| 2. | College of Arts and Sciences | .8126 |
| 3. | School of Business | .7764 |
| 4. | College of Education | .8309 |
| 5. | School of Health, Physical Education, and Recreation | .7992 |
| 6. | School of Music | .8075 |
| 7. | School of Nursing | .7443 |

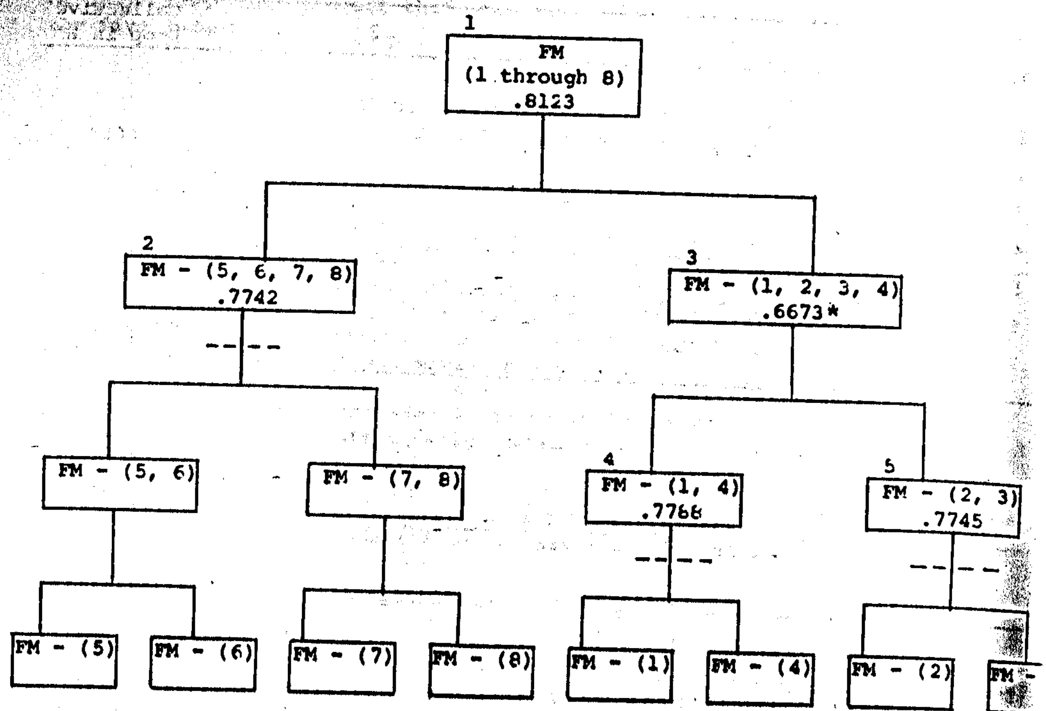
TABLE 4
Stages of the JAN Procedure for the Seven Judges

| Stage | Judges | $R^2$ | Collective Drop in $R^2$ |
|---|---|---|---|
| 1 | 1, 2, 3, 4, 5, 6, 7 | .8141 | |
| 2 | (2, 3), 1, 4, 5, 6, 7 | .8137 | .0004 |
| 3 | (2, 3), (5, 6), 1, 4, 7 | .8132 | .0009 |
| 4 | (1, 4), (2, 3), (5, 6), 7 | .8121 | .0019 |
| 5 | (1, 4), (2, 3, 7), (5, 6) | .8099 | .0042 |
| 6 | (1, 4, 2, 3, 7), (5, 6) | .8064 | .0077 |
| 7 | (1, 4, 2, 3, 7, 5, 6) | .7893 | .0248 |

TABLE 5
Subjective Hierarchy of Variables

| | |
|---|---|
| Methodology: | |
| Teacher's interest and enthusiasm for course | (1) |
| Ability to interest and motivate students | (4) |
| Ability to adequately answer questions | (2) |
| Ability to communicate subject matter effectively | (3) |
| Humanistic: | |
| Fairness in testing and grading | (5) |
| Personal interest and adaptation to student's needs | (6) |
| Organizational: | |
| Course objectives are clearly stated | (7) |
| Course objectives are met | (8) |

FIGURE 1
Seven Judged (Subjective Hierarchy)




*Significant drop in $R^2$.

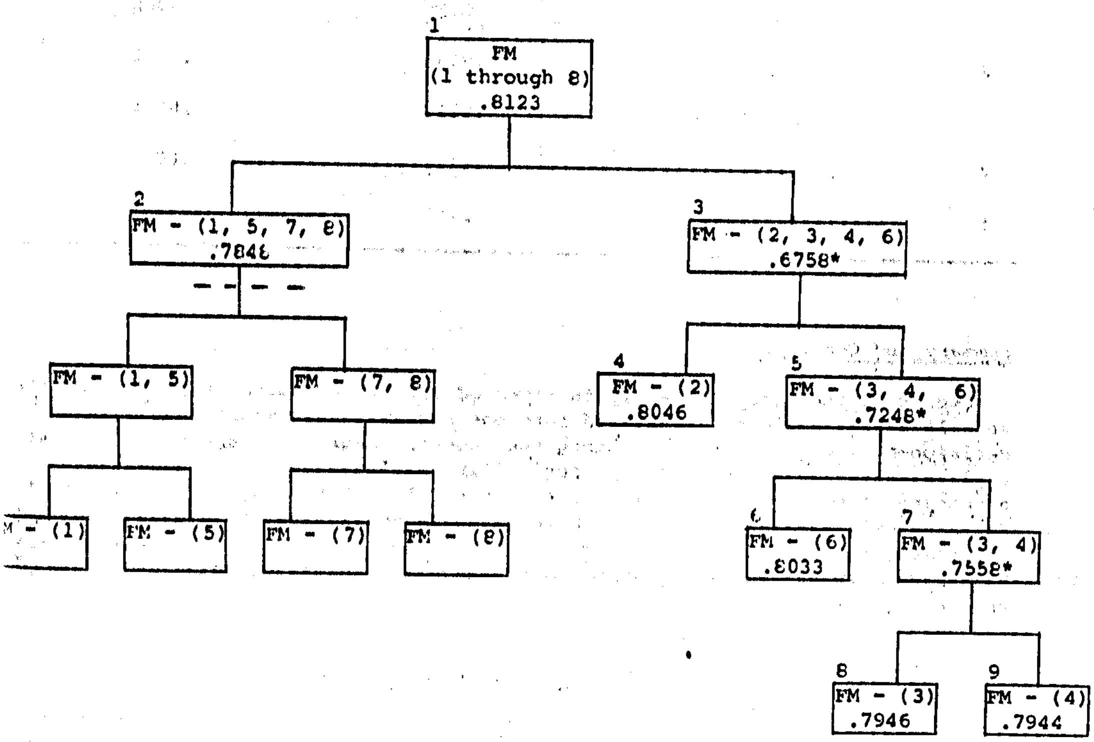
TABLE 6
Correlations of Predictor and Criterion Variables

| Variable | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 1. | IEth | | | | | | | | |
| 2. | AnsQ | .580 | | | | | | | |
| 3. | CSub | .606 | .696 | | | | | | |
| 4. | MoSt | .646 | .621 | .746 | | | | | |
| 5. | TeGr | .426 | .471 | .492 | .522 | | | | |
| 6. | Snds | .558 | .566 | .613 | .688 | .582 | | | |
| 7. | CObS | .477 | .507 | .580 | .550 | .467 | .532 | | |
| 8. | CObM | .532 | .564 | .633 | .618 | .510 | .578 | .794 | |
| 9. | GenR | .688 | .715 | .716 | .804 | .604 | .728 | .623 | .699 |

TABLE 7
Hierarchy of Variables Based on Correlations

| | |
|---|---|
| Subset 1 | |
| Sub-subsets: | |
| Ability to interest and motivate students | (4) |
| Ability to communicate subject matter effectively | (3) |
| Personal interest and adaptation to student's needs | (6) |
| Ability to adequately answer questions | (2) |
| Subset 2 | |
| Sub-subsets: | |
| Course objectives are met | (8) |
| Teachers interest and enthusiasm for course | (1) |
| Course objectives are clearly stated | (7) |
| Fairness in testing and grading | (5) |

FIGURE 2
Seven Judges (Hierarchy Based on Correlations)




*Significant drop in $R^2$.

TABLE 8

$R^2$ Values for All Judges from Regression Models

| | Judges | $R^2$ |
|---|---|---|
| 1. | Freshmen | .7988 |
| 2. | Sophomores | .7954 |
| 3. | Juniors | .8165 |
| 4. | Seniors | .8344 |
| 5. | Graduates | .8276 |

TABLE 9

Stages of the JAN Procedure for the Five Judges

| Stage | Judges | $R^2$ | Collective Drop in $R^2$ |
|---|---|---|---|
| 1 | 1, 2, 3, 4, 5 | .8136 | .0000 |
| 2 | (1, 2), 3, 4, 5 | .8134 | .0002 |
| 3 | (1, 2), (3, 4), 5 | .8131 | .0005 |
| 4 | (1, 2, 3, 4), 5 | .8121 | .0015 |
| 5 | (1, 2, 3, 4, 5) | .8106 | .0030 |

## Summary and Conclusions

Results of the first JAN analysis revealed the seven judges, representing the schools and/or colleges, clustered into one system. This meant that only one decision-making policy existed among the judges. Regression analysis was used t explain this single judgmental policy and it was found that the judges were attending primarily to variables 3, 4, and 6. An interesting finding was that t RM using only variables 3, 4, and 6 resulted in predictive efficiency significan equivalent to that of the FM. Judges representing the five grade levels were al clustered into one system as a result of the hierarchical grouping procedure of second JAN analysis.

## 5. EVALUATING THE EVALUATORS VIA JAN

What is now presented is an application of JAN to indicate how it might be use evaluate evaluators.

The League of Cooperating Schools (LCS) was launched in May 1966, as a 5-year project to study and promote planned change in American education. It was sponsored by a partnership of the University of California at Los Angeles, the Institute of Development of Educational Activities, Inc., and eighteen independent school districts in Southern California. Each school district contributed one League school and these districts ranged in size from the massive Los Angeles City system to a small district of only three schools. The districts and schools were selected in such a way as to represent, hopefully, a true microcosm of American elementary schools. It was the aim of this joint enterprise to develop a cohesive program of research, development, innovation, and dissemination of information in order to narrow the chasm between current educational theory and practice.

In order to effect educational change, a rationale was needed that would serve as a basis for research design while at the same time serving the interests of the cooperating schools. The result was the creation of a new social system in which principals and teachers in the LCS were to be challenged by I/D/E/A to fashion new norms, roles, supports and rewards for themselves.

Four members of the Intervention Staff were requested to score on a 5-point scale each of eighteen schools on eight characteristics deemed essential for effective schools. A list of these characteristics with explanations appears in Table 10 (variables 1-8). In addition, the Intervention Staff members were asked to rank the eighteen schools in terms of overall effectiveness. The rankings were used as the criterion variable in the JAN process. This procedure represents a slight modification of the usual JAN procedure in that the judges generated their own profiles by the scores they gave on variables 1-8.

In Table 11 appears the intercorrelations between all the variables. The means and standard deviations are presented in Table 12. A multiple linear regression equation was developed for each Intervention Staff member who served as judge. Table 13 contains the correlations of each predictor variable and the criterion variable (school rank). Also included for each rater is his multiple correlation coefficient.

Table 14 summarizes intercorrelations of judgmental policies. It appears that judges 3 and 4 have the most homogeneous policy as the correlation coefficient rating their rankings of effective schools is 0.90. This is borne out in Table 15 which gives the stage values for the JAN technique. In Stage 2, two groups have been formed and judges 3 and 4 have been first to be grouped. The investigators conclude that there are essentially two policies present. The justification for this stems from the fact that the collective drop in $R^2$ from Stage 1 to Stage 3 is just 0.0361 while the drop from Stage 3 to Stage 4 results in a loss of 0.0678 making the collective drop 0.1060. From Table 15 one can see in Stage 3 that judges 1 and 2 comprise one policy group while judges 3 and 4 form the second policy group.

In analyzing the policies one might wish to refer to Table 13 which reports the correlations between the school characteristics and judges. However, one finds a distressing situation in that all the intercorrelations are high. This means that the judges may have been guilty of the "halo effect" as they generated their profile scores for the eighteen schools.

The investigators were interested in determining the unique contribution of proper subsets of the predictor variables, 1-8, to the prediction of the criterion, JANCr, in both policies to compensate for multicollinearity.

For an explanation of the two judgmental policies, the investigators first made a subjective analysis of the predictors and conjectured that they formed a hierarchical pattern as displayed in Table 16.

Presented in Table 17 is a schematic to guide the sequence of tests associated with the single policy of Judges 1 and 2.

In summary the eight predictor variables were very efficient in predicting the criterion since the $R^2$ was reported to be 0.8672. Policy I as expressed by Judges 1 and 2 could basically be explained as a concern for the competence of the professional team (variables 1, 2, and 3).

In Table 18 appears a schematic which illustrates the second policy, namely the of judges 3 and 4. From blocks 2, 3, and 4, it can be seen that each of the three subsets in the subjective hierarchy was making a significant unique contribution to predicting the criterion.

TABLE 10
List of Variables

| Number | Variable | Abbr. |
|---|---|---|
| 1. | Extent professional team (principal and teachers) shows enthusiasm about their school program | PEnt |
| 2. | Extent professional team is action-oriented; i.e., they put their ideas into practice | PAct |
| 3. | Extent professional team is inquiring and searching intullecutally and self-critical | PInq |
| 4. | Extent children are involved in educational activity (can observe and talk to children) | CInv |
| 5. | Extent teacher concerns are with each child as an individual. (One can gain information from children, teachers, or parents.) | TChC |
| 6. | Extent the district supports and shows pride in the school program | DSup |
| 7. | Extent of community support (the program is supported by participation in school life, publicity, etc.) | CSup |
| 8. | The quality of the educational program vis-a-vis individualization of instruction is evident (alternatives, conferences, different grouping procedures, etc.) | QEdPr |
| 9. | JAN criterion--rank of school | JANCr |

TABLE 11
Intercorrelations

| Variable | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| PEnt | 1 | | | | | | | | |
| PAct | 2 | .83 | | | | | | | |
| PInq | 3 | .56 | .79 | | | | | | |
| CInv | 4 | .66 | .71 | .71 | | | | | |
| TChC | 5 | .70 | .74 | .72 | .74 | | | | |
| DSup | 6 | .80 | .60 | .64 | .73 | .60 | | | |
| CSup | 7 | .74 | .76 | .84 | .77 | .77 | .67 | | |
| CEdPr | 8 | .58 | .66 | .65 | .79 | .73 | .46 | .67 | |
| JANCr | 9 | .57 | .74 | .82 | .75 | .71 | .56 | .59 | .71 |

TABLE 12
Means and Standard Deviations (N = 18)

| Variable | | Mean | Standard Deviation |
|---|---|---|---|
| 1 | PEnt | 2.333 | .594 |
| 2 | PAct | 1.944 | .872 |
| 3 | PInq | 1.722 | .826 |
| 4 | CInv | 1.388 | .698 |
| 5 | TChC | 1.833 | .707 |
| 6 | DSup | 1.777 | .878 |
| 7 | CSup | 1.611 | .650 |
| 8 | CEdPr | 1.666 | .686 |
| 9 | JANCr | 9.500 | 5.338 |

TABLE 13
Correlations Between Judges and School Characteristics

| Judge | School Characteristics | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | PEnt | PAct | PInq | GInv | TChC | DSup | CSup | CEdPr | R |
| 1 | 0.56 | 0.74 | 0.82 | 0.75 | 0.71 | 0.56 | 0.59 | 0.71 | 0.95 |
| 2 | 0.57 | 0.59 | 0.62 | 0.77 | 0.69 | 0.63 | 0.63 | 0.59 | 0.81 |
| 3 | 0.87 | 0.88 | 0.69 | 0.77 | 0.83 | 0.66 | 0.76 | 0.63 | 0.94 |
| 4 | 0.85 | 0.85 | 0.71 | 0.73 | 0.80 | 0.69 | 0.82 | 0.69 | 0.93 |

TABLE 14
Intercorrelations of Judges

| Judge | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 1.00 | 0.68 | 0.71 | 0.63 |
| 2 | 0.68 | 1.00 | 0.69 | 0.66 |
| 3 | 0.71 | 0.69 | 1.00 | 0.90 |
| 4 | 0.63 | 0.66 | 0.90 | 1.00 |

TABLE 15
Stages of the JAN Procedure

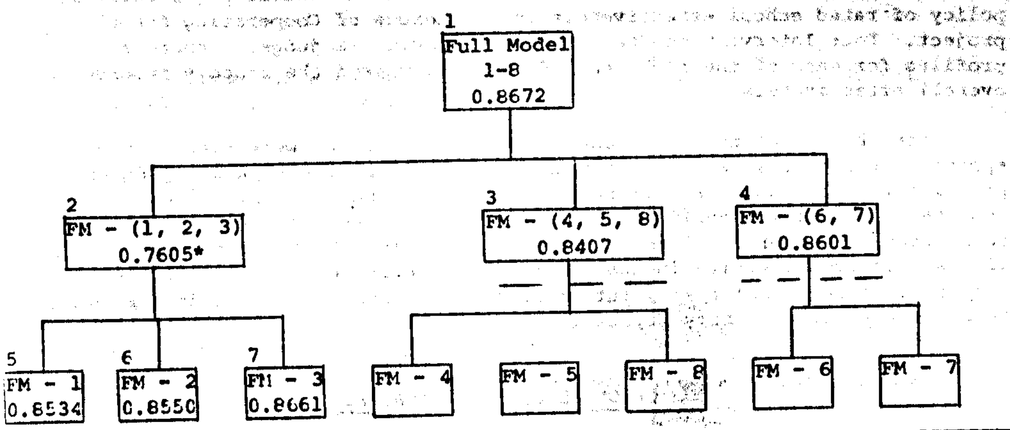
| Stage | Judges | $R^2$ | Collectiv[e] Drop in [R²] |
|---|---|---|---|
| 1 | 1,2,3,4 | .8302 | |
| 2 | (3,4), 1,2 | .8222 | .0080 |
| 3 | (3,4), (1,2) | .7921 | .0381 |
| 4 | (1,2,3,4) | .7242 | .1060 |

TABLE 16
Subjective Hierarchy of Variables

| | | |
|---|---|---|
| Professional staff competence: | Extent professional team is enthusiastic | (1) |
| | Extent professional team is action-oriented | (2) |
| | Extent professional team is inquiring and self-critical | (3) |
| Concern for children: | Extent children are involved in educational activity | (4) |
| | Extent teacher concerns are with child as individual | (5) |
| | Extent of individualized instruction | (8) |
| Outside support: | Extent of district support | (6) |
| | Extent of community support | (7) |

TABLE 17

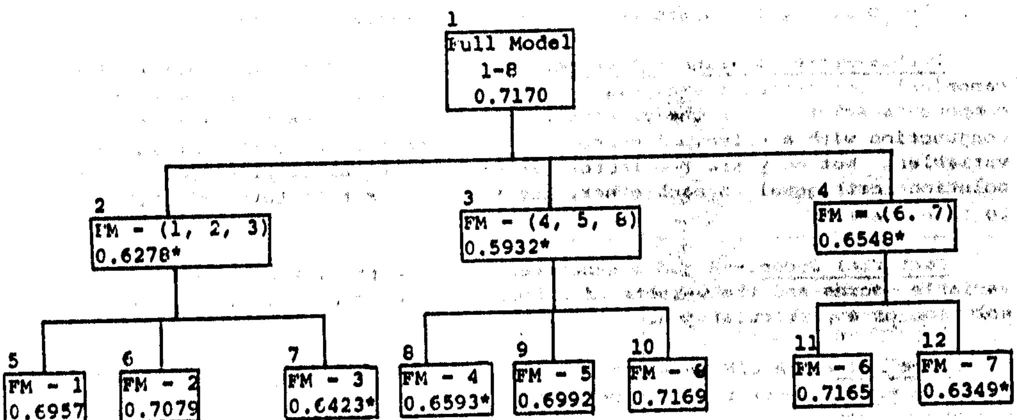
Flowchart of Regression Analysis of Policy I (Judges 1 and 2)




*Significant drop in $R^2$.

TABLE 18

Flowchart of Regression Analysis of Policy II (Judges 3 and 4)




*Significant drop in $R^2$.

In summary, the eight predictor variables were efficient in predicting the criterion for judges 3 and 4, though not as efficient as in Policy I. Policy II differed from Policy I in that each of the three hypothetical subsets made a significant unique contribution.

Summary.--In this study, an attempt was made to demonstrate the feasibility of utilizing a modified form of JAN as a vehicle for identifying a policy of rated school effectiveness in the League of Cooperating Schools project. Four Intervention Staff members, serving as judges, generated profiles for each of the eighteen LCS and then ranked the schools in order of overall effectiveness.

With the use of the JAN technique, the four judges were placed into appropriate clusters, and it was found that at least two separate judgmental policies were present. A regression analysis of the two policies was undertaken. Policy I could be explained basically as a concern for the competence of the professional team in the schools. On the other hand, Policy II was more comprehensive in that it not only reflected a concern for a competent professional staff, but it included a concern for children as well as a concern for community support.

## 6. CANONIAL JUDGMENT ANALYSIS

What is now proposed is a strategy in which the JAN technique can be extended to include the ratings of judges on two or more criterion variables or dimensions. The technique is identified as Canonical Judgment Analysis or C-JAN. The C-JAN technique was successfully used by Johnson and King (1973) in a team doctoral dissertation at the University of Northern Colorado.

### Definition of Terms

The following terms are defined in the development of C-JAN:

Double-Barreled Principal Components Solution.--A factor solution for a canonical correlational analysis. In this type of factor solution a principal components solution for the predictor (profile) variables is given in conjunction with a principal components solution for the criterion (judgment) variables. Not only are the factors in each of the above principal component solutions orthogonal to each other, but the cross-set factors are orthogonal to each other.

Factorial Judge.--A judge generated from the predictor and criterion variable scores and the weights of a double-barreled principal components solution of a particular judge.

Type A JAN.--A JAN in which all the judges give ratings on the same subjects with respect to the same criterion variable and predictor variables.

Type B JAN.--A JAN in which the judges do not rate the same subjects with respect to the same criterion and predictor variables.

teps in C-JAN Process

tep 1

For each judge run a canonical correlation analysis using Veldman's 1967) CANONA program. Let the judges be $J_k$ for k = 1,...,m

tep 2

For each judge, $J_k$, determine the number of factorial judges, $J_{k,F1}, J_{K,F2}, \ldots J_{k,Fn_F}$.

iis is where $J_k$, would be the ith factorial judge generated from the ith ictor for the kth judge. Also, $n_F$ = the number of significant factors.

1. Let $\underline{Z}_{Fi}$ be the canonical predictor factor score vector for the ith factor for the kth judge..

2. Let $\underline{U}_{Fi}$ be the canonical criterion factor score vector associated with $\underline{Z}_{Fi}$ for the kth judge.

3. Let $(a_{1,Fi})_{i=1}$ be the weight vector for the jth predictor factor for the kth judge.

4. Let $(b_{i,Fj})_{i=1}^{t}$ be the weight vector for the jth criterion factor for the kth judge.

5. Let the following model be used in the JAN process for the factoral judge $J_{k,Fi}$ for i=1, ... , $n_F$.

   The criterion vector: $(\underline{Z}_{Fi}', \underline{U}_{Fi}')'$

   The profile matrix:

| $a_{1,Fi}{}^{*}X_1$ | $a_{2,Fi}{}^{*}X_2$ | ... | ... | $a_{s,Fi}{}^{*}X_s$ | $b_{1,Fi}{}^{*}Y_1$ | ... | ... | $b_{t,Fi}{}^{*}Y_T$ |
|---|---|---|---|---|---|---|---|---|
| xx | xx | ... | ... | xx | | | | |
| xx | xx | ... | ... | xx | | $O_{Nxt}$ | | |
| xx | xx | ... | ... | xx | | | | |
| | | | | | xx | ... | ... | xx |
| | | $O_{Nxs}$ | | | xx | ... | ... | xx |
| | | | | | xx | ... | ... | xx |

N = number of subjects for $J_k$.

Step 3

Determine the judges who should be retained. Judges who identify at least one significant canonical factor should be retained in the analysis. Any judge who is unable to identify at least one significant factor should be eliminated as he is failing to relate any predictor variable set to any criterion variable set. After eliminating inconsistent judges, a Type A or Type B (JAN) should be completed on all of the factorial judges identified in the study.

Step 4

For every policy captured in Step 3 form a matrix in which each column represents the respective factorial judge's original factor loadings. These loadings will be obtained from the CANONA printout for the judge from which the factorial judge was generated. Include along with this matrix the corresponding vector of canonical correlations for the original CANONA printout.

Step 5

At this point aided with the data presented in Step 4, the researcher should make an intuitive analysis of each of the captured factorial policies in order to determine relationships between predictor variable sets and criterion variable sets.

A limitation in this approach to C-JAN is that a single judge may be allowed to express more than one policy as more than one canonical correlation associated with his judgments may be significant. Unfortunately this full C-JAN technique is so complex that it has rarely been used.

Instead we propose a simplified C-JAN methodology which may be suitable for use in many practical situations and avoids much of the complexity of the full C-JAN methodology. Essentially, the canonical analysis will only be used as a data reduction technique to reduce the multiple criterion variables to a single criterion variable. This then allows use of the standard JAN analysis. This approach would be suitable for the case in which judge's rankings on the multiple criterion variables display a degree of redundancy. The basic steps are as follows:

1. Give a set of N profiles to the K judges and have them rank the profiles on the specified criterion variables.

2. Use canonical correlation analysis to produce a set of canonical functions for each judge using the judge's rankings as one canonical set and the profile variables as the second canonical set.

3. Check the canonical correlation between the first and second two canonical functions for each judge. To continue with the simplified C-JAN procedure, it would be necessary for the first canonical functions to be of practical significance and the second and further

possible canonical functions to be of little or no practical significance. If even the first canonical R is of no significance for a particular judge, the judge should not be used in further analysis. If more than the first canonical functions are highly important, the more complex C-JAN procedure must be used.

4. Use the first criterion canonical function to produce a new canonical variate for each judge. Substitute the new canonial variate for the original set of criterion ranking variables for each judge. Substitute the new canonical variate for the original set of criterion ranking variables for each judge.

5. Proceed with the standard JAN analysis as described in the previous section.

6. If multicollinearity of the profile variable set is not a problem, then regression analysis can be used to capture the judgment policies as usual. If multicollinearity is a problem, then canonical correlation analysis may be used to help determine the judgmental policies.

The logic behind this procedure is quite straightforward. The first ınonical criterion function is the linear combination of the criterion ıriables which extracts the maximum possible variance of the criterion ıriables and has the maximum covariance with the first canonical function of ıe profile variables. We are attempting to maximize the simplicity of ıbsequent data analysis while minimizing the loss of information.

## ıplication Example

Many institutions of higher education have internal funds which are used ɔ support the beginning stages of research which may lead to outside funding nd publishable journal articles. It is typical for such funds to be llocated by committee decision. Several interesting questions might be aised about such decisions:

1. Given a set of profile descriptors of a research proposal, how many different judgmental policies exist among the committee members in determining the quality of the research proposals?

2. Which descriptors do the differing judgmental policy groups emphasize in determining proposal quality?

The following example illustrated the C-JAN approach in answering the stated questions. We first constructed a set of 32 hypothetical descriptions of proposals by use of simulation techniques. A sample profile appears in Table 19.

TABLE 19

Sample Research Proposal Profile

| Profile Variable ID Numbers and Descriptors | Weak 1 | 2 | 3 | Average 4 | 5 | 6 | 7 | Strong 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. Need | | | 3 | | | | | | | |
| 2. Feasibility | | | | | | | | 8 | | |
| 3. Cost benefit | | | | 4 | | | | | | |
| 4. Quality of writing | | 2 | | | | | | | | |
| 5. Originality | | | | | | 6 | | | | |
| Judges' Overall Rating (repeated rankings not allowed) | Rank Profile from 1st (strongest) to 32nd (weakest) | | | | | | | | | |
| Possibility of generating outside funding | | | | | | | | | | |
| Possibility of leading to publishable journal research | | | | | | | | | | |

The set of 32 profiles was then submitted to each of four members of a hypothetical proposal funding committee. The judges were required to independently rank their set of profiled from strongest (1st) to weakest (32nd) based on the profile descriptor values. This ranking had to be accomplished first for the possibility that the proposed research would lead to outside funding, and secondly, for the possibility the proposed research would generate journal publication. The rankings for each of the criterion variables should be carried out at separate times in order to minimize halo effect. Tied rankings were not allowed for any particular criterion variable.

Tables 20 and 21 show means, standard deviations and intercorrelations of the five simulated profile variables. The simulated profiles appear to be quite good with consistent means, standard deviations, and low intercorrelations between the profile variables.

TABLE 20
Means and Standard Deviations (N = 32)

| Variable | Mean | Standard Deviation |
|---|---|---|
| 1 | 6.25 | 2.54 |
| 2 | 5.69 | 2.76 |
| 3 | 5.34 | 2.73 |
| 4 | 5.72 | 3.15 |
| 5 | 5.25 | 2.80 |

TABLE 21
Intercorrelations of the Profile Variables

| | Research Proposal Profile Variables | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 1.00 | -.28 | -.23 | -.24 | .23 |
| 2 | -.28 | 1.00 | -.03 | -.19 | -.13 |
| 3 | -.23 | -.03 | 1.00 | .09 | -.06 |
| 4 | -.24 | -.19 | .09 | 1.00 | .01 |
| 5 | .23 | -.13 | -.06 | .01 | 1.00 |

The set of two criterion variable rankings and the five profile variables were then subjected to canonical correlation analysis for each judge. The canonical correlations for this analysis are displayed in Table 22.

TABLE 22
Canonical Correlations Between the Ranking and Profile Variable Sets by Judge

| Judge Number | Canonical R | |
|---|---|---|
| | 1st | 2nd |
| 1 | .959 | .272 |
| 2 | .899 | .541 |
| 3 | .916 | .367 |
| 4 | .915 | .329 |

In each case the first canonical correlation is very strong while the second is comparatively weak. We therefore proceeded with the simplified C-JAN procedure. The first canonical function for the criterion variable set was used to produce a single canonical variable for each judge. The original set of two criterion variable rankings was replaced by the single canonical variable.

The modified data were then analyzed by means of the JAN procedure which computes a regression equation for each judge and then hierarchically clusters the judges based on the homogeneity of their prediction equations. A general idea of which judges will cluster together can be determined by looking at Table 23 which shows the intercorrelations of the judges.

TABLE 23
Intercorrelations of Judge's Ratings

| Judge | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 1.00 | .46 | .39 | .49 |
| 2 | .46 | 1.00 | .95 | .94 |
| 3 | .39 | .95 | 1.00 | .95 |
| 4 | .49 | .94 | .95 | 1.00 |

stages of the JAN process are displayed in Table 24.

TABLE 24
Stages of the JAN Procedure for the Four Judges

| ıge | Judges | System $R^2$ | Total System $R^2$ Drop |
|---|---|---|---|
| 1 | 1, 2, 3, 4 | .8507 | |
| 2 | (2, 4), 1, 3 | .8497 | .0011 |
| 3 | (2, 3, 4), 1 | .8472 | .0035 |
| 4 | (1, 2, 3, 4) | .6864 | .1643 |

ing an *a priori* criterion of an $R^2$ drop of .05 or more as indicating a parture from linearity, the clustering of judges is easily determined. The op in overall system $R^2$ for stages one through three are of little nsequence. Judges which cluster together are indicated by parentheses. The drop from stage 3 to 4 is considerably larger than the .05 criterion and dicates a substantial loss of predictive efficiency. We therefore conclude at two policies were present in the committee. Judge 1 has Policy I while ıdges 2, 3 and 4 have Policy II.

To explain the two policies, all possible subsets regression was used. A ıugh idea of the profile variables the judges were attending to while making ıeir ranking can be gained from Table 25.

TABLE 25
Correlations Between Judges and
Research Proposal Profile Variables

| Judge | Research Proposal Variables | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | -.48 | .27 | -.11 | -.60 | -.46 |
| 2 | .08 | -.13 | -.75 | -.31 | -.26 |
| 3 | -.13 | -.24 | -.75 | -.26 | -.26 |
| 4 | .04 | -.17 | -.72 | -.33 | -.29 |

To explain Policy I, the use of Table 26 is required. Table 26 indicates all possible combinations of profile variables ordered by their $R^2$ values for predicting the canonical variables of Judge 1.

TABLE 26

Results from All Possible Subsets
Regression for the Single Judge Cluster (Judge 1)

| Profile Variables in Equation | RSQ |
|---|---|
| 1, 2, 3, 4, 5 | .919 |
| 1, 3, 4, 5 | .909 |
| 1, 2, 4, 5 | .874 |
| 1, 4, 5 | .868 |
| 1, 2, 3, 4 | .817 |
| 1, 3, 4 | .810 |
| 1, 2, 4 | .775 |
| 1, 4 | .771 |
| 2, 3, 4, 5 | .584 |
| 2, 4, 5 | .577 |
| 3, 4, 5 | .574 |
| 4, 5 | .567 |
| 1, 2, 3, 5 | .420 |
| 1, 3, 5 | .411 |
| 2, 3, 4 | .390 |
| 2, 4 | .387 |
| 1, 2, 5 | .372 |
| 3, 4 | .366 |
| 4 | .362 |
| 1, 5 | .358 |
| 1, 2, 3 | .293 |
| 1, 3 | .278 |
| 2, 3, 5 | .272 |
| 2, 5 | .255 |
| 1, 2 | .248 |
| 3, 5 | .229 |
| 1 | .228 |
| 5 | .211 |
| 2, 3 | .082 |
| 2 | .072 |
| 3 | .012 |

We again look for a jump in $R^2$ using the a priori .05 criterion. This jump occurs when going from the equation with variables 1, 4 and 5 to the equation with variables 1, 2, 3, and 4. Judge 1 was attending to variables 1, 4 and 5. We can also see that major emphasis was placed on variable 4. In other words, the Policy I judge was primarily considering need, quality of writing, and originality while ranking the proposals and essentially ignoring feasibility and cost benefit.

Policy II can be explained in a similar manner using Table Table 27 shows the all possible subsets regression for Judges 2, 3 and 4 bined as a single data set.

TABLE 27

Results from All Possible Subsets Regression for the Three Judge Cluster (Judges 2, 3, 4)

| Profile Variables in Equation | RSQ |
|---|---|
| 1, 2, 3, 4, 5 | .824 |
| 2, 3, 4, 5 | .790 |
| 1, 2, 3, 4 | .729 |
| 1, 2, 3, 5 | .718 |
| 2, 3, 5 | .709 |
| 1, 3, 4, 5 | .708 |
| 3, 4, 5 | .702 |
| 2, 3, 4 | .667 |
| 1, 3, 5 | .649 |
| 3, 5 | .648 |
| 1, 3, 4 | .624 |
| 1, 2, 3 | .612 |
| 3, 4 | .603 |
| 2, 3 | .588 |
| 1, 3 | .554 |
| 3 | .547 |
| 1, 2, 4, 5 | .240 |
| 2, 4, 5 | .239 |
| 1, 4, 5 | .167 |
| 4, 5 | .162 |
| 1, 2, 4 | .155 |
| 2, 4 | .149 |
| 1, 2, 5 | .129 |
| 2, 5 | .120 |
| 1, 5 | .095 |
| 1, 4 | .090 |
| 4 | .090 |
| 5 | .073 |
| 1, 2 | .034 |
| 2 | .033 |
| 1 | .007 |

In this case we see that a major jump in $R^2$ occurs when going from variables 2, 3, 4, and 5 to 1, 2, 3 and 4. It is obvious that variable 3 was of major importance. That is, the Policy II judges were attending to feasibility, cost benefit, quality of writing and originality with a primary emphasis on cost benefit while ranking the proposal profiles. Need was not viewed as important. It is interesting to note that neither of the policy groups attended to all the profile variables.

Although JAN and C-JAN are useful and innovative procedures, they do have some general problems. As with any statistical procedure, it would oftentimes be advisable to validate the results by use of split sample techniques or replication. Since the JAN procedure is based on regression, it suffers from the same problems encountered with regression. For example, JAN must have a sufficient ratio of profiles to profile variables to avoid overfit which results in inflated and unstable $R^2$s. Since JAN clusters on the basis of homogeneity of prediction equations, multicollinearity of the profile variables is also a serious problem. High multicollinearity will lead to questionable clustering results and make the interpretation of the captured policies quite difficult. However, if utilized properly, JAN and C-Jan are promising tools for evaluation methodologists to be used as additional techniques in decision-making and policy-capturing situations.

# BIBLIOGRAPHY


ck, D. E.
1973 Development of the E-2 weighted airman promotion system. AFHRL-TR-73-3, AD-767 195. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

tenberg, Robert A. and Raymond Christal
1968 Grouping criteria - a method which retains maximum predictive efficiency. The Journal of Experimental Education 36, 4: 28-34.

ing, N-K.
1970 Hierarchical groupings of judges according to selected criteria for financial aid awards. Unpublished doctoral dissertation, University of Northern Colorado.

ristal, Raymond E.
1968a JAN: a technique for analyzing group judgment. The Journal of Experimental Education 36, 4: 24-27.

1968b Selecting a harem - and other applications of the policy-capturing model. The Journal of Experimental Education 36, 4: 35-41.

dycha, A. L.
1970 A monte carlo evaluation of JAN: a technique for capturing and clustering raters' policies. Organizational Behavior and Human Performance, 5: 501-506.

ff, W. L.
1969 A quantitative study of teacher selection and evaluation of policies at a suburban elementary school district. Unpublished doctoral dissertation, University of California at Los Angeles.

och, L. L.
1972 Policy capturing with local models: the application of the AID technique in modeling judgment. Unpublished Ph.D. Dissertation, The University of Texas, Auston.

tt, C. D.
1974 Development of the weighted airman screening system for the air reserve forces. AFHRL-TR-74-18, AD-781 747. Lackland AFB, TX: Computational Sciences Division, Air Force Human Resources Laboratory.

uston, Judith A., Houston, Samuel R. and R. LaMonte Chlson
1974 On determining pornographic material. The Journal of Psychology, 86: 277-287.

Houston, Samuel R.

1967 The Judgment Analysis regression technique applied to the admission variables for doctoral students at Colorado State College, 1963-1966. Unpublished Ph.D. Dissertation, University of Northern Colorado, Greeley.

1968 Generating a projected criterion of graduate school success via normative Judgment Analysis. The Journal of Experimental Education 37, 2: 53-58.

1974a Classification of Judgment Analysis. In: Judgment Analysis: tool for decision makers, edited by Samuel R. Houston, pp. 52-53. New York: MSS Information Corp.

1974b Issues associated with the use of Judgment Analysis. In: Judgment Analysis: tool for decision makers, edited by Samuel R. Houston, pp. 69-73. New York, MSS Information Corp.

1974c Faculty policies of teaching effectiveness. In Judgment Analysis: tool for decision makers, edited by Samuel R. Houston, pp. 140-147. New York, MSS Information Corp.

-----, and James T. Bolding, Jr.

1974 The general linear model and Judgment Analysis. In: Judgment Analysis: tool for decision makers, edited by Samuel R. Houston, pp. 54-60. New York, MSS Information Corp.

-----, and Joseph W. Gilpin

1971 Hierarchical groupings of students according to their policy of rated teacher effectiveness. SPATE 10, 2: 28-53.

-----, and Gary C. Stock

1973 Judgment Analysis (JAN): tool for education decision-makers. SRIS Quarterly 6, 2: 22-24.

Holmes, George P. and Sheldon Zedick

1973 Judgment analysis for assessing paintings. The Journal of Experimental Education 41, 4: 26-30.

Johnson, J. W., and King, F. S.

1973 Multiple criterion judgment analysis for the educational researcher. Unpublished team doctoral dissertation, University of Northern Colorado.

Jones, K. M., Mennis, L. S., Martin L. R., Summers, J. L., and G. R. Wagner

1976 Judgment modeling for effective policy and decision making. Research Report for Air Force Office of Scientific Research Grant No. AFOSR-74-2656, AD-A033 186.

Keelan, J. A., Houston, T. R., and Houston, S. R.

1973 Leadership policies as perceived by firemen. Colorado Journal of Education Research, 12: 20-23.

plyay, J. B.
1970 Extension of the weighted airman promotion system to grades E-8 and E-9. AFHRL-TR-70-2, AD-703 687. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

plyay, J. B., Albert, W. G., and D. E. Black
1976 Development of a senior NCO promotion system. AFHRL-TR-76-46, AD-A030 607. Lackland AFB, TX: Computational Sciences Division, Air Force Human Resources Laboratory.

llins, C. J. and E. Usdin
1970 Estimation of validity in the absence of a criterion. AFHRL-TR-70-36, AD-716 809. Lackland AFB, TX: Personnel Division, Air Force Human Resources Laboratory.

aylor, J. C., and Wherry, R. L.
1965 The use of simulated stimuli and the JAN technique to capture and cluster the policies of raters. Educational and Psychological Measurement, 25: 969-986.

tock, C. C.
1969 Judgment analysis for the educational researcher. Unpublished doctoral dissertation, Colorado State College.

uits, D. B.
1957 Use of dummy variables in regression equations. Journal of the American Statistical Association 52: 548-551.

orgunrud, E. A.
1971 Criteria guiding curriculum decisions of selected school superintendents. Unpublished doctoral dissertation, University of California at Los Angeles.

ard, Joe H., Jr.
1961 Hierarchical grouping to maximize payoff. Lackland Air Force Base, Texas: Personnel Laboratory, Wright Air Development Division.

1963 Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58: 236-244.

-----, and K. Davis
1963 Teaching a digital computer to assist in making decisions. PRL-TDR-63-16, AD-407 322. Lackland AFB, TX: 6570th Personnel Research Laboratory, Aerospace Medical Division.

-----, and Marion E. Hook
1963 Applications of an hierarchical grouping procedure to a problem of grouping profiles. Education and Psychological Measurement 23: 69-81.

Williams, J. D., Gab, D., Linden, A.
1969 Judgment analysis for assessing doctoral admission policies. Journal of Experimental Education, 38: 92-96.

# The Use of MLR Models to Analyze Partial Interaction: An Educational Application

John W. Fraas
Ashland College
Mary Ellen Drushal
Ashland Theological Seminary
Ashland College

Abstract

Certain research questions found in educational studies require partial interaction effects to be tested. This paper presents an application of the method of using MLR models to test a partial interaction hypothesis.

## Introduction

Newman, Deitchman, Burkholder, Sanders, and Ervin (1976) addressed the issue of the importance of matching the statistical analysis with the question posed by the researcher. The use of multiple linear regression (MLR) models allows the researcher the flexibility of analysis needed to address research questions that require the testing of partial interaction (see McNeil, Kelly and McNeil; 1975). This paper presents the MLR models and the technique used to test a partial interaction research hypothesis posed in an educational study.

## Research Design

A study by Drushal (1986) examined the impact of various participative decision making (PDM) techniques. The techniques examined in the study were Delphi Survey Technique (DST), Social Judgment Analysis (SJA), Nominal Group Technique (NGT), and a control group. The students in the control group were not exposed to any of the PDM techniques.

Seminary students were randomly assigned to one of the four groups. Through participation in a decision making technique, students selected the criteria to be considered in making a curriculum choice for a Sunday school. After experiencing the assigned decision making technique, participants responded to the Participative Management Survey (PMS). The PMS is a survey composed of research-based statements on leadership, trust, communication and participative decision making (see Drushal, 1986). Each student in the study received a total score on the PMS instrument. These total scores served as the values of the dependent variable for the MLR models used to test the partial interaction research question presented in the next section of this paper.

## Research Hypothesis

One of the research hypotheses of interest to the researchers was:

> $H_1$: The difference between the average of the mean PMS scores for females in the PDM groups and the mean PMS score for females in the control group will exceed the difference between the average of the mean PMS scores for males in the PDM groups and the mean PMS score for males in the control group.

To test this research hypothesis, a test of partial interaction was required. The construction and analysis of MLR models readily allowed the researchers to test this partial interaction hypothesis.

## Full MLR Model

The full MLR model used to test the partial interaction hypothesis contains the interaction effect between the two independent variables--instructional techniques and gender. There were four instructional techniques and the two levels of gender. The full MLR model, which is a full interaction model, was:

$$y = au + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 + b_5 x_5 + b_6 x_6 + b_7 x_7 + e$$

where:

$y$ = PMS score for each student
$x_1$ = 1 if student in DST group and female; 0 otherwise
$x_2$ = 1 if student in SJA group and female; 0 otherwise
$x_3$ = 1 if student in NGT group and female; 0 otherwise
$x_4$ = 1 if student in Control group and female; 0 otherwise
$x_5$ = 1 if student in DST group and male; 0 otherwise
$x_6$ = 1 if student in SJA group and male; 0 otherwise
$x_7$ = 1 if student in NGT group and male; 0 otherwise
$x_8$ = 1 if student in Control group and male; 0 otherwise
$a$ = constant term
$e$ = error term
$u$ = unit vector

It is interesting to note that the $R_2$ value of this full model will equal the $R_2$ value generated by a oneway ANOVA of the scores of the eight groups.

Since the computer program used to compute the parameters for the full MLR model includes a unit vector, the variable $x_8$ was not included in the model. Thus, the value for a--the constant term--represents the mean PMS score for the males in the control group. The $b_1$ value represents the difference between the mean PMS score for females in the DST group and the value for the constant term a, which is the mean PMS score for males in the control group. The other b values contained in the full MLR model would be interpreted in a similar fashion.

## Restriction

The restriction made on the full model to obtain the restricted MLR model required that the difference between the average of the mean PMS scores of the females in the PDM groups and the mean PMS score for females in the control group be equal to the difference between the average of the mean PMS scores of the males in the PDM groups and the mean PDM score for males in the control group. Thus, the restriction was:

$$(b_1 + b_2 + b_3)/3 - b_4 = (b_5 + b_6 + b_7)/3$$

The left-hand side of the restriction represents the difference between the PMS mean scores of the females assigned to the PDM groups and the mean score of the females in the control group. The right-hand side of the restriction represents the difference between the average of the mean PMS scores for males in the PDM groups and the mean PDM score for the males in the control group.

Again, it is interesting to note that in view of the fact that the $R_2$ value of the full model corresponds to the $R_2$ value that would be generated by an ANOVA of the scores, this restriction can be thought of as a contrast of the eight group means. The restriction specifies the contrast. Williams (1976 and 1979) discussed the use of MLR models to conduct contrasts of group means.

The restriction can be more clearly explained by referring to a graph of the interaction effect between the instructional methods and gender, which was estimated by the regression coefficients of the full MLR model. Gender was placed along the X axis of Figure 1. Recall that each of the regression coefficients of the full MLR model represents the differences between the mean

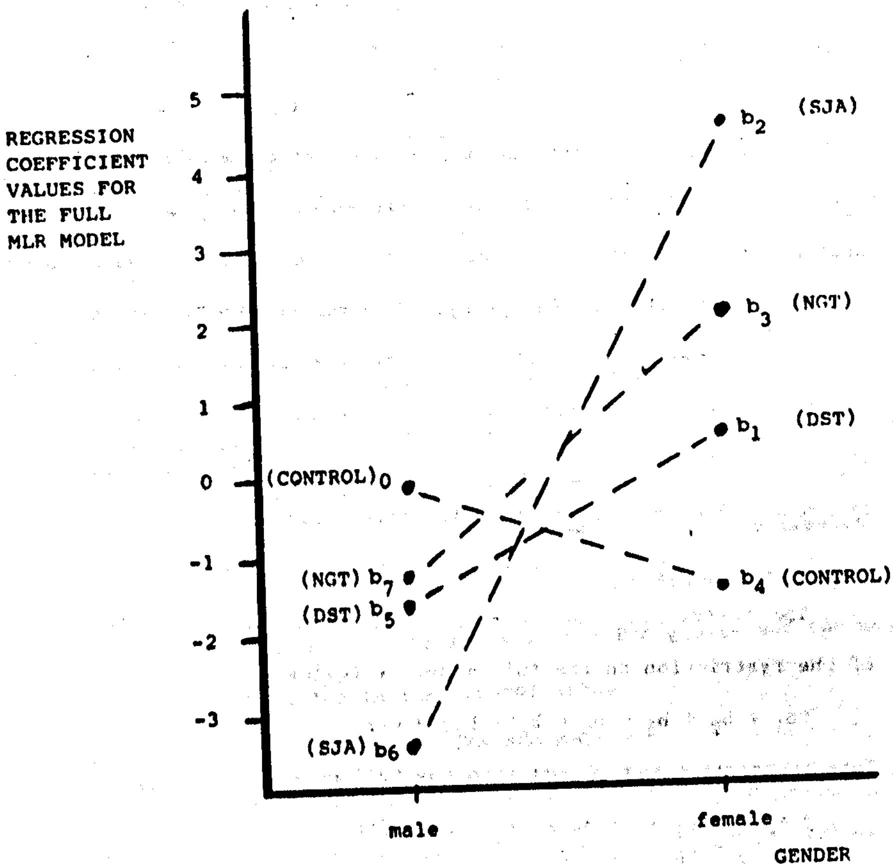



Figure 1

Interaction Effect Estimated by the Full MLR Model

PMS score for a given instructional group and gender, and the mean PMS score for males in the control group. Thus, the Y axis of Figure 1 represents the differences in the mean PMS scores of the various combinations of groups and gender, and the value for the constant term a, which is the mean PMS score of the males in the control group.

In Figure 1 the distance between average of points $b_1$, $b_2$ and $b_3$, and point $b_4$ represents the difference between the mean PMS scores for females in the three PDM groups and the mean PDM score for females in the control group. The restriction requires that this distance equal the distance between the average of points $b_5$, $b_6$ and $b_7$, and the 0 point, which is the difference between the average of the mean PMS scores of the males in the PDM groups and the mean PMS score for males in the control group.

### Restricted MLR Model

The restriction was manipulated to facilitate the placement of the restriction on the full model as follows:

$$(b_1 + b_2 + b_3 - b_5 - b_6 - b_7)/3 = b_4$$

This restriction was placed into the full model as follows:

$$y = au + b_1 x_1 + b_2 x_2 + b_3 x_3 + ((b_1 + b_2 + b_3 - b_5 - b_6 - b_7)/3)\, x_4 + b_5 x_5 + b_6 x_6 + b_7 x_7 + e$$

Multiplying the restriction by $x_4$ and collecting like regression coefficients produced the following restricted model:

$$y = au + b_1 \left(x_1 + \frac{x_4}{3}\right) + b_3 \left(x_2 + \frac{x_4}{3}\right) + b_3 \left(x_3 + \frac{x_4}{3}\right) +$$

$$b_5 \left(x_5 - \frac{x_4}{3}\right) + b_6 \left(x_6 - \frac{x_4}{3}\right) + b_7 \left(x_7 - \frac{x_4}{3}\right) + e$$

To facilitate the analysis of the restricted MLR model by the computer, the following variables were calculated:

$$x_9 = x_1 + x_4/3$$

$$x_{10} = x_2 + x_4/3$$

$$x_{11} = x_3 + x_4/3$$

$$x_{12} = x_5 - x_4/3$$

$$x_{13} = x_6 - x_4/3$$

$$x_{14} = x_7 - x_4/3$$

Thus, the restricted model took the form:

$$y = au + b_9 x_9 + b_{10} x_{10} + b_{11} x_{11} + b_{12} x_{12} + b_{13} x_{13} + b_{14} x_{14} + e$$

Due to the nature of the restriction, this restricted model requires that the difference between the average PMS scores for females in the PDM groups and the mean PMS score of the females in the control group be equal to the difference between the average of the mean PMS scores of the males in the PDM groups and the mean PMS score of the males in the control group.

### Test of the MLR Models

To determine whether the data supported the researcher hypothesis , an F test of the difference between the $R^2$ values of the full and restricted models was required. The results of the analysis are presented in Table 1. Since the research hypothesis was directional, the critical F value of 2.75 for the alpha level of .05 corresponded to the critical value of a directional or

## Table 1

### F Test of the Partial Interaction Research Hypothesis

| Hypothesis and Models | $R^2$ | df | F | Critical F | Sig |
|---|---|---|---|---|---|
| $H_1$: The difference between the average of the mean PMS scores for females in the PDM groups and the mean PMS score for females in the control group will exceed the difference between the average of the mean PMS scores for males in the PDM groups and the mean PMS score for males in the control group. | | | | | |
| Full Model: $y = au + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 + b_5 x_5 + b_6 x_6 + b_7 x_7 + e$ | .1263 | | | | |
| Restriction: $(b_1 + b_2 + b_3)/3 - b_4 = (b_5 + b_6 + b_7)/3$ | | 1/129 | 6.02 | 2.75 | Sig |
| Restricted Model: $y = au + b_9 x_9 + b_{10} x_{10} + b_{11} x_{11} + b_{12} x_{12} + b_{13} x_{13} + b_{14} x_{14} + e$ | .0855 | | | | |

one-tailed test. The F test revealed that the calculated F value of 6.02 did exceed the critical F value of 2.75.

Even though the calculated F value exceeded the critical value, the researchers had to check the signs of the regression coefficients contained in the restriction before it could be determined whether the directional research hypothesis was supported by the data. That is, the difference between the average of the mean PMS scores for females in the PDM groups and the mean PMS score for the females in the control group had to exceed the difference between the average of the mean PMS scores for males in the PDM groups and the mean PDM score for the males in the control group.

The regression coefficient values for the full MLR model were as follows:

$b_1 = .78$
$b_2 = 4.92$
$b_3 = 2.07$
$b_4 = -1.47$
$b_5 = -1.59$
$b_6 = -3.22$
$b_7 = -1.07$

To support the directional statement contained in the research hypothesis, the left-hand side of the restriction had to be greater than the right-hand side of the restriction. That is:

$$(b_1 + b_2 + b_3)/3 - b_4 > (b_5 + b_6 + b_7)/3$$

The regression coefficients indicated that the value of 4.06 for the left-hand side of the restriction was indeed greater than the value of -1.96 for the right-hand side of the restriction. Therefore, the signs of the regression coefficients as well as the

F test of the difference between the $R^2$ values of the full and restricted MLR models supported the research hypothesis.

## Summary

Researchers should not be hesitant to include partial interaction questions in research projects because of the perceived difficulty of testing such hypotheses. As indicated by the procedures presented in this paper, the use of MLR models allows researchers to analyze partial interaction questions in a versatile and straightforward manner.

## References

Drushal, M.E. Attitudes toward participative decision making among church leaders: A comparison of the influences of nominal group technique, delphi survey techniques, and social judgment analysis. Unpublished doctoral dissertation, Vanderbilt University, 1986.

McNeil, K., Kelly, F. and McNeil, J. Testing research hypotheses using multiple linear regression. Carbondale, Illinois: Southern Illinois University Press, 1975.

Newman, I., Deitchman, R., Burkholder, J., Sanders, R. and Ervin, L. Type VI error: inconsistency between the statistical procedure and the research question. Multiple Linear Regression Viewpoints, 1976, 6(4), 1-19.

Williams, J.D. Multiple comparisons by multiple linear regression. Multiple Linear Regression Viewpoints, Monograph Series #2, 1976, 7.

Williams, J.D. Contrasts with unequal N by multiple linear regression. Multiple Linear Regression Viewpoints, 1979, 9(3) 1-7.

# Conducting an 86-variable Factor Analysis on a Small Computer and Preserving the Mean Substitution Option

Irvin Sam Schonfeld
The City College of New York and
New York State Psychiatric Institute
Candace Erickson
Columbia University
College of Physicians and Surgeons

Abstract

This paper shows how we overcame limitations imposed on us by the memory capacity of the relatively small mainframe we used in conducting a factor analysis in which means are substituted for missing values. Insufficient memory did not permit us to employ SPSSX, with its mean substitution feature, in conducting a factor analysis of 86 variables reflecting ways in which parents cope with the hospitalization of their children. Instead, we employed a two-step solution: (1) we ran SPSSX Condescriptive to create *z*-score equivalents of the 86 variables and recoded the *z* variables' system missing values to zeros; (2) the output of the Condescriptive run constituted the input of a BMDP P4M factor analysis run.

Frequently researchers who choose to conduct factor analyses will take advantage of software available in the SPSSX (SPSSX) package. There are several advantages that the SPSSX package offers over previous releases. SPSSX can handle more variables and it can substitute means for missing values. The latter feature is helpful because with it a case is not deleted when a missing variable is encountered.

A disadvantage of SPSSX is that it uses a great of deal of memory. This disadvantage came home to us when we attempted to factor analyze a data set consisting of 86 variables and 271 cases[1]. The variables consisted of parents' responses to 86 of 173 questionnaire items describing behaviors adults use to cope with the problem of having a child in the hospital. Subjects' response choices ranged from "not at all" (0) to "very much" (3). Examples of coping questionnaire items are presented in Figure 1.

If we were to permit the program to delete cases with any missing values, our data set would have been reduced substantially. Of the 271 cases 137 subjects, or 51%, had no missing values; therefore, we would have lost 49% of our subjects. The loss of subjects would have been extremely wasteful since about 27% of the parents failed to complete only 1% of the questionnaire items; 4%, 2% of the items; and another 4%, 3% of the items. About 11% of the parents failed to complete between 4 and 14% of the items. We therefore elected to use the mean substitution option in the SPSSX Factor procedure in order to avoid subject loss.

Unfortunately the four megabyte IBM 4331 computer we used at New York State Psychiatric Institute did not provide sufficient memory to execute the job. The program listing returned the "insufficient storage" error message. We think our solution to the problem might be of interest to readers who face similar storage obstacles to running large factor analyses and other statistical procedures on small systems. In order to deal effectively with this problem we linearly transformed our original values, and then submitted the new transformed values to a factor analysis program supplied by a software package that uses computer memory more economically than SPSSX.

The data originally resided in an SPSS system file (Nie et al., 1975). Since SPSSX reads SPSS system files, we wrote an SPSSX program to read the system file. The program invoked a series of procedures the first of which, the Condescriptive procedure, created a new set of 86 variables (ZV1 to ZV86). The 86 new (ZV) variables corresponded one-to-one to variables (V1 to V86) in the original data set. Each new variable was the equivalent to the z-score transformation of the corresponding variable in the original data set. The Condescriptive procedure assigns a system missing value to any new (ZV) variable when the corresponding old variable is missing. Thus a parent who did not respond to questionnaire item V30 would receive a system missing value for new variable ZV30. Immediately after the Condescriptive routine was invoked the Recode command was employed to convert all system missing values in the new (ZV) variables to zero. The Recode command in effect substituted means for missing values since zero is, necessarily, the mean of a set of z-scores. Next the Write Outfile procedure was called upon to write out all the new (ZV) variables into a raw data file. Figure 2 depicts the SPSSX program that operated upon the original 86 variables.

BMDP (Brown et al., 1983) provides the user an economical alternative to SPSSX. When the user runs a BMDP job, one program out of the BMDP library of programs is called up. By contrast, when SPSSX is run, the entire SPSSX library of programs is called up. The advantage inherent in the SPSSX approach is that multiple procedures can be invoked in a single run. The disadvantage is that a great deal of memory is required to store the program

Figure 1

Circle the number that corresponds to the response that best describes your experience in the last week. If your child has been in the hospital for less than a week, circle the number that corresponds to the response that best describes your experience since your child entered the hospital.

| | very much | pretty much | just a little | not at all |
|---|---|---|---|---|
| 1. I think the doctors have made a mistake and that my child doesn't really need to be in the hospital...................... | 3 | 2 | 1 | 0 |
| 2. I watch myself doing things, and it feels like I'm watching someone else........ | 3 | 2 | 1 | 0 |
| 3. I want someone around to hold or comfort me................................ | 3 | 2 | 1 | 0 |
| 4. Something ironic or humorous usually breaks the tension......................... | 3 | 2 | 1 | 0 |

Figure 2

SPSSX program to output data

```
COMMENT        SPSSX PROGRAM TO OUTPUT DATA TO BE READ BY BMDP PROGRAM.
FILE HANDLE    SYSFILE/NAME='HOSP SYSFILE A'
FILE HANDLE    ZDATA/NAME='Z DATA A'
GET FILE       SYSFILE
COMMENT ***********************************************************************
     THE PURPOSE OF THE NEXT 6 STATEMENTS IS TO INCLUDE ONLY THOSE
     SUBJECTS WHO HAVE FEWER THAN 20% MISSING VALUES ON ALL 173 VARIABLES.
*******************************************************************************
DO REPEAT      A = V1 TO V173/B=CT1 TO CT173
COUNT          B = A (9)
END REPEAT
COMPUTE        TOT9 = SUM (CT1 TO CT173)
COMPUTE        TOT9PER = TOT9/173
SELECT IF      (TOT9PER LT .20)
COMMENT ***********************************************************************
     THE PURPOSE OF OPTION 3 OF THE CONDESCRIPTIVE PROCEDURE IS TO CREATE A
     SET OF NEW VARIABLES, ZV1 TO ZV86, WHICH ARE Z-SCORE TRANSFORMATIONS OF
     OLD VARIABLES, V1 TO V86.  WHEN A SUBJECT RECEIVED A MISSING VALUE FOR
     ONE OF THE OLD VARIABLES, S/HE IS ASSIGNED A SYSTEM MISSING VALUE ON THE
     CORRESPONDING NEW VARIABLE.
*******************************************************************************
CONDESCRIPTIVE           V1 TO V86
OPTIONS                  3
```

Figure 2, continued

```
COMMENT ************************************************************************
      THE PURPOSE OF THE RECODE STATEMENT IS TO CONVERT THE SYSTEM MISSING
      VALUES FOR THE NEW 'ZV' VARIABLES TO ZEROS.
 ******************************************************************************
RECODE ZV1 TO ZV86 (MISSING = 0)
COMMENT ************************************************************************
      THE PURPOSE OF THE WRITE OUTFILE STATEMTNENT IS TO WRITE OUT A
      RECTANGULAR DATA FILE THAT CAN BE READ BY A BMDP PROGRAM.
 ******************************************************************************
WRITE OUTFILE = ZDATA TABLE
 /ZV1 TO ZV6
 /ZV7 TO ZV12
 /ZV13 TO ZV18
 /ZV19 TO ZV24
 /ZV25 TO ZV30
 /ZV31 TO ZV36
 /ZV37 TO ZV42
 /ZV43 TO ZV48
 /ZV49 TO ZV54
 /ZV55 TO ZV60
 /ZV61 TO ZV66
 /ZV67 TO ZV72
 /ZV73 TO ZV78
 /ZV79 TO ZV84
 /ZV85 TO ZV86
 EXECUTE
 FINISH
```

library, rendering insufficient memory for jobs like ours that are conducted on small systems. We could not run the SPSSX driven factor analysis even when we created a two or a three megabyte virtual machine. We, therefore, elected to use the output of the SPSSX Write Outfile procedure, that is, the coping items rescaled as *z*-scores with zeros having replaced missing values, as the input for the BMPD Factor Analysis program, P4M. We successfully ran BMDP P4M with storage defined at 1.5 megabytes. Figure 3 shows the BMDP factor analysis program.

We thus overcame a disadvantage of the BMDP Factor Analysis program, namely, that P4M does not include a mean substitution option. The listing of the BMDP program provides a check on the adequacy of the procedure just employed. The listing included the means and standard deviations of each ZV variable. The listing showed that each of the ZV means was within rounding error of zero, and that each standard deviation attained a value of one or, as would be expected from the additional zero scores, values slightly less than one.

Figure 3

BMDP program to read output from SPSSX program and perform the factor analysis

```
COMMENT      BMDP PROGRAM TO BE RUN UNDER P4M.
/PROBLEM     TITLE IS 'HOSPITALIZATION STUDY'.
/INPUT       VARIABLES ARE 86.
                 FORMAT IS FREE.
                 CASE = 271.
/VARIABLE NAMES ARE ZV1 TO ZV86.
                 USE = 1 TO 86.
/FACTOR      NUMB = 10.
/END
-----------------------------DATA IS PLACED HERE-----------------------------
```

## References

Brown, M.B., Engelman, L., Frane, J.W., Hill, M.A., Jennich, R.I., & Toporek, J.D. (1983). BMDP Statistical Software. Berkeley, CA: University of California Press.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., & Bent, D.H. (1975). Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill.

SPSSX (1983). SPSSX User's Guide. New York: McGraw-Hill.

## Footnote

[1] We recognize that it would have been desirable to have perhaps 130 additional subjects in conducting the factor analysis. Actually the factor analysis was not our primary vehicle for studying the ways parents coped with having children in the hospital. The factor analysis was conducted as an adjunct to and a check on a more important set of analyses we had performed earlier. In the earlier analyses we constructed a priori scales by combining items clinical experience suggested went together. Typically, the scales we constructed had satisfactory internal consistency reliabilities as measured by the coefficient alpha. Generally, the items factored in ways anticipated by our a priori scales.

AR REGRESSION VIEWPOINTS
NUMBER 2, WINTER 1987

# The Use of Multiple Regression in Evaluating Alternative Methods of Scoring Multiple Choice Tests

Gerald J. Blumenfeld
Isadore Newman
The University of Akron

Echternacht (1972) has reviewed a substantial body of literature in the field of confidence testing. Confidence testing refers to methods of weighing responses so as to reflect the examinee's belief in the correctness of the options selected. The intent is to maximize the amount of information gained from a given set of test items. Lord and Novick (1968) state that maximizing this information involves the manner in which the examinees respond to the items, specifying an item scoring rule, and combining items scores into a weighted total score.

Coombs, Milholland, and Womer (1956) and Ebel (1965) report higher reliabilities for the confidence testing methods they employed when compared to traditional scoring procedures. Echternacht's review (1972)suggests that while higher reliabilities have been found, some researchers have reported lower reliabilities (Hambleton, Roberts, and Traub, 1970; Jacobs, 1971; and Koehler, 1971).

In most studies only increase in reliability has been used to evaluate confidence testing. Minimal attention has been

Presented at the American Psychological Association Convention, at Montreal Canada, August, 1973

given to validity. Archer (1962) has reported lower validity while Hambleton, Roberts, and Traub (1970) have reported higher validity. The purpose of this paper is to provide specific examples of how multiple regression analysis could be used to analyze item discrimination, item validity, and test validity when confidence testing is employed. Current practices tend to utilize apriori scoring formulas rather than maximize the predictiveness possible with the obtained data. We will also suggest that the application of these methods may require the development of multivariate techniques for assessing test reliability.

## Method: Data Collection

Subjects and Measures. During the spring quarter, 1973, two sections, 40 students per section, of one of the author's undergraduate test and measurements classes were used to collect the data reported. Students were required to pass 25 M-C item exams covering objectives from each of 6 instructional modules. Each module included initial and remedial exams. A score of 80 percent correct was required. A teaching project was also required, and two of the assignments associated with that project were used as independent criteria for estimates of validity. Only the initial exam of the first three modules was used.

Modules 1, 2, and 3 involved a) types of tests and classification of educational objectives, b) objective test items, and c) anecdotal records, rating scales, and check lists (including the analytical scoring of essays), respectively. The two assignments used as criteria for assessing validity were 1) the precise statement of a "higher-than-knowledge" behavioral objective; and 2) a three-column table containing a) a higher-than-knowledge behavioral objective, b) a description of an instructional procedure appropriate for the objective, and c) a measurement device which agreed with both the objective and the specific instruction proposed.

Success in developing such a three-column table is one of the major objectives of the course. Therefore, use of these project scores as a criterion for assessing the validity of the exams is appropriate.

Scoring Procedure. Students were required to respond to each four- or five-option multiple choice item twice. They indicated the option they thought least likely to be correct. If the correct option was selected as most likely to be correct, the item was scored two points; if the correct option was selected as least likely to be correct, the item was scored zero points; if the correct option was neither selected as most likely correct nor least likely correct, the item was scored one point.

The statement of a behavioral objective was scored on a zero to five point scale. The objective had to be stated in behavioral terms to receive at least one point. Inclusion of stimulus conditions and required standard of excellence added one point each. If the objective was at the higher-than-knowledge level, this received one point and the omission of any reference to instruction received one point.

The three-column table was scored on a zero to three point scale. The objective had to describe a higher-than-knowledge level behavior or task to receive at least one point. If the proposed instruction agreed with the objective, a second point was awarded. If the measurement procedure and device agreed with both the objective and the instructional procedure, a third point was awarded.

The authors scored the objectives and the three-column tables independently. Discrepancies were discussed until a common score could be agreed upon. The independent scoring resulted in agreement on more than 80 percent of the papers. Discussion was needed on the other 20 percent.

## Results and Discussion

Validity estimates were calculated on two separate criteria. The first criterion was objectives that the students wrote which received grades ranging from 0 through 5. The second criterion for validity estimates was the students project score. This project consisted of writing a behavioral objective, describing how the objective would be taught and how it would be tested. (See method section for more details).

Validity estimates for each of the two criteria were calculated four different ways. These four methods were applied to each of the three tests. The first method (the traditional method) simply correlated (r) the subject's total score on each test separately with the score they received on criterion one (objectives). Under this condition, the test scores were arrived by traditional grading. Each item was graded either 1 if correct, 0 otherwise.

The second method was identical to the first except in this case each test item was graded in the experimental manner so that the subject could receive for any one item either 0, 1, or 2 points. (See method section for further details). Here, as in the first method, r was used to obtain an estimate of the predictive validity.

The third method used a multiple linear regression procedure to estimate the predictive validity for the experimental procedure. This method differed from the second in that in the second method, each student received only one total score for each of the tests. This score was arrived at by summing the total points earned on each test, separately. In the third method, instead of having one predictor variable, the total test score, three predictor variables were constructed by taking a frequency count of the number of questions each student received full credit (2 points) for, the number of questions on which each received partial credit (1 point), and the number of questions on which each received no credit (0 points). In this manner, information was collected on how many items on each test each student received full, partial, or no credit for. This information then was utilized in the following equation:

Model 1 $Y_1 = a_0U + a_1X_1 + a_2X_2 + a_3X_3 + E_1$

Where $Y_1$ = the score received on the objectives

$X_1$ = the number of 0's each student received

$X_2$ = the number of 1's each student received

$X_3$ = the number of 2's each student received

$U$ = 1 if the subject is in the sample, 0 if otherwise

$a_0$, $a_1$, $X_3$ = partial regression weights

$E_1$ = error vector $(Y_1-Y_1)$

Method four was exactly the same as the third method except that a correction for shrinkage was calculated for the multiple regression formula. The shrinkage formula used was:

$$R^2_s = 1 - (1-R^2)\frac{N-1}{N-K}$$

Where: $R^2_s$ = the corrected shrunken $R^2$

$R^2$ = the calculated $R^2$

$N$ = the number of independent observations

$K$ = the number of predictor variables

Methods one through four were duplicated exactly using as the criterion, scores on the project in place of scores obtained on the objectives. These results are presented in Tables 1 and 2.

Inspection of Tables 1 and 2 indicates that method two produced a higher predictive validity estimate than did method one, four out of six times. (This was found not to be significant as a Sign Test was used). Method three, the employment of the multiple regression technique, was found to produce higher predictive validity estimates than both methods one and two, six out of six times. This was considered significant since the pròbability of the Sign Test was $p = .0156$. Method four, in which the R was corrected for shrinkage, was also found to produce higher predictive validity estimates than method one, six out of six times ($p = .0156$) and higher validity estimates than method two, five out of six times ($p = .0938$). This was found to be non-significant at alpha = .05. However, one should keep in mind that the Sign Test is highly conservative.

Seventy-five additional analyses were computed in which each item (25 items per test, on three tests) was used as the predictor variable, predicting the scores on the objectives using methods one and three (traditional scoring and experimental scoring 0, 1, or 2, respectively). Another seventy-five analyses were computed exactly the same way predicting the project score. The results of these analyses can be found in Appendix A. They were not presented in the body of the paper because Tables 1 and 2 are conceptually a composite of all of the separate analyses which are of most theoretical and practical importance.

In addition to estimating the validity of the experimental grading procedures compared to the traditional procedure in predicting the two criteria (objectve and project scores), item discriminations were calculated for each of the items on each of the three tests, comparing both the traditional and experimental grading.

Item discrimination for the traditional method was calculated by correlating (r) the score on each item (graded 1 or 0) with the total score on the test graded in the traditional manner. Therefore, there were twenty-five item discrimination estimates for each of the three tests.

Item discrimination was calculated for the experimental method by using multiple regression analysis to predict the total score for each separate test. These total scores were arrived at by using the experimental grading system (0, 1, or 2 points) and summing these scores for all items to get the total for each test. The predictor variables (the experimental score, or 0, 1, or 2 for each item, was placed into one of three vectors as shown in Model 2.

Model 2: $Y_2 = a_0U + a_1X_4 + a_2X_5 + a_3X_6 + E_2$

Where: $Y_2$ = the total score for Test 1 using the experimental grading procedure

$X_4$ = 1 if the subject received no points for item #1 on Test 1, 0 otherwise

$X_5$ = 1 if the subject received one point for item #1 on Test 1, 0 otherwise

$X_6$ = 1 if the subject received two poinst for item #1 on Test 1, 0 otherwise

$U$ = 1 if the subject was in the sample, 0 otherwise

$a_0, a_1, a_2, a_3$ = partial regression weights

$E_2$ = error vector ($Y_2 - \hat{Y}_2$)

Seventy-five such models were calculated, one for each of the twenty-five items on each of the three tests.

The results of the item discrimination analyses calculated for both the traditional and experimental grading systems are presented in Tables 3, 4, and 5. Table 3 presents the item discriminations for the twenty-five items in Test 1. As can be seen, when comparing these methods, the experimental method produced higher absolute item discrimination values fifteen out of the twenty-five items on Test 1 (Sign Test not significant).

Table 4 presents the item discriminations for Test 2. Here the experimental method only produced higher absolute item discrimination values ten out of the twenty-five times. (Sign Test not significant). Table 5 presents item discriminations for Test 3. In twenty out of twenty-five item discriminations, the absolute value was higher for the experimental scoring procedure. Unfortunately, one cannot truly interpret these item discrimination results since the computer program employed for calculating R only prints out $R^2$. To arrive at R, the square root of $R^2$ was taken; therefore, all of the R presented in Tables 3, 4, and 5 are positive values and we did not determine if any of these values should have been negative. Since negative item discrimination values are not desirable, and since we could not discern which items, if any, should have been negative for the experimental method of grading, the results in Tables 3, 4, and 5 should be looked at cautiously. (However, one should note that only 2 items of the 75 scored traditionally produce negative values).

Since the experimental method of grading required that the students respond twice to every test item, it was felt that this method may have produced a different testing situation which would result in different overall test scores. This was originally hypothesized by one of the authors while administering the test. He observed students verbal and non-verbal behavior indicating that they found the experimental testing procedure to be much more difficult. In the summer, 1973, to check on this possible effect, the authors randomly assigned the two different grading procedures to each of half of the two class sections of undergraduate tests and measurements. In each section, half of the students were taking the test traditionally and the other half of the students were taking it experimentally. Both tests were then graded, using the traditional grading procedures. These results are presented in Table 6.

The mean number of right answers for both procedures was approximately 18, and the standard deviation for the traditional procedure was approximately 3.4, and 3.0 for the experimental. These results indicate that the two procedures are not producing different testing situations.

The results of this study may have been unable to fully demonstrate the potential increase in effectiveness of the experimental grading over the traditional method because some of the validity criteria (objectives and project) were lost. This loss was partially due to the students being given access to their projects which resulted in some just taking their project. A quick evaluation indicated that the projects that tended to be taken were the ones receiving the lowest test grades. This may have seriously affected our range of scores. Since our theoretical position was that the experimental method would be more sensitive in detecting partial knowledge and would therefore be better able to detect differing ability levels, then restricted ranges would severely handicap the experimental method's ability to demonstrate its effectiveness.

One should note when reading the results that shrinkage estimates were employed for the total test validity results, but they were not calculated for item validities that were reported. This should be taken into account when interpreting the results. The item validities can be found in Appendix A, and it was felt that the total test validities were of greater importance.

One should also note that the item discrimination using the multiple regression procedures were not corrected for shrinkage. This was not done because of a time factor but they theoretically should be calculated. However, one should also consider that the standard method (r) used to calculate item discrimination and item validities have not been, and generally are not corrected for shrinkage.

Another consideration, as pointed out by Uhl and Eisenberg (1970) and Newman (1973), is that there are variations between shrinkage estimate formulas. Wherry's formula, which is most commonly used, was employed for calculating shrinkage estimates for this study. One should consider using Lord's (1950) formula for a shrinkage estimate for both R and r.

In this study, an attempt was made to develop a multivariable approach for improving item validities. It seems that if such an approach is further explored one would also have to develop multivariable and multivariate[1] methods for determining reliability. If one developed a multivariate technique for improving item discrimination and item validity and still used the traditional univariable technique for calculating reliability, this would be highly inconsistent. We would like to suggest that a modification of the canonical correlation procedure may be appropriate for developing a multivariate techniquo for estimating reliability which would be consistent with the approach suggested in the paper for improving validity.

In conclusion, we believe that multiple regression procedures will allow one to maximally use the available existing information produced by the probabilistic responses from examinees to determine validity estimates. The traditionally-used univariable technique will only produce one weight which is calculated to maximize it's prediction. Therefore, it is potentially much less effective than a technique that is capable of calculating a number of separate weights for maximizing prediction. In addition, working with univariable techniques may tend to fixate researchers to thinking in univariable terms, while in our estimation, multivariate and multivariable techniques are less confining and therefore are more likely to facilitate more creative and potentially more useful research. We believe multiple regression gave us the freedom which helped us conceptually derive a potentially useful method of grading and analyzing our results.

Table #2

Validity Criterion (Project Scores)

| Test | Method 1 (Trad. r) | Method 2 (Exp. r) | Method 3 (Exp. R) | Method (Exp. R |
|---|---|---|---|---|
| 1 (N=54) | .110 | .37 | .267 | .187 |
| 2 (N=52) | .041 | .204 | .299 | .229 |
| 3 (N=55) | .237 | .198 | .315 | .253 |

Note: See Table #1 for descriptions of methods

Table #3
Item Discriminations for Test #1

| Items | Traditional Scoring (r) pt. Bis | Experimental Scoring (R) | Items | Traditional Scoring (r) pt. Bis. | Experimental Scoring (R) |
|---|---|---|---|---|---|
| 1 | .339 | .309 | 14 | .421 | .489 |
| 2 | .159 | .143 | 15 | .404 | .381 |
| 3 | .265 | .301 | 16 | .157 | .261 |
| 4 | .202 | .297 | 17 | .066 | .103 |
| 5 | .430 | .356 | 18 | .076 | .238 |
| 6 | .218 | .281 | 19 | .275 | .427 |
| 7 | .437 | .317 | 20 | .066 | .179 |
| 8 | .437 | .340 | 21 | .347 | .320 |
| 9 | .260 | .087 | 22 | .456 | .394 |
| 10 | .212 | .214 | 23 | .479 | .547 |
| 11 | .282 | .293 | 24 | .360 | .432 |
| 12 | .454 | .484 | 25 | .390 | .354 |
| 13 | .425 | .441 | | | |

Note: N=75

Table #4
Item Descriminations for Test #2

| Items | Traditional Scoring (r) pt. Bis. | Experimental Scoring (R) | Items | Traditional Scoring (r) pt. Bis. | Experi Scori (R) |
|---|---|---|---|---|---|
| 1 | .055 | .444 | 14 | .101 | .412 |
| 2 | .355 | .334 | 15 | .150 | .244 |
| 3 | .358 | .309 | 16 | .392 | .222 |
| 4 | .437 | .391 | 17 | -.040 | .246 |
| 5 | .438 | .380 | 18 | .436 | .315 |
| 6 | .435 | .181 | 19 | .318 | .311 |
| 7 | .058 | .200 | 20 | .433 | .367 |
| 8 | .226 | .214 | 21 | .585 | .408 |
| 9 | .517 | .337 | 22 | .431 | .520 |
| 10 | .375 | .348 | 23 | .417 | .218 |
| 11 | .111 | .352 | 24 | .481 | .401 |
| 12 | .354 | .260 | 25 | .133 | .141 |
| 13 | .131 | .309 | | | |

Note: N=75

Table #5
Item Discriminations for Test #3

| Items | Traditional Scoring (r) pt. Bis | Experimental Scoring (R) | Items | Traditional Scoring (r) pt. Bis. | Experimental Scoring (R) |
|---|---|---|---|---|---|
| 1 | .185 | .180 | 14 | .206 | .441 |
| 2 | .148 | .786 | 15 | .0 | .649 |
| 3 | .188 | .183 | 16 | .285 | .333 |
| 4 | .279 | .553 | 17 | .294 | .413 |
| 5 | .172 | .757 | 18 | .315 | .248 |
| 6 | .402 | .353 | 19 | .379 | .670 |
| 7 | .229 | .794 | 20 | .431 | .493 |
| 8 | .370 | .766 | 21 | .069 | .232 |
| 9 | .523 | .796 | 22 | .162 | .626 |
| 10 | .604 | .527 | 23 | .370 | .637 |
| 11 | .054 | .783 | 24 | .323 | .653 |
| 12 | .478 | .520 | 25 | .306 | .669 |
| 13 | .155 | .798 | | | |

Note: N=75

Table #6

Data from Summer Session 1, 1973
Controlling for Testing Situation Effect
for Sections 1 and 2 Combined

| | Traditional Testing Situation | Experimental Testing Situation |
|---|---|---|
| S | 3.4280 | 3.0220 |
| $\bar{X}$ | 18.4137 | 18.1515 |
| N | 29. | 33. |

Note: No test of significance was run since the data obviously would be nonsignificant at our alpha level of .05.

# APPENDIX A

## APPENDIX A

### Item Validity-Criterion: Objective

### Test 3

| # | Pt. Bis. | r | R | $R^2$ | # | Pt. Bis. | r | R | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|
| | -.1702 | -.1773 | .178 | .0315 | 14 | .1024 | .1544 | .195 | .0380 |
| | -.1201 | -.1201 | .120 | .0144 | 15 | 0.0 | 0.0 | | 0.0 |
| 3 | -.0380 | .0306 | .135 | .0182 | 16 | .1376 | .0721 | .233 | .0541 |
| 4 | .1847 | .0854 | .316 | .0988 | 17 | .1007 | .0383 | .201 | .0403 |
| 5 | .0863 | .0863 | .087 | .0075 | 18 | -.0208 | -.0143 | .027 | .0007 |
| 6 | -.0334 | -.0806 | .120 | .0143 | 19 | .1365 | .1365 | .136 | .0186 |
| 7 | .1169 | .0764 | .136 | .0185 | 20 | .2015 | .1244 | .264 | .0700 |
| 8 | -.0847 | -.1354 | .176 | .0311 | 21 | -.0156 | .1287 | .148 | .0219 |
| 9 | .1913 | .2576 | .363 | .1314 | 22 | .3117 | .3117 | .312 | .0972 |
| 0 | .1782 | .1226 | .180 | .0325 | 23 | .1278 | .0277 | .292 | .0854 |
| 1 | -.0388 | -.0388 | .039 | .0015 | 24 | .1058 | .0226 | .149 | .0223 |
| 2 | -.0169 | .0657 | .178 | .0317 | 25 | .1807 | .1975 | .174 | .0327 |
| 3 | .0072 | .0072 | .010 | .0001 | | | | | |

## APPENDIX A

### Item Validity-Criterion Objective

### Test 1

| # | Pt. Bis. | r | R | $R^2$ | # | Pt. Bis. | r | R | R |
|---|---|---|---|---|---|---|---|---|---|
| 1 | -.0644 | -.0661 | .006 | .0044 | 14 | .2081 | .1678 | .223 | .04 |
| 2 | -.2544 | -.2489 | .256 | .0657 | 15 | .2294 | .1959 | .236 | .95 |
| 3 | -.0203 | -.0203 | .02 | .0004 | 16 | .9363 | .2032 | .360 | .12 |
| 4 | -.2570 | .2470 | .257 | .0661 | 17 | -.0503 | -.0462 | .050 | .00 |
| 5 | .1941 | .2628 | .313 | .0980 | 18 | .0455 | .289 | .07 | .00 |
| 6 | .1368 | .2256 | .315 | .0991 | 19 | .0156 | .9483 | .06 | .00 |
| 7 | .0456 | .0971 | .224 | .0503 | 20 | -.1191 | -.1326 | .134 | .01 |
| 8 | -.0156 | -.0228 | .05 | .0025 | 21 | -.0201 | -.0775 | .157 | .02 |
| 9 | .0714 | .1053 | .32 | .0074 | 22 | -.0610 | -.0175 | .110 | .01 |
| 10 | .0465 | .0223 | .035 | .0073 | 23 | -.0063 | -.0395 | .078 | .006 |
| 11 | -.0538 | -.0296 | .082 | .0067 | 24 | .2235 | .1737 | .232 | .053 |
| 12 | .1315 | .1817 | .242 | .0583 | 25 | .2514 | .3249 | .353 | .124 |
| 13 | .3629 | .3324 | .364 | .1322 | | | | | |

## APPENDIX A

### Item Validity-Criterion: Objective

### Test 2

| Pt. Bis. | r | R | $R^2$ | # | Pt. Bis. | r | R | $R^2$ |
|---|---|---|---|---|---|---|---|---|
| .3314 | .3314 | .365 | .1332 | 14 | .1884 | .0411 | .151 | .0227 |
| -.1254 | -.0683 | .121 | .0146 | 15 | -.0993 | .1890 | .137 | .0189 |
| -.1502 | -.144 | .250 | .0627 | 16 | -.0021 | -.0539 | .335 | .1123 |
| .0324 | .1031 | .248 | .0615 | 17 | -.0394 | -.0394 | .074 | .0054 |
| -.0569 | -.0537 | .106 | .0113 | 18 | -.0638 | -.1038 | .303 | .0917 |
| .0587 | .0836 | .107 | .0148 | 19 | .0213 | .0849 | .210 | .0441 |
| -.0199 | -.0072 | .076 | .0057 | 20 | -.2729 | -.2019 | .105 | .0110 |
| -.0246 | .0761 | .187 | .0350 | 21 | -.0747 | .0130 | .161 | .0260 |
| .1696 | .1233 | .052 | .0027 | 22 | .1820 | .0830 | .166 | .0277 |
| -.2151 | -.1690 | .112 | .0126 | 23 | .2537 | .2010 | .391 | .1527 |
| .0089 | -.1639 | .204 | .0411 | 24 | -.0605 | -.0904 | .047 | .0022 |
| -.1261 | -.1525 | .135 | .0183 | 25 | -.1850 | -.1403 | .201 | .0404 |
| .0090 | -.1351 | .273 | .0743 | | | | | |

## APPENDIX A

### Item Validity-Criterion: Projects

Test 1

| # | Pt. Bis. | r | R | $R^2$ | # | Pt. Bis. | r | R | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | .1947 | .1115 | .310 | .0963 | 14 | .2463 | .2463 | .246 | .( |
| 2 | -.0677 | -.0064 | .201 | .0405 | 15 | .0189 | .0398 | .075 | .( |
| 3 | .1104 | .1104 | .110 | .0122 | 16 | .2119 | .2342 | .234 | .( |
| 4 | -.1374 | -.1765 | .227 | .0515 | 17 | -.2748 | -.2415 | .281 | .0 |
| 5 | .1804 | .1804 | .181 | .0326 | 18 | .1383 | .1223 | .140 | .0 |
| 6 | .0719 | .0719 | .072 | .0052 | 19 | -.1033 | .0063 | .179 | .0 |
| 7 | .3221 | .3222 | .326 | .1062 | 20 | .1080 | .1495 | .182 | .0 |
| 8 | .0189 | .0528 | .140 | .0196 | 21 | .0947 | .0939 | .096 | .0 |
| 9 | .1336 | .0928 | .033 | .0011 | 22 | .0509 | .0860 | .111 | .0 |
| 10 | -.2101 | -.1630 | .232 | .0538 | 23 | -.1100 | -.0822 | .105 | .0 |
| 11 | -.0599 | -.0566 | .060 | .0036 | 24 | .0220 | .0666 | .110 | .0 |
| 12 | .0036 | .0114 | .0002 | .0006 | 25 | -.0111 | -.0017 | .003 | .0( |
| 13 | .0899 | .0862 | .090 | .0081 | | | | | |

## APPENDIX A

### Item Validity-Criterion: Projects

### Test 2

| # | Pt. Bis. | r | R | $R^2$ | # | Pt. Bis. | r | R | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | .2512 | .2361 | .318 | .1001 | 14 | .1201 | .0884 | .098 | .0096 |
| 2 | -.0341 | .0635 | .296 | .0874 | 15 | .1423 | .1923 | .179 | .0321 |
| 3 | .1357 | .1232 | .094 | .0088 | 16 | .0275 | .0275 | .0027 | .0007 |
| 4 | .2010 | .2119 | .224 | .0501 | 17 | -.0038 | -.0038 | .047 | .0022 |
| 5 | .1623 | .1524 | .171 | .0291 | 18 | -.2535 | -.1848 | .189 | .0357 |
| 6 | .2450 | .2555 | .242 | .0586 | 19 | -.1313 | -.0570 | .179 | .0321 |
| 7 | .1423 | .1423 | .179 | .0321 | 20 | -.1710 | -.2092 | .164 | .0268 |
| 8 | -.0340 | .0239 | .233 | .0540 | 21 | -.1167 | -.0251 | .284 | .0808 |
| 9 | .3040 | .2774 | .180 | .0325 | 22 | .1175 | .0612 | .098 | .0096 |
| 10 | -.0965 | -.1528 | .172 | .0297 | 23 | .1486 | .1683 | .127 | .0162 |
| 11 | -.2525 | -.3512 | .230 | .0527 | 24 | -.0747 | -.1095 | .106 | .0112 |
| 12 | -.0596 | -.1635 | .193 | .0372 | 25 | -.0384 | -.2323 | .217 | .0470 |
| 13 | .1222 | .0075 | .321 | .1031 | | | | | |

# APPENDIX A

Item Validity-Criterion: Projects

Test 3

| # | Pt. Bis. | r | R | $R^2$ | # | Pt. Bis. | r | R | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | .1656 | .353 | .201 | .0405 | 14 | .3654 | .3678 | .373 | .13 |
| 2 | -.1187 | -.1416 | .148 | .0219 | 15 | 0 | 0 | | 0.0 |
| 3 | .0166 | 0 | .033 | .0011 | 16 | .2160 | .1397 | .307 | .09 |
| 4 | .0431 | -.0325 | .211 | .0445 | 17 | .2274 | .2561 | .260 | .06 |
| 5 | .1017 | .1017 | .101 | .0103 | 18 | .1929 | .0791 | .335 | .11 |
| 6 | -.0371 | -.0441 | .045 | .0020 | 19 | .1744 | .1744 | .174 | .030 |
| 7. | -.1783 | -.1009 | .232 | .0538 | 20 | .1944 | .2171 | .219 | .047 |
| 8 | -.0967 | -.1169 | .148 | .0222 | 21 | -.0455 | .0651 | .065 | .004 |
| 9 | .0352 | .0946 | .143 | .0205 | 22 | .1744 | .1744 | .174 | .030 |
| 10 | .2474 | .2095 | .250 | .0624 | 23 | .1379 | .0052 | .186 | .034 |
| 11 | .1744 | .1744 | .174 | .0304 | 24 | -.0755 | -.0290 | .066 | .004 |
| 12 | .1258 | .0858 | .178 | .0315 | 25 | .0422 | -.0020 | .108 | .011 |
| 13 | -.0455 | -.0455 | .213 | .0455 | | | | | |

## Reference


Archer, N. S., "A Comparison of the Conventional and Two Modified Procedures for Responding to Multiple-Choice Items with Respect to Test Reliability, Validity, and Item Characterists." Unpublished Doctorial Dissertation. Syracuse University, 1962.

Coombs, Milholland, Womer, "The Assessment of Partial Knowledge." Educational and Psychological Measures, 1965, 16, 13-37.

Ebel, "Confidence Weighing and Test Reliability." Journal of Educational Measurement, 1965, 2, 49-57.

Echternacht, B., "The Use of Confidence Testing in Objective Tests." Review of Educational Research, 1972, 42, 2, 217-236.

Hambleton, Roberts, Traub, "A Comparison of the Reliability and Validity of Two Methods for Assessing Partial Knowledge on a Multiple-Choice Test." Journal of Educational Measurement. 1970, 7, 75-82.

Jocobs, Stanly S., "Correlation of Unwarranted Confidence in Responses to Objective Items." Journal of Educational Measurement, Vol. 8, sp 1971.

Kelly, Beggs, McNeil, Eichelberger and Lyon, Research Design in the Behavioral Sciences: Multiple Regression Approach, Southern Illinois University Press, 1969.

Koehler, Roger O., "A Comparison of the Validities of Conventional Choice Tests and Various Confidence Marking Procedures." Journal of Educational Measurement, Vol. 8, No. 4, Winter, 1971.

Lord, Novick, Statistical Theories of Mental Test Scores. Addison-Wesley, 1968.

Newman, Isadore, "Variations Between Shrinkage Estimation Formulas and the Appropriateness of Their Interpretation." Multiple Linear Regression Viewpoints. Vol. 4, 2, 45-48, 1973.

Uhl, N. and Eisenbert, T., "Predicting Shrinkage in the Multiple Correlation Coefficient." Education and Psychological Measurement, 30, 487-489, 1970.

# A Simple Multiple Linear Regression Test for Differential Effects of a Given Independent Variable on Several Dependent Measures

Jerry A. Colliver
Steven J. Verhulst
Paul Kolm
Southern Illinois University
School of Medicine

ABSTRACT

Multiple linear regression may be used to determine whether an independent variable of interest has a differential effect on two or more dependent variables. The initial step involves the separate standardization of each dependent variable. The values of the standardized dependent variables are pooled and treated for purposes of the analysis as constituting a single dependent variable. A within subjects independent variable is created and the levels of the variable are used to denote the different dependent variables. The data are analyzed with a split-plot analysis of variance for which the independent variable of interest is the between groups factor and the independent variable which distinguishes the dependent variables is the within subjects factor. The test of the interaction of these two factors provides a statistical determination of whether the independent variable of interest has a differential effect on the two or more dependent variables.

A problem we have encounted on several occassions can be dealt with easily by using an interesting "twist" on multiple linear regression procedures. The problem involves the determination of whether a given independent variable has different effects on several dependent measures. For example, most recently, we were asked to determine if the dosage of a given drug administered to animals injected with tumor cells had different effects on tumor size and body weight. To make this determination, we separately standardized each of the two dependent variables, tumor size and body weight, pooled these standardized values, and treated the two standardized variables as if they constituted one dependent measure. The two standardized dependent variables were distinguished via a within subjects, independent variable (called Outcome Measure), which we created for the purpose. This within subjects, independent variable had two levels which denoted the two standardized dependent variables, respectively. A split-plot analysis of variance (ANOVA) was performed and the test of the Dosage X Outcome Measure interaction provided a simple test of whether Dosage had different effects on the two outcome measures, tumor size and body weight.

## PROCEDURE

The procedure can be illustrated with a set of simulated data used to stimulate "solutions" for discussion purposes at a recent Multiple Linear Regression Special Interest Group session (Leitner, 1986). (See Appendix A.) Data were generated for n = 30 hypothetical subjects on five continuous variables (Y, X, U, V and W) and three dummy variables (D1, D2 and D3). For the purpose of illustrating the procedure, the five continuous variables were regarded as dependent variables. Each was standardized, and the five standardized variables were subsequently treated for purposes of the analysis as representing one dependent variable. The five variables were distinguished by considering each variable as if it represented one level of an artificially created independent variable, Outcome Measure.

The three dummy variables, D1, D2, and D3, were treated as if they represented one independent variable called Treatment with levels represented by the binary code expressed by the three dummies. Using this procedure the independent variable was found to have four levels represented by the binary codes, 000, 010, 100, and 111. Thus, the four levels of the Treatment independent variable were 0, 2, 4, and 8.

A 4 X 5 split-plot analysis of variance with one between subjects variable (Treatment with four levels, 0, 2, 4, and 8) and one within subjects variable (Outcome Measure with five levels, Y, X, U, V, and W) was performed on the simulated data. Treatment represented the independent variable of interest and Outcome Measure represented the independent variable used to distinguish the five standardized dependent variables.

## RESULTS

The results showed a significant Treatment X Outcome Measure interaction, indicating that Treatment had different effects on the different outcome measures, $F(12,104) = 2.21$; $p = 0.0448$. Simple interaction effects tests showed that the effect of Treatment on the dependent variable W differed significantly from the effects of Treatment on the other four dependent variables, Y, X, U, and V, and that the effects of Treatment on the four dependent variables, Y, X, U, and V, did not differ significantly. A graph of the relationship between Treatment and the five dependent variables is presented in Figure 1, which shows that variable W decreased from Treatment level 0 to level 2 to level 4 and remained fairly stable from level 4 to level 8. Variables Y, X, U, and V decreased from level 0 to level 2, increased from level 2 to level 4 to level 8.

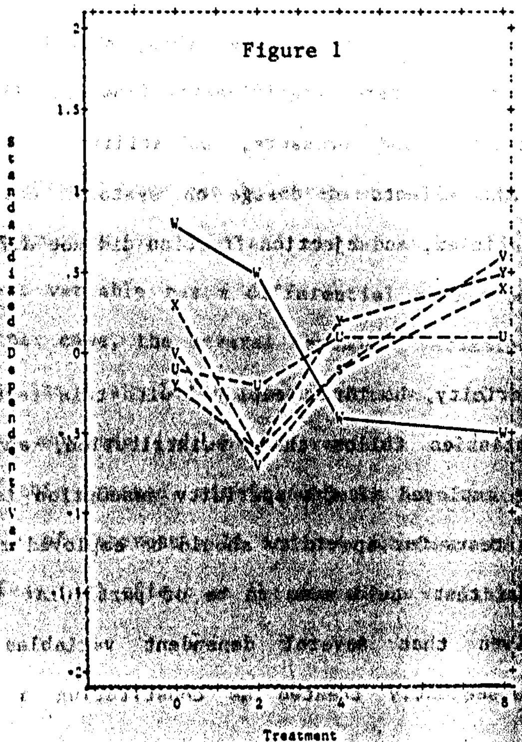


Figure 1

## DISCUSSION

The results showed that the independent variable, Treatment, had significantly different effects on the five dependent variables, Y, X, U, V, and W. To give substance to this example, suppose that the Treatment independent variable with four levels represented the dosage of some drug such as ethanol, epinepherine, streptokinease, etc. and that the four dosages were 0, 2, 4, and 8 units. Further suppose that the five dependent variables were as follows: Y, systolic blood pressure; X, diastolic blood pressures; U, pulsatility index; V, ejection fraction; and W, heart rate. The research hypothesis, then, would state that drug dosage has a differential effect on the five dependent variables, and the null hypothesis would be $H_0$: $\sigma^2$(interaction) = $\sigma^2$(error). Our results, then, showed that the effect of drug dosage on heart rate differed significantly from the effects of dosage on systolic and diastolic blood pressure, pulsatility index, and ejection fraction but that the effects of dosage on systolic and diastolic blood pressure, pulsatility index, and ejection fraction did not differ significantly from one another.

The test for spericity should be employed with this test to determine if the computed F statistics follow the F distribution, and an appropriate adjustment should be employed if the spericity assumption is violated (Kirk, 1982). Although the tests for spericity should be employed routinely with any split-plot ANOVA, the test would seem to be of particular importance in the present context given that several dependent variables are separately standardized and subsequently treated as constituting a single dependent variable.

The reader will undoubtedly notice the similarity between the procedure outlined here and the more commonly known profile analysis (Morrison, 1967). The difference in emphasis and orientation between this procedure and profile analysis, however, would seem to warrant separate consideration of the procedure described here. Profile analysis focuses on the comparison of profiles of means of several variables for two or more groups. The typical example involves the comparison of profiles of means on psychological tests in a test battery for groups of patients with different psychiatric diagnoses. The typical graphic representation depicts a profile of test (dependent variable) means plotted separately for each group. The procedure outlined here, on the other hand, involves the comparison of the effects of an independent variable on several dependent variables, with a graphic representation that depicts the effect of the independent variable on each dependent variable separately (see Figure 1).

The procedure outlined here can be extended to designs with more than one between groups, independent variable and can be used to determine if a within subjects independent variable has a differential effect on several dependent variables. In either case, the several dependent variables are standardized, treated as constituting a single dependent variable, and distinguished by the levels of a within subjects independent variable created for that purpose. The interaction of this created, within subjects independent variable and the independent variable of interest will indicate whether the latter independent variable has a differential effect on the dependent variables.

## REFERENCES


Kirk, R.E. (1982). Experimental design: Procedures for the Behavioural Sciences (2nd ed.). Belmont, CA: Brooks/Cole.

Leitner, D.W. (1986). Data set for you to analyze. Call for papers for Multiple Linear Regression Special Interest Group, American Education Research Association, San Francisco, CA.

Morrison, D.F. (1967). Multivariate statistical methods. New York: McGraw-Hill.

| OBS | Y | X | U | V | W | D1 | D2 | D3 |
|---|---|---|---|---|---|---|---|---|
| 1 | 54 | 47 | 49 | 62 | 41 | 0 | 1 | 0 |
| 2 | 42 | 55 | 64 | 56 | 66 | 1 | 0 | 0 |
| 3 | 64 | 61 | 47 | 81 | 49 | 0 | 1 | 0 |
| 4 | 48 | 45 | 63 | 55 | 46 | 0 | 0 | 0 |
| 5 | 5 | 21 | 93 | 31 | 62 | 1 | 1 | 1 |
| 6 | 42 | 46 | 11 | 50 | 53 | 1 | 0 | 0 |
| 7 | 40 | 55 | 14 | 54 | 67 | 1 | 1 | 1 |
| 8 | 62 | 55 | 26 | 74 | 44 | 1 | 1 | 1 |
| 9 | 45 | 56 | 13 | 59 | 63 | 1 | 1 | 1 |
| 10 | 47 | 43 | 52 | 52 | 44 | 1 | 0 | 0 |
| 11 | 61 | 69 | 96 | 83 | 63 | 0 | 1 | 0 |
| 12 | 62 | 69 | 11 | 84 | 62 | 1 | 0 | 0 |
| 13 | 54 | 41 | 30 | 58 | 33 | 0 | 0 | 0 |
| 14 | 47 | 47 | 83 | 55 | 49 | 1 | 0 | 0 |
| 15 | 48 | 38 | 69 | 51 | 36 | 0 | 1 | 0 |
| 16 | 87 | 78 | 49 | 47 | 47 | 0 | 1 | 0 |
| 17 | 47 | 51 | 31 | 58 | 55 | 1 | 0 | 0 |
| 18 | 73 | 49 | 35 | 40 | 70 | 1 | 1 | 1 |
| 19 | 49 | 49 | 73 | 58 | 50 | 0 | 0 | 0 |
| 20 | 40 | 43 | 92 | 46 | 53 | 1 | 0 | 0 |
| 21 | 54 | 44 | 47 | 50 | 37 | 0 | 0 | 0 |
| 22 | 52 | 49 | 54 | 61 | 47 | 0 | 0 | 0 |
| 23 | 48 | 47 | 70 | 56 | 48 | 0 | 1 | 0 |
| 24 | 40 | 45 | 10 | 65 | 34 | 0 | 1 | 0 |
| 25 | 37 | 43 | 96 | 43 | 56 | 1 | 1 | 1 |
| 26 | 39 | 40 | 26 | 43 | 48 | 1 | 1 | 1 |
| 27 | 46 | 47 | 56 | 54 | 49 | 1 | 1 | 1 |
| 28 | 62 | 58 | 48 | 76 | 48 | 0 | 1 | 0 |
| 29 | 46 | 44 | 53 | 52 | 47 | 1 | 1 | 1 |
| 30 | 35 | 40 | 63 | 39 | 54 | 1 | 1 | 1 |

Appendix A. Simulated Data from Multiple Linear Regression Special Interest Group Session (Leitner, 1986).

If you are submitting a research article other than notes or comments, I would like to suggest that you use the following format if possible:

Title
Author and affiliation
Indented abstract (entire manuscript should be single spaced)
Introduction (purpose—short review of literature, etc.)
Method
Results
Discussion (conclusion)
References

All manuscripts should be sent to the editor at the above address. (All manuscripts should be camera-ready.)

It is the policy of the M.L.R. SIG-multiple linear regression and of *Viewpoints* to consider articles for publication which deal with the theory and the application of multiple linear regression. Manuscripts should be submitted to the editor as original, double-spaced, *camera-ready copy.* Citations, tables, figures and references should conform to the guidelines published in the most recent edition of the *APA Publication Manual* with the exception that figures and tables should be put into the body of the paper. A cost of $1 per page should be sent with the submitted paper. Reprints are available to the authors from the editor. Reprints should be ordered at the time the paper is submitted, and 20 reprints will cost $.50 per page of manuscript. Prices may be adjusted as necessary in the future.

A publication of the Multiple Linear Regression Special Interest Group of the American Educational Research Association, *Viewpoints* is published primarily to facilitate communication, authorship, creativity and exchange of ideas among the members of the group and others in the field. As such, it is not sponsored by the American Educational Research Association nor necessarily bound by the association's regulations.

"Membership in the Multiple Linear Regression Special Interest Group is renewed yearly at the time of the American Educational Research Association convention. Membership dues pay for a subscription to the *Viewpoints* and are either individual at a rate of $5, or institutional (libraries and other agencies) at a rate of $18. Membership dues and subscription requests should be sent to the executive secretary of the M.L.R. SIG."

THE UNIVERSITY OF AKRON
AKRON, OH 44325

ROGERS, BRUCE G.
DEPT OF ED PSYCH
UNIV OF NO IOWA
CEDAR FALLS, IOWA 50613

AKRON APR-9'87 OHIO
U.S. POSTAGE 0.69
P.B. METER 6035302

BOOKS - SPECIAL 4th CLASS RATE

## TABLE OF CONTENTS



ISSN 0195-7171

The University of Akron is an Equal Education and Employment Institution