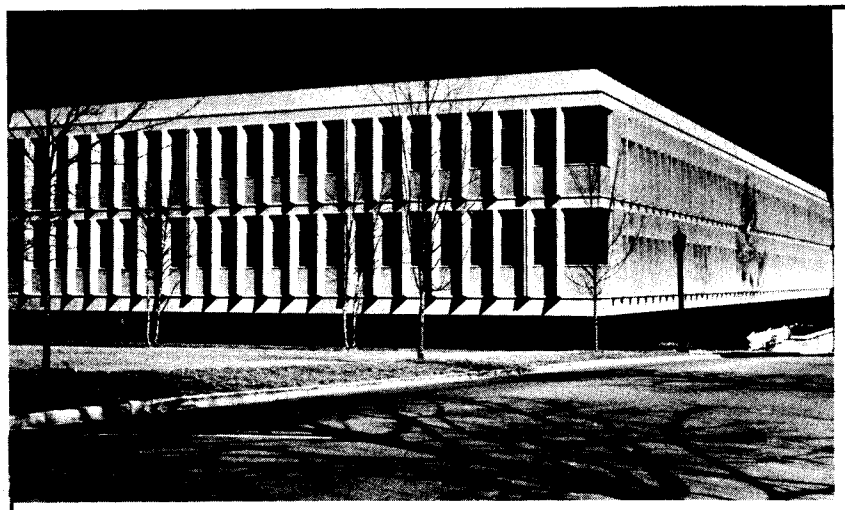

Multiple and Canonical Regression with a Mix of Qualitative and Quantitative Variables: An Alternating Least Squares Method with Optimal Scaling Features

Forrest W. Young

Jan de Leeuw

and

Yoshio Takane

November, 1975

Report Number 146

THE L. L. THURSTONE
PSYCHOMETRIC LABORATORY
UNIVERSITY OF NORTH CAROLINA

CHAPEL HILL, N. C.
27514

Abstract

A method is discussed which extends canonical regression analysis to the situation where the variables may be measured at a variety of levels (nominal, ordinal, or interval), and where they may be either continuous or discrete. There is no restriction on the mix of measurement characteristics (i.e., some variables may be discrete-ordinal, others continuous-nominal, and yet others discrete-interval). The method, which is purely descriptive, scales the observations on each variable, within the restriction imposed by the variable's measurement characteristics, so that the canonical correlation is maximal. The alternating least squares algorithm is discussed. Several examples are presented. It is concluded that the method is very robust. Inferential aspects of the method are not discussed.

Current addresses:

Jan de Leeuw, Datatheorie, Central Rekeninstituut, Wassensaarseweg 80, Leiden, The Netherlands

Yoshio Takane, Department of Psychology, University of Tokyo, Tokyo, Japan

---

This research was partially supported by grants MH10006 and MH26504 from the National Institute of Mental Health to the Psychometric Laboratory of the University of North Carolina. Portions of the research reported here were presented to the spring meeting of the Psychometric Society, 1975. We wish to thank John B. Carroll and Elliot M. Cramer for their critical evaluations of an earlier draft of this report, and Jack Hoadley and John B. Carroll for letting us use their data. Copies of the paper may be obtained from the first author.

0.0 Summary

In many areas of social and behavioral sciences, the investigator has obtained observations on a variety of variables from a number of respondents and wishes to investigate their structure with the use of multiple (or, occasionally, canonical) correlation techniques. But the investigator has a problem. He knows that his variables do not measure up, as it were, to the assumptions required by the multiple correlation technique. Some of his variables are measured at the interval level, as is conventionally required, but others are ordinal and yet others are (heaven forbid) only nominal. What does he do in the face of this dilemma? He proceeds as though there were no problem at all. After all, what other alternatives are there? There is no completely appropriate alternative, as the only multiple correlation techniques which permit qualitative variables assume that all variables are qualitative, or that the independent variables are quantitative while the dependent variable is qualitative. In the former situation the variables must all be nominal (Hayashi, 1950), or all ordinal (Lingoes, 1973; de Leeuw, 1973), with no mixing of levels and no quantitative variables. In the latter situation the dependent variable may be ordinal (de Leeuw, 1969; Carroll, 1972; Srinivasan, 1974), nominal (discriminant analysis) or binary (de Leeuw, 1968). With the exception of discriminant analysis, no procedures have been proposed which permit several dependent variables which may be qualitative.

The work presented in this paper is designed with the above situation in mind. With the Multiple Optimal Regression by Alternating Least Squares (MORALS) technique, and the corresponding

canonical regression technique (CORALS), the investigator with variables defined at a variety of measurement levels can investigate their structure while he respects the various levels of measurement. MORALS optimizes the multiple correlation between a single criterion variable and a set of predictor variables where any of the variables (criterion included) may be nominal, ordinal or interval. The variables do not all have to be measured at the same level; any mixture will do. Also, the process assumed to have generated the data may be either discrete or continuous. As will be explained, MORALS obtains an optimal scaling for each variable within the restrictions imposed by the regression model, the measurement level, and the generating process. The scaling is optimal in the Fisher (1938) sense of optimal scaling: the multiple correlation is maximized.

CORALS is very similar to MORALS, except that it optimizes the canonical correlation between two sets of variables. Since MORALS is a special case of CORALS (where one of the sets of variables consists of a single variable) we refer to the algorithm as the MORALS/CORALS algorithm.

In the companion paper (de Leeuw, Young & Takane, 1975) we discussed in detail the simple additive situation where one has obtained qualitative data in a factorial design and where one is only interested in the main effects. The previous work, then, was restricted to univariate qualitative data obtained in factorial experiments, and was restricted to the additive model with no interaction terms. In this paper we remove all of these limitations. We discuss the general linear model as it applies to multivariate

data where the variables may be defined at any mixture of measurement levels.

As was discussed in the companion paper, the analysis of additivity has usually been introduced in the context of a statistical model which places very stringent and oftentimes unrealistic assumptions on the nature of the data. As a result of these assumptions it is possible to develop inferential procedures. Our purpose is not to develop an inferential procedure, but rather to develop a descriptive procedure to provide the investigator with tools for investigating the additive structure of his qualitative data. While it would be desirable to develop a descriptive procedure based on assumptions commensurate with qualitative variables, such a development is not presented in this paper.

1.0 Mathematical Developments

1.1 Notation

We use bold-face capital letters to represent matrices (i.e., $\underline{X}$); bold-face lower case letters for vectors ($\underline{x}$); and regular lower case letters for scalars (x). Note that all vectors are assumed to be column vectors, with a row vector denoted as the transpose of a column vector ($\underline{x}'$). We refer to a specific column vector of a matrix as $\underline{x}_i$; a specific element of a matrix as x_{ai}; and a specific element of a vector as x_a. We further reserve Greek letters for parameters (i.e., $\underline{\beta}$ is a vector of parameters) and script letters for functions (i.e., $\mathcal{t}$ is a transformation).

1.2 The problem

Let there be two matrices of observation variables $\underline{X}$ and $\underline{Y}$, there being n variables $\underline{x}_i$ and m variables $\underline{y}_j$, all having k

observations (we will consider missing data later). We assume that each $\underline{x}_i$ and $\underline{y}_j$ is measured at some known measurement level, with each element being subject to certain measurement restrictions. There need be no particular relationship of measurement levels between variables, although we assume that all observations on a single variable are at the same measurement level (an assumption to be relaxed later).

Let us further define two vectors of parameters $\underline{\alpha}$ and $\underline{\beta}$, and two matrices $\underline{X}^*$ and $\underline{Y}^*$, where $\underline{\alpha}$ has n elements α_i, $\underline{\beta}$ has m elements β_i, $\underline{X}^*$ has n columns and k rows, and $\underline{Y}^*$ has m columns and k rows. The columns $\underline{x}_i^*$ and $\underline{y}_j^*$ correspond to the observation variables $\underline{x}_i$ and $\underline{y}_j$. Furthermore, the columns $\underline{x}_i^*$ and $\underline{y}_j^*$ have two important characteristics: a) all $\underline{x}_i^*$ and $\underline{y}_j^*$ are defined at the interval level of measurement; and b) each $\underline{x}_i^*$ and $\underline{y}_j^*$ is related to its observation variable $\underline{x}_i$ and $\underline{y}_j$ by a transformation which completely satisfies the measurement characteristics of the observation variable. That is

$$(1) \qquad \begin{aligned} \underline{x}_i^* &= t_i[\underline{x}_i], \\ \underline{y}_j^* &= t_j[\underline{y}_j] \end{aligned}$$

where t_i and t_j are called the measurement transformations, and are collectively referred to as t when there is no resulting confusion. Of course, the t are subject to restraints by the measurement level and process of their variables. These restraints are discussed in the next section. We may correctly think of $\underline{x}_i^*$ and $\underline{y}_j^*$ as being the observation variables rescaled at the interval level of measurement so that the correlation is optimized. Thus, we will oftentimes refer to $\underline{x}_i^*$ and $\underline{y}_j^*$ as the optimally scaled observations.

The problem we wish to solve can now be stated: We wish to obtain transformations t of each observation variable $\underline{x}_i$ and $\underline{y}_j$, as well as regression weights $\underline{\alpha}$ and $\underline{\beta}$ so that the canonical correlation between $\underline{X}^*$ and $\underline{Y}^*$ is as high as possible. Since maximizing the canonical correlation is equivalent to minimizing the sums of squared differences between two composite variables (under suitable normalization assumptions), we define composite variables $\underline{a}$ and $\underline{b}$ such that

$$\begin{aligned} \underline{a} &= \underline{X}^*\underline{\alpha} \\ \underline{b} &= \underline{Y}^*\underline{\beta} \end{aligned} \tag{2}$$

and state our goal as the minimization of

$$\begin{aligned} \lambda^2 &= (\underline{a}-\underline{b})'(\underline{a}-\underline{b}) \\ &= (\underline{X}^*\underline{\alpha}-\underline{Y}^*\underline{\beta})'(\underline{X}^*\underline{\alpha}-\underline{Y}^*\underline{\beta}) \end{aligned} \tag{3}$$

where the normalization restrictions are

$$\underline{1}'\underline{x}_i^* = \underline{1}'\underline{y}_j^* = 0 \quad (i=1,\ldots,n;\ j=1,\ldots,m) \tag{a}$$

$$\frac{1}{k}\underline{x}_i^{*\prime}\underline{x}_i^* = \frac{1}{k}\underline{y}_j^{*\prime}\underline{y}_j^* = 1 \tag{b}$$

(4)

$$\frac{1}{k}\underline{a}'\underline{a} = 1 \tag{c}$$

$$\frac{1}{k}\underline{b}'\underline{b} = 1. \tag{d}$$

Of course, $\underline{x}_i^*$ and $\underline{y}_j^*$ are also subject to the measurement restrictions of equation (1). Note that we minimize λ and not λ normalized by by the variance of $\underline{a}$ (or $\underline{b}$), restriction (c) making such a normalization unnecessary as has been shown by de Leeuw (1975). Note that the canonical correlation between the two sets of variables is defined as

$$R = \frac{1}{k}\underline{a}'\underline{b}.$$

1.3 Restrictions

In this section, we discuss the types of restrictions imposed on the relationship between the raw observations in matrices $\underline{X}$ and $\underline{Y}$ and the rescaled observations represented by matrices $\underline{X}^*$ and $\underline{Y}^*$; that is, we discuss restrictions of the transformations in equation (1). We will see that the restrictions are of three types; a) those concerned with identifying the parameters $\underline{\alpha}$ and $\underline{\beta}$ (the normalization restrictions); b) those concerning the measurement level of the observation variables; and c) those concerning the underlying process which generated the observations on each variable. Of course, it does not make sense to impose no restrictions on the several t, for if we do not, they are free to be of any form, in which case we could trivially solve equation (3) by setting $\underline{a}=\underline{b}$ (which is, of course, not unique).

The minimal set of restrictions always involves the normalization restrictions imposed by equation (4). Though these restrictions are trivial and not very interesting, they are of crucial importance since without them unique values for $\underline{\alpha}$ and $\underline{\beta}$ are undefinable. If we did not invoke these restrictions, then we could set all X* and Y* to zero, with a resulting perfect (but trivial) fit. While these restrictions are crucial, they are generally called non-restrictive constraints as they have no essential effect on the value of λ. There is no "best" normalization, there is only the "conventional" normalization herein adopted.

Of much greater interest are the measurement restrictions placed on the transformations t (equation 1). These restrictions concern two different aspects of the measurement situations, the level of measurement of the observation variable (i.e., nominal, ordinal,

interval or ratio) and the nature of the process which generated the observations (discrete or continuous).

To fully understand the several level and process restrictions, we must first introduce a concept which is crucial to our work: It is our view that all observations are categorical. That is, we view an observation variable as consisting of observations which fall into a variety of categories, such that all observations in a particular category are empirically equivalent. Furthermore, we take this "categorical" view regardless of the variable's measurement level and regardless of the nature of the process which generated the observations. Put more simply, it is our view that the observational process delivers observations which are categorical because of the finite precision of the measurement and observation process, if for no other reason. For example, if one is measuring temperature with an ordinary thermometer (which is likely to generate interval level observations reasonably assumed to reflect a continuous process) it is doubtful whether the degrees are reported with any more precision than whole degrees. Thus, the observation is categorical: there are a very large (indeed infinite) number of uniquely different temperatures which would all be reported as say, 40°. Thus, we say that the observation of 40° is categorical.

As we will see, the two types of measurement restrictions (level and process restrictions) concern two different aspects of the observation categories. The process restrictions concern the relationships among all the observations <u>within</u> a single category, whereas the level restrictions concern the relationships among all the observations <u>between</u> different categories. We will first take up the process restrictions, and then the level restrictions.

There are two types of process restrictions, one invoked when we assume that the generating process is discrete, and the other when we assume a continuous process. One or the other assumption must always be made. If we believe that the process is discrete then all observations within a particular category should be represented by the same real number after the transformation t has been made. On the other hand, if we adopt the continuous assumption then each of the observations within a particular category should be represented by a real number selected from a closed interval of real numbers. In the former case the discrete nature of the process is reflected by the fact that we choose a single (discrete) number to represent all observations in the category, whereas in the latter case, the continuity of the process is reflected by the fact that we choose real numbers from a closed (continuous) interval of real numbers. Formally, we define the two restrictions as follows: the discrete restriction is

$$(5) \qquad t^d: \begin{cases} (x_{ai} \sim x_{bi}) \rightarrow (x^*_{ai} = x^*_{bi}) \\ (y_{aj} \sim y_{bj}) \rightarrow (y^*_{aj} = y^*_{bj}) \end{cases}$$

where $\sim$ indicates empirical equivalence (i.e., membership in the same category) and where the superscript on t^d indicates the discrete assumption. The continuous restriction is represented as

$$(6) \qquad t^c: \begin{cases} (x_{ai} \sim x_{bi}) \rightarrow \begin{cases} (x^-_{ai} = x^-_{bi} \leq x^*_{ai} \leq x^+_{ai} - x^+_{bi}) \\ (x^-_{ai} = x^-_{bi} \leq x^*_{bi} \leq x^+_{ai} = x^+_{bi}) \end{cases} \\ (y_{aj} \sim y_{bj}) \rightarrow \begin{cases} (y^-_{aj} = y^-_{bj} \leq y^*_{aj} \leq y^+_{aj} = y^+_{bj}) \\ (y^-_{aj} = y^-_{bj} \leq y^*_{bj} \leq y^+_{aj} = y^+_{bj}) \end{cases} \end{cases}$$

where x_{ai}^{-} and x_{ai}^{+} are the lower and upper bounds of an interval of real numbers. Note that one of the implications of empirical (categorical) equivalence is that the upper and lower boundaries of all observations in a particular category are the same for all the observations. Thus, the boundaries are more correctly thought of as applying to the categories rather than the observations, but to denote this would involve a somewhat more complicated notational system. Note also that for all observations in a particular category the corresponding rescaled observations are required to fall in the interval but need not be equal.

We now turn to the final set of restraints on the several measurement transformations t, the measurement level restraints. With these restraints we can modify the nature of the allowable transformations so that they correspond to the assumed level of measurement of the observation variables. There are, of course, a variety of different restraints which might be of interest, but we only mention three here. With these three, we can satisfy the characteristics of the most frequently mentioned measurement levels (nominal, ordinal and interval). We do not treat the ratio level of measurement since it is indistinguishable from the interval level for the regression model.

For nominal variables we introduce no level restraints since the characteristics of nominal variables are completely specified by the previously mentioned process restraints. We note that there is a special type of nominal variable which has only two categories, and which is referred to as a binary variable. We can think of this as a separate level of measurement, with there being both discrete-binary and continuous-binary variables. While the binary

measurement level is like any other nominal variable in that no additional constraints are present beyond those stated in Eqs. (5) and (6), there is a subtle difference in the nature of the discrete-binary case. In this case the optimally scaled variable consists of only two values, with their values being determined solely by the normalization conventions adopted in Eq. (4).

For ordinal variables, we require, in addition to the process restraints, that the real numbers assigned to observations in different categories represent the order of the empirical observations:

$$(7) \qquad t^{o}: \ (x_{ai} \lessdot x_{bi}) \rightarrow (x^{*}_{ai} \leq x^{*}_{bi})$$

where the superscript on t^{o} indicates the order restriction, and where $\lessdot$ indicates empirical order. Note that we require weak order, i.e., the assigned numbers are permitted to be equal even if the observations are not. The problem of what to do about ties has already been handled by our previous discussion of the process restrictions. If the variable is discrete-ordinal (t^{do}) then tied observations remain tied after transformation, whereas for continuous-ordinal (t^{co}) variables tied observations may be untied after transformation.

For quantitative (interval) variables, we require that the real numbers assigned to the observations be linearly related to the observations:

$$(8) \qquad t^{\ell}: \ x^{*}_{ai} = \delta_{o} + \delta_{1} x_{ai}.$$

More generally, we might require that the assigned numbers be related to the observations by a polynomial of known degree:

$$(9) \qquad t^p: \quad x^*_{ai} = \sum_{q=0}^{p} \delta_q x^q_{ai}.$$

Note that we still think of the observations as being categorical even if the measurement level is quantitative, although this is not very illuminating since each category will generally have only one observation (i.e., there are usually no ties). Thus the discrete-continuous distinction is usually only of academic interest with quantitative variables and will not be pursued further.

1.4 Model subspace

The final notion to be introduced is that of the model subspace, represented by the matrices $\tilde{\underline{X}}$ and $\tilde{\underline{Y}}$, which are of the same order as the observation matrices $\underline{X}$ and $\underline{Y}$. The model subspace notion is most easily introduced by expanding Eq. (3) to

$$(10) \qquad \lambda^2 = [\underline{x}^*_\ell \alpha_\ell - (\underline{b}-\underline{a}+\underline{x}^*_\ell \alpha_\ell)]'[\underline{x}^*_\ell \alpha_\ell - (\underline{b}-\underline{a}+\underline{x}^*_\ell \alpha_\ell)].$$

If we define

$$(11) \qquad \tilde{\underline{x}}_\ell \alpha_\ell = [\underline{b} - (\underline{a}-\underline{x}^*_\ell \alpha_\ell)]$$

then

$$(12) \qquad \lambda^2 = [\underline{x}^*_\ell \alpha_\ell - \tilde{\underline{x}}_\ell \alpha_\ell]'[\underline{x}^*_\ell \alpha_\ell - \tilde{\underline{x}}_\ell \alpha_\ell]$$
$$= \alpha_\ell^2 (\underline{x}^*_\ell - \tilde{\underline{x}}_\ell)'(\underline{x}^*_\ell - \tilde{\underline{x}}_\ell).$$

Note that the model subspace vector $\tilde{\underline{x}}_\ell$ is proportional to the difference between the two linear components $\underline{a}$ and $\underline{b}$ when the optimally scaled observations on variable ℓ are removed from the equation:

$$(13) \qquad \tilde{\underline{x}}_\ell = [\underline{b} - (\underline{a}-\underline{x}^*_\ell \alpha_\ell)]\alpha_\ell^{-1}.$$

thus explaining the name of these variables. Of course, we may define $\underline{y}_\ell$ correspondingly.

2.0 Algorithm

The MORALS/CORALS algorithm is an alternating least squares (ALS) algorithm. A very closely related ALS algorithm which is appropriate to the ANOVA situation (ADDALS) has been described in the companion paper (de Leeuw, Young, and Takane, 1975). An ALS algorithm for individual differences multidimensional scaling (ALSCAL) has also been discussed by Takane, Young, and de Leeuw (1975). These same investigators are also developing ALS algorithms for principal components analysis. All of these algorithms have in common the fact that the data being analyzed may be at a variety of measurement levels. The ALS approach is related to the work of Wold (1966), de Leeuw (1970; 1973) and Young (1972).

As is implied by the name, an ALS algorithm is an iterative algorithm which alternates back and forth between two phases, each of which is least squares. In one of the phases least squares estimates for the model parameters are obtained while the data transformations are held constant, whereas in the other phase, least squares estimates of the transformations are obtained while the model parameters are held constant. It has been shown by de Leeuw, Young and Takane (1975) that under the appropriate conditions such an iterative process is convergent. The MORALS/CORALS algorithm is such an algorithm. In one of the phases the least squares estimates of the model parameters $\underline{\alpha}$ and $\underline{\beta}$ are obtained while the transformations t are held constant, while in the other phase the least squares estimate of a single transformation t_i is obtained while $\underline{\alpha}$, $\underline{\beta}$, and the remaining t are held fixed. Since both phases minimize Eq. (3) the algorithm is convergent. In the next two sections we discuss the

model estimation phase (which estimates $\underline{\alpha}$ and $\underline{\beta}$) and then the optimal scaling phase (which estimates t_i).

We should point out that the model estimation phase is entirely superfluous. It is possible to recast our problem in the indicator matrix format used in our work with the additive model (de Leeuw, Young & Takane, 1975) and to avoid the transformations notation entirely. When viewed in this light it is apparent that the regression weights are unnecessary, being absorbed into the scaling of each variable. However, it is our opinion that the presentation mode used here is, pedagogically, more desirable in the present context.

2.1 Model Estimation Phase

In the model estimation phase we desire to obtain least squares estimates (relative to λ^2) of the model parameters $\underline{\alpha}$ and $\underline{\beta}$ under the assumption that all the optimally scaled variables $\underline{X}^*$ and $\underline{Y}^*$ are held constant. When it is recalled that the $\underline{X}^*$ and $\underline{Y}^*$ represent the observations rescaled at the interval level, then we see that this phase is no different than the classical canonical correlation situation (Hotelling, 1935) where the variables are the $\underline{x}_i^*$ and $\underline{y}_j^*$ (not the $\underline{x}_i$ and $\underline{y}_j$). If we define $\underline{R}_{xx}$ to be the correlations among the variables in $\underline{X}^*$, with R_{xy}, and R_{yy} defined correspondingly, then the least squares solution to the canonical correlation equation

$$\underline{X}^*\underline{\alpha} \tilde{=} \underline{Y}^*\underline{\beta} \tag{14}$$

may be obtained by finding the square root of the largest latent root μ of

$$(\underline{R}_{yy}^{-1}\underline{R}_{yx}\underline{R}_{xx}^{-1}\underline{R}_{xy} - \mu\underline{I})\underline{\beta} = \underline{0}. \tag{15}$$

It is then the case that the vector of weights $\underline{\beta}$ is the characteristic vector associate with the largest root of (15). The weight vector $\underline{\alpha}$ is obtained from

(16) $\underline{\alpha} = (\underline{R}_{xx}^{-1}\underline{R}_{xy}\underline{\beta})^{-1/2}$

In the multiple correlation situation where, for example, there is only one variable $\underline{x}^*$, the least squares estimates of

(17) $\underline{x}^* = \underline{Y}^*\underline{\beta}$

for _ reduces to

(18) $\underline{\beta} = (\underline{Y}^{*\prime}\underline{Y}^*)^{-1}\underline{Y}^{*\prime}\underline{x}^*$.

2.2 Optimal Scaling Phase

In the optimal scaling phase of the ALS process we desire to obtain least squares estimates (relative to λ^2) of the optimal scaling transformation t_i under the assumption that the remaining transformations $t_\ell (\ell \neq i)$ and the model parameters $\underline{\alpha}$ and $\underline{\beta}$ are all held constant. The actual procedure by which t_i is obtained depends on the specific measurement assumptions for each variable; that is, whether the transformation is of the type t^d, t^c, t^{do}, t^{co}, t^ℓ, or t^p. We discuss each of these in turn.

For the two nominal level transformations t^d (discrete-nominal) and t^c (continuous-nominal) the estimation process is very simple and, at least for t^d, quite well known. Fisher, in 1938 (pp 285-298), first discussed the t^d procedure, which was christened "optimal scaling" by Bock (1960) (a name we use to refer to all of the t estimation procedures). The t^d procedure consists, simply enough, of defining an element x^*_{ai} (or, implicitly, y^*_{aj}) as the mean of all the $\tilde{x}_{ai}$ which correspond to observations x_{ai} in a particular category. Thus, if there are, say, five observations x_{ai} in a particular category we define each of the five corresponding x^*_{ai} as the mean of the appropriate five subspace values $\tilde{x}_{ai}$. Since the x^*_{ai} are the mean of their corresponding $\tilde{x}_{ai}$, we minimize λ^2 (Eq. 12) given the restrictions placed by the measurement

characteristics on t^d (Eq. 5). Formally, x^*_{ai} is stated, under restrictions t^d, as

$$(19) \qquad t^d[\underline{x}_a, \tilde{\underline{x}}_a] \equiv \underline{X}^*_a = \underline{Z}(\underline{Z}'\underline{Z})^{-1}\underline{Z}'\tilde{\underline{X}}_a$$

where $Z = [z_{bci}]$; and where z_{bci} indicates category membership:

$$z_{bci} = \begin{cases} 1 & \text{iff } x_{ai} \in \text{category } c \\ 0 & \text{otherwise} \end{cases} .$$

(Note that the transformation t is most properly thought of as a function of both the observations vector $\underline{x}_i$ and the subspace vector $\tilde{\underline{x}}_i$. Thus, the definition of t given in equation 1 is a simplification made to ease the presentation).

The continuous-nominal situation is a bit more complex. The added complexity is introduced because the continuous-nominal situation, as discussed to this point, involves no measurement restrictions. For t^c (Eq. 6) we just have the requirement that each optimally scaled observation x^*_{ai} should reside in some interval, and we have placed no restrictions on the formation of the interval. Thus we could select arbitrarily large upper and lower boundaries x^+_{ai} and x^-_{ai} such that all elements of $\underline{x}^*$ are set equal to all elements of $\tilde{\underline{x}}$, thus minimizing Eq. (12) trivially and totally.

Naturally, the process proposed in the previous paragraph is meaningless. Therefore, we propose an alternative process which utilizes the nominal information in the observation vector $\underline{x}_i$ (to define the categories) and the ordinal information in the subspace vector $\tilde{\underline{x}}_i$ (to define the upper and lower boundaries). Since only the nominal characteristics of $\underline{x}_i$ are used we do not violate the measurement assumptions made about the data. However, since we use additional ordinal information from the model, we term this

transformation the pseudo-ordinal transformation. We still denote it as t^c, however, since it fully satisfies the continuous-nominal restrictions (Eq. 6) without using more than nominal information in the observation vector.

The pseudo-ordinal transformation t^c involves the following process: As the first step, we follow the procedure for the discrete-nominal transformation t^d to form the mean of the subspace elements for all observations in a given category. We then order the categories according to these means, noting that we have not violated the assumed nominal level of $\underline{x}_i$ since we are basing the order on the model subspace $\underline{\tilde{x}}_i$. As the third step, we order all the observations in a given category according to each observations' subspace value $\underline{\tilde{x}}_{ai}$ noting that the subspace values for the observation category with the smallest mean will tend to be smaller than the subspace values of the observation category with the next smallest mean, etc. At this point, we have determined an order for each of the observations on the basis of the model subspace. This order dictates that all those observations which are in the category with the smallest mean should be optimally scaled with x^*_{ai} which are smaller (or, at least, not larger) than all the x^*_{ai} for the next larger category, etc. Furthermore, the order dictates that all the optimally scaled x^*_{ai} <u>within</u> a given category ought to be in the same order (at least weakly) as their corresponding subspace value $\tilde{x}_{ai}$. Of course, if we simply defined the optimally scaled x^*_{ai} to be equal to their corresponding $\tilde{x}_{ai}$ we would probably have order violations since it is likely that the largest $\tilde{x}_{ai}$ in the smallest category are larger than the smallest $\tilde{x}_{ai}$ in the next larger category. Thus, as the final step in defining t^c we apply Kruskal's

least squares monotonic transformation to the subspace $\tilde{x}_i$ and the observation order defined from the subspace in the manner just described.

A brief example will help clarify the procedure. Say that we have observation and subspace vectors as follows:

$$\underline{x}_i' = [M,B,B,Q,M,Q,B]$$

$$\underline{\tilde{x}}_i' = [1,6,4,3,3,7,2]$$

Then the vector of category means is

$$\underline{c}_i' = [2,4,4,5,2,5,4]$$

which would be $\underline{x}^*$ under the discrete-nominal transformation t^d. However, for the continuous-nominal transformation t^c we sort this vector into order ($\underline{s}_i$) and rearrange the subspace accordingly ($\underline{d}_i$):

$$\underline{s}_i' = [2,2,4,4,4,5,5]$$

$$\underline{d}_i' = [1,3,6,4,2,3,7].$$

As the third step, we sort the elements of $\underline{d}_i$ <u>within</u> category blocks (to get $\underline{e}_i$):

$$\underline{e}_i' = [1,3,2,4,6,3,7].$$

The final step is to apply Kruskal's least squares monotonic transformation to get

$$\underline{f}_i' = [1,2\tfrac{1}{2},2\tfrac{1}{2},4,4\tfrac{1}{2},4\tfrac{1}{2},7],$$

which, when rearranged to correspond with the observations, gives

$$\underline{x}_i^{*\prime} = [1,4\tfrac{1}{2},4,4\tfrac{1}{2},2\tfrac{1}{2},7,2\tfrac{1}{2}],$$

with the vectors $\underline{x}_i^+$ and $\underline{x}_i^-$ being defined (implicitly) as

$$\underline{x}_i^{-\prime} = [1,2\tfrac{1}{2},2\tfrac{1}{2},4\tfrac{1}{2},1,4\tfrac{1}{2},2\tfrac{1}{2}]$$

and

$$\underline{x}_i^{+\prime} = [2\tfrac{1}{2},4\tfrac{1}{2},4\tfrac{1}{2},7,2\tfrac{1}{2},7,4\tfrac{1}{2}].$$

(Note that an equivalent procedure would be to apply Kruskal's primary least squares monotonic transformation to the vector $\underline{d}_i$, which, as its first step, transforms $\underline{d}_i$ into $\underline{e}_i$. Also, note that were we to apply Kruskal's secondary monotonic least squares transformation to $\underline{d}_i$ we would obtain the $\underline{x}_i^*$ obtained under the discrete-nominal transformation.) It can be shown that

$$(20) \qquad t^c[\underline{x}_a, \tilde{\underline{x}}_a] \equiv \underline{X}_a^* = \underline{Z}(\underline{Z}'\underline{Z})^{-1}\underline{Z}'\tilde{\underline{X}}_a$$

where $\underline{Z}$ is a dummy binary matrix indicating elements of $\tilde{\underline{X}}_a$ which must be tied to satisfy the ordinal properties developed as stated above. The matrix $\underline{Z}$ is defined by Kruskal's least square monotonic transformation procedure (Kruskal, 1964; Ayer, et.al., 1955).

The two ordinal level transformations t^{do} (discrete-ordinal) and t^{co} (continuous-ordinal) are defined by Kruskal's least squares monotonic transformations where our discrete process corresponds to his secondary procedure, and our continuous process to his primary procedure. Young (1975) has shown that the discrete-ordinal transformation may be formally stated as

$$(21) \qquad t^{do}[\underline{x}_a, \tilde{\underline{x}}_a] \equiv \underline{X}_a^* = \underline{Z}(\underline{Z}'\underline{Z})^{-1}\underline{Z}'\tilde{\underline{X}}_a$$

where $\underline{Z}$ is a binary matrix of dummy variables indicating the $\tilde{\underline{X}}_a$ which must be tied to satisfy both categorical equivalence and the ordinal restrictions. The continuous-ordinal transformation is formally

$$(22) \qquad t^{co}[\underline{x}_a, \tilde{\underline{x}}_a] \equiv \underline{X}_a^* = \underline{Z}(\underline{Z}'\underline{Z})^{-1}\underline{Z}'\tilde{\underline{X}}_a.$$

where $\underline{Z}$ is a binary matrix of dummy variables indicating the $\tilde{\underline{X}}_a$ which must be tied to satisfy the ordinal restrictions. Kruskal (1964) has shown that t^{do} is least square relative to λ given the constraints specified by Eq. (5) and (7), and de Leeuw (1975) has shown that t^{co} is least squares under constraints (6) and (7).

The least squares solution for $\underline{x}_i^*$ (or $\underline{y}_j^*$) under the restrictions of the t^ℓ and t^p transformations are well known and are summarized here. The t^p transformation (equation 9) can be written, in matrix notation, as

$$(23) \qquad t^p[\underline{x}_i, \tilde{\underline{x}}_i] \equiv \underline{x}_i^* = \underline{X}^{pi}\underline{\delta}$$

where $\underline{X}^{pi}$ is a matrix with k columns and p+1 rows, each row being an integer power of the vector $\underline{x}_i$ of observations. The first row is the zero'th power (i.e., all ones), the second row is the first power (i.e., is $\underline{x}_i$ itself), the third row is the squares $\underline{x}_i^2$, etc. (for t^ℓ p=1). The least squares estimate of δ is

$$(24) \qquad \tilde{\delta} = (\underline{X}^{pi'}\underline{X}^{pi})^{-1}\underline{X}^{pi}\tilde{\underline{x}}_i$$

Note that all of the transformations have been stated in a form involving linear combinations of the model subspace vector $\underline{x}_i$ (See Eqs. 19 through 23), thus allowing each transformation to be represented as a projection operator. This point has been discussed fully by Takane, Young & de Leeuw (1975).

2.3 Initial values

As with any iterative algorithm we must supply the procedure with values to initiate the process, but with MORALS, unlike some algorithms, there is a particularly compelling initialization procedure. We simply assume that the matrices $\underline{X}$ and $\underline{Y}$ (the raw data) are actually the matrices $\underline{X}^*$ and $\underline{Y}^*$. This is equivalent to assuming, for the initialization process, that the raw data are measured on an interval scale. (Due to their nature, we must assign arbitrary values to the observation categories when a variable is nominal.) We then enter the model estimation phase and solve for $\underline{\alpha}$ and $\underline{\beta}$ by equations (15) and (16) (or equation 18, as the case may be). We

then use these estimates of the model parameters to begin the iterative procedure whose flow is discussed in the next section. Note that the initialization procedure is simply the classical canonical or multiple regression procedure. Thus, the very first step of the procedure corresponds to the analysis the researcher would have obtained had he decided that his variables were all quantitatively measured. By comparing the results of the entire iterative process with the results of the very first step of the process the user can determine what he has gained by the use of his measurement assumptions (except for nominal variables). If, for example, he assumed that his variables were ordinal and discovers that the canonical correlation is essentially the same after the iterative process as it was before, and he also discovers that his monotonic transformations are all essentially linear, then he could safely conclude that he has gained nothing by the assumptions of ordinality, and that it would have been proper to assume that his variables are measured at the interval level. Thus, we see that by using the classical procedure as the initialization process the user can easily investigate whether his assumptions concerning measurement levels are correct. More will be said on this important point later.

2.4 Algorithm Flow

As described thus far, the MORALS/CORALS procedure consists of two phases, one to estimate the model parameters $\underline{\alpha}$ and $\underline{\beta}$, and the other to estimate the data transformation t_i. The obvious ALS procedure would involve alternating these two phases until convergence is obtained. If you have followed the discussion closely, however, you will have detected a subtle imbalance between the two phases. Whereas the model estimation phase obtains estimates of the

weights $\underline{\alpha}$ and $\underline{\beta}$ for all variables, the optimal scaling phase obtains estimates of the transformation one variable at a time. Thus, the obvious ALS procedure must be modified somewhat. Several of the alternatives that we might choose are discussed in the remainder of this section.

One alternative is an iterative procedure where a single iteration is defined as follows: First select a particular variable and optimize λ^2 relative to that variable (as expressed by equation 13) by repetitively alternating the two phases until convergence is obtained. Then repeat this procedure for another variable and another, etc., until all variables have been subjected to this procedure. This then completes a single iteration, with the iterations being repeated until convergence is obtained. We have not investigated this procedure as it seems to us to be computationally inefficient relative to some of the procedures discussed below.

Another alternative would define an iteration as follows: Obtain the transformation t which yields the optimal $\underline{x}_1^*$, replace the previous $\underline{x}_1^*$ with the new one, repeat this process for $\underline{x}_2^*$, then $\underline{x}_3^*$, etc., until all the variables in both sets have been optimally scaled. There is now an entirely new set of $\underline{X}^*$ and $\underline{Y}^*$, and the transformation process is repeated again for this new set of $\underline{X}^*$ and $\underline{Y}^*$. We repeat the transformation estimation until convergence is obtained at which point we have already obtained estimates of $\underline{\alpha}$ and $\underline{\beta}$ as a by-product of the procedure. We have investigated this algorithm and have found it to be very slow. Thus, we do not pursue it further, although it can be shown to converge on the desired $\underline{X}^*$ and $\underline{Y}^*$ and on the appropriate weights $\underline{\alpha}$ and $\underline{\beta}$.

The two algorithms that we have investigated most thoroughly are each rather similar to the one just discussed. The iterative structure for the first of these two procedures is as follows: First obtain the transformation t which yields the optimal $\underline{x}_1^*$, replace the previous $\underline{x}_1^*$ with the new one, then repeat this process for each of the other $\underline{x}_i^*$ (i=1,2,3...n). Now solve for weights $\underline{\alpha}_i$ and $\underline{\beta}_j$. Follow this with estimation and replacement for each $\underline{y}_j^*$ (j=1,2,...m) and another estimation of $\underline{\alpha}_i$ and $\underline{\beta}_j$. The other algorithm that we have investigated is exactly the same except that the weights $\underline{\alpha}$ and $\underline{\beta}$ are only estimated once on each iteration. Thus, the structure of an iteration for this procedure is to estimate and replace each $\underline{x}_i^*$ (i=1,2,...n), and then each $\underline{y}_j^*$ (j=1,2,...m), and then to estimate the $\underline{\alpha}$ and $\underline{\beta}$. Both of these algorithms entail an iteration which is much quicker than the first possibility discussed in the previous paragraphs and slightly slower than the second possibility. Many fewer iterations are required than with the second alternative. More important, however, is the fact that these two algorithms place more nearly equal emphasis on the two phases than the former two algorithms. We have been unable to determine any characteristics of the latter two algorithms which would allow us to select one over the other except for the fact that the one involving only one model estimation is slightly more efficient. Thus, we have chosen it to define the iterative flow of MORALS/CORALS.

2.5 Missing data

Missing data are allowed for in a manner which does not destroy the ALS property of the MORALS/CORALS algorithm. If some observation x_{ai} (or, implicitly, y_{aj}) is missing, then the computation

of the initial weights $\underline{\alpha}_i$ and $\underline{\beta}_j$ is changed in a minor manner: We simply estimate each missing x^*_{ai} as being the mean of the non-missing observations in the fector $\underline{x}_i$. The computation of initial $\underline{\alpha}_i$ and $\underline{\beta}_j$ then proceeds as stated.

The computation of the subspace vector $\underline{\tilde{x}}_i$ is as stated above except for the element $\tilde{x}_{ai}$ corresponding to the missing observation. The value of this subspace element is determined by either a) setting it equal to the mean of $\underline{x}^*_i$; or b) setting it equal to the element x^*_{ai}. The choice of method is left to the user and corresponds to whether he views his missing observations as having been caused by a discrete or continuous process, respectively. If the discrete assumption seems appropriate then all missing observations on a given variable are assumed to have been caused by the same (discrete) set of events, and are all assigned the mean of the non-missing values (the least squares discrete estimate). If the continuous assumption is appropriate then all missing observations on a given variable are assumed to have been caused by different (continuous) sets of events and are all set equal to their corresponding optimally scaled observation (the least squares continuous estimate). Alternatively, we may view the choice as concerning whether all missing observations represent a single category (form means) or separate categories (don't form means).

2.6 Partitions

It may sometimes be the case that not all of the observations made on a single variable are comparable. For example, it may be that some of the observations of a variable were made at a different time or under somewhat different conditions than some of the other observations with the result being that a specific observation

made at one time (or place) cannot be said to be larger or smaller than one made at the other time (or place) even though the observations have the same measurement characteristics. Or, as another example, the measuring device may have broken during the observation process and been repaired in such a way that the measurements before and after the break-down are no longer directly comparable. If it is the case that not all observations on a specific variable are directly comparable, then we wish to partition the observations into mutually exclusive and exhaustive subsets, and permit separate transformations within each partition.

The partition notion is also useful in precisely the opposite situation. It may be that we have two (or more) variables which are measured on the same scale and that we wish to obtain identical optimal transformations of both variables. In this case we wish to collect the observations on the two variables into a single superset of observations and obtain a single transformation for the superset.

It should be apparent that defining partitions as either subsets of the observations on a variable, or as supersets of observations creates no particular problem. We need only to substitute the desired partitioning of $\underline{X}$ and $\underline{Y}$ for the partitioning which has been implicit in our previous developments. (i.e., partitioning by variables.

2.7 Step size

Since the MORALS/CORALS algorithm uses algebraic expressions which permit explicit least squares parameter estimation, there is no step to estimate. Therefore, there is no step size problem.

2.8 Convergence

A formal proof of the convergence of an ALS process is available in de Leeuw, Young and Takane (1975). An intuitive grasp of the proof can be obtained from the following simple argument: In either of the phases we obtain the best possible estimates for the relevant subset of parameters, given that the remaining parameters are constant. Each phase optimizes λ^2 by manipulating parameters which are completely disjoint from those involved in the other phase. Furthermore, each phase must necessarily reduce λ^2 by some amount (or, at the very least, λ^2 can <u>never</u> increase) and λ^2 can never be negative. Therefore, it follows that some value of λ^2 must be found such that the amount of further decrease in λ^2 is of negligible importance. It is in this sense that an ALS algorithm is convergent, with the critical notion being λ^2 can never increase. While it is true that monotone decreasing bounded sequences (the type of sequence just discussed) converge, they do not necessarily converge on a stable point unless the conditional minima are unique and λ^2 is continuous. While these conditions may not be satisfied in our situation, we have never observed instabilities. It might be pointed out at this point that while λ^2 can never increase, the sum of squared deviations between a given optimally scaled variable $\underline{x}_\ell^*$ and its subspace variable $\tilde{\underline{x}}_\ell$ may occasionally increase. That is, while

$$\lambda^2 = \alpha_\ell^2(\underline{x}_\ell^* - \tilde{\underline{x}}_\ell)'(\underline{x}_\ell^* - \tilde{\underline{x}}_\ell)$$

cannot increase, the stress (Kruskal, 1964) of a variable

$$s_\ell^2 = (\underline{x}_\ell^* - \tilde{\underline{x}}_\ell)'(\underline{x}_\ell^* - \tilde{\underline{x}}_\ell)$$

may increase (s_ℓ^2 is normalized via equation 4). The increase in s_ℓ^2

will be more than compensated by the decrease in α_ℓ^2, thus permitting λ^2 to decrease. This emphasizes the fact that MORALS/CORALS optimizes the overall canonical correlation but not the fit for each variable, just as in the classical regression situation. Finally, we note that technically we optimize λ and not λ^2 since $\alpha'=-\alpha$ and $\beta'=-\beta$ also optimize λ^2 but not λ. Thus the optimal λ is uniquely defined, whereas the optimal λ^2 is not.

2.9 Local minimum

It might appear as though an ALS algorithm would have no problems with local minimum since it has no step size to estimate and since it is convergent. Unfortunately, this is not entirely true. It is the case, however, that an ALS algorithm has fewer sources of local minimum problems than a gradient procedure (since once the process is initiated there is no step size estimation and the process is convergent). It is the case, however, that the point of convergence can be influenced by the initialization process with, perhaps, a superior value of λ^2 obtainable from one set of initial values than from another set. While in practice we have never seen this occur we have been unable to show that it cannot occur. However, we are on firm grounds, both logically and empirically, when we assert that a convergent procedure with no step size to estimate will be less prone to local optimum problems than a divergent procedure with a step size. Thus, since we use a rational initialization procedure, it is not inaccurate to describe the solution obtained by MORALS as having conditional global optimum features, with the globalness of the optimum being conditional only on the initialization procedure.

2.10 Termination

The termination rule is simple: If the improvement in λ^2 is less than some desired amount, or if a prespecified number of iterations has been exceeded, we stop.

3.0 Evaluation

In this section we present three evaluations of the MORALS/CORALS algorithm. The first evaluation involves a Monte Carlo study in which we find that the algorithm can recover known information in the face of systematic and random error. In the second evaluation we obtain useful and interpretable results when the algorithm is applied to a set of real survey data. In the third and final evaluation we obtain a meaningful interpretation of a multidimensional scaling solution in a special case of MORALS/CORALS which corresponds to a previously proposed procedure for interpreting such solutions.

3.1 Monte Carlo study

A small Monte Carlo study designed to investigate the robustness of MORALS/CORALS in the face of both systematic and random error is presented in this section. We do not claim that this is a complete or definitive study of the algorithm's behavior in situations likely to be encountered by the typical investigator. It is not. Rather, it is simply a small study designed to demonstrate the algorithm's behavior in one common situation, that in which there is a single dependent variable and two dependent variables, all three of which are ordinal. The design of the study is as follows.

Two "true" independent variables were generated, each with 64 observations, each observation being sampled from a random uniform

distribution on the interval (0-1). A "true" dependent variable was then generated by simply adding together corresponding elements of the independent variables.

Five degrees of nonlinear monotonic distortion, and three levels of random error were then defined. The nonlinear distortion involved the three transformations:

$$y_i = \bar{y}_i^k$$

$$x_{i1} = \bar{x}_{i1}^{1/k}$$

$$x_{i2} = (\bar{x}_{i2} - \bar{x}_{.2})^k \cdot \mathrm{SIGN}(\bar{x}_{i2} - \bar{x}_{.2})$$

where k varies from 1 to 5, where a "bar" indicates "true" values, and where $\bar{x}_{.1}$ indicates the mean of the true values. Thus when k=1 all variables remain undistorted, and when k varies up to 5 the variables become increasingly more nonlinear, with each variable being distorted in its own unique manner. Note, however, that all distorted variables are equivalent to the undistorted variables at the ordinal level, since each transformation is strictly monotonic. Random error involved adding a random normal deviate to each element of each of the three distorted variables, where the standard deviation of the deviate is proportional to the standard deviation of the undistorted variable. The proportionality constant defined the three levels of random error, and was equal to zero for the lowest level (no random error), .10 for the intermediate level, and .50 for the highest level. These systematically distorted and randomly perturbed variables were then input to MORALS/CORALS, under the assumption that each variable was ordinal. The results are summarized in Table 1 and Figure 1. The table presents the classical multiple correlation coefficient R_c (which is the one

Table 1

Monte Carlo Study

Random Error Level		Systematic Distortion Level				
		1	2	3	4	5
.00	R_c	1.0000	.9213	.8081	.7079	.6255
	R_o	1.0000	1.0000	1.0000	.9999	.9999
	r_y	1.0000	.9917	.9861	.9833	.9831
	r_{x1}	1.0000	.9869	.9792	.9761	.9763
	r_{x2}	1.0000	.9943	.9918	.9904	.9898
	iter	1	5	7	8	9
.10	R_c	.9882	.9069	.7832	.6939	.5921
	R_o	.9995	.9999	.9994	.9996	.9998
	r_y	.9942	.9867	.9810	.9517	.9869
	r_{x1}	.9931	.9820	.9796	.9139	.9800
	r_{x2}	.9906	.9851	.9814	.9748	.9864
	iter	5	6	13	11	12
.50	R_c	.8486	.7043	.5944	.6498	.5228
	R_o	.9544	.9269	.9998	.9545	.9997
	r_y	.9177	.8753	.3339	.9214	.2721
	r_{x1}	.8801	.8709	.8927	.8805	.8126
	r_{x2}	.8849	.9058	.2785	.9099	.2380
	iter	6	5	21	6	23

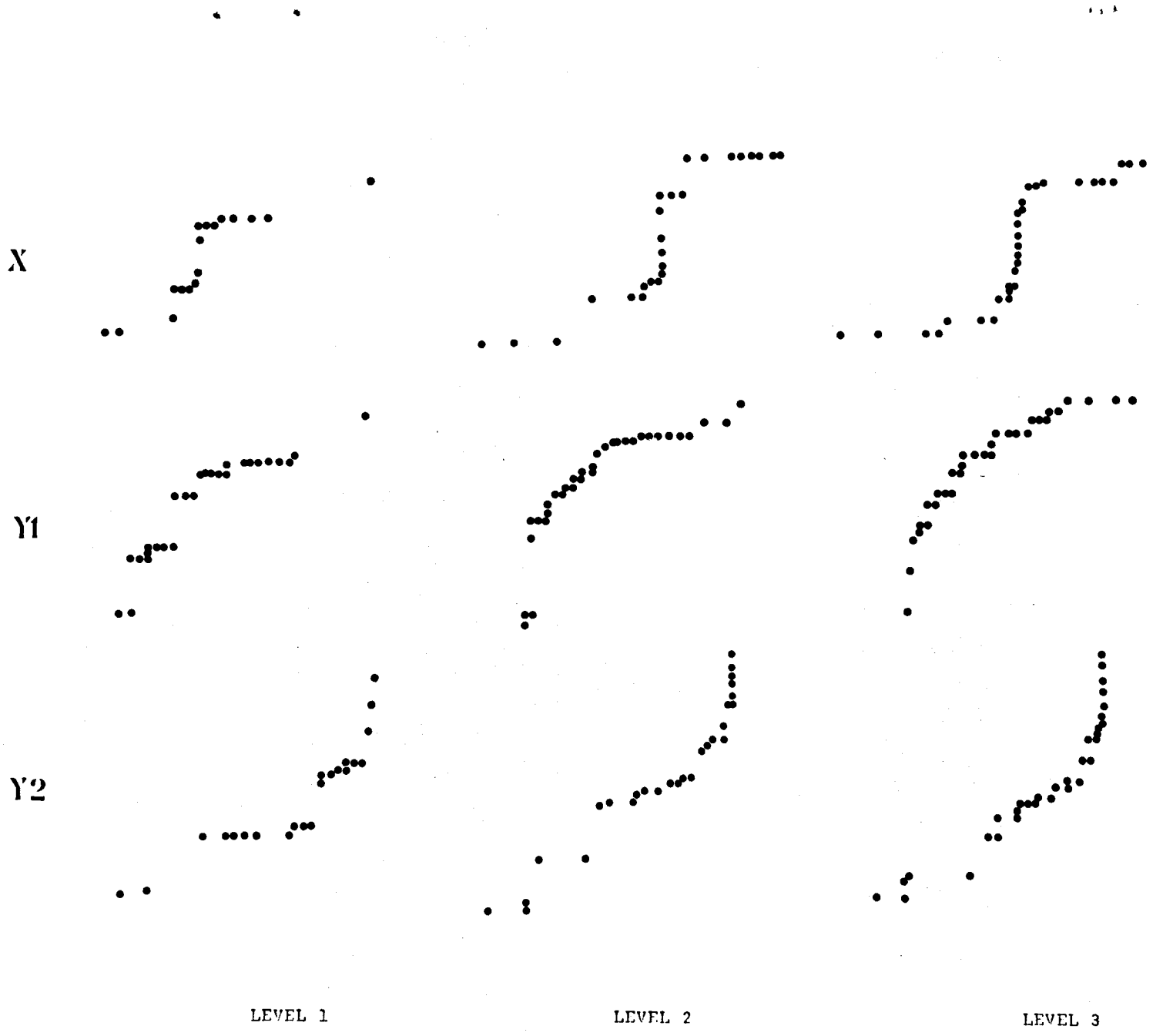

Figure 1

Recovered transformations for three levels of random error
at the fourth level of systematic distortion.

obtained under the interval measurement level assumptions made in the standard multiple correlation analysis, and which is also the initial correlation obtained by MORALS/CORALS), the optimal multiple correlation coefficient R_o (obtained under ordinal measurement level assumptions), the simple correlation of each optimally scaled variable with its corresponding "true" variable (r_y, r_{x1}, and r_{x2}), and the number of iterations required for convergence of the algorithm. The figure presents the optimal transformations obtained for each variable for the fourth degree of systematic distortion and for all three levels of random perturbation. The horizontal axes are the distorted and perturbed values being analyzed, and the vertical axes are the optimally transformed values. Ideally, the optimal transformations presented in Figure 1 should be the inverse of the transformations used to define the systematic distortion.

We observe that for all levels of systematic distortion the algorithm can perfectly recover the underlying true structure when there is no random error (all optimal multiple correlations are 1.0 and all simple correlations are at least .976). This implies that the algorithm is able to obtain the desired inverse of each transformation defined above, otherwise the simple correlations would be lower. Note also that in these (no random error) cases the classical

multiple correlation progressively decreases from 1.0 in the no-distortion case to .626 in the case of most distortion, as is expected.

We also observe that when random error is present the degree to which the true information is recovered deteriorates, as is expected. For the intermediate level of random error the amount of deterioration does not seem to be affected much by the amount of systematic distortion, but there does seem to be some effect of systematic error for the greatest degree of random error, with the overall poorest recovery occurring when there is a large amount of random error, and severe non-linearities in the data. In particular we note that for high random error and for systematic error levels 3 and 5 (but not 4) the solution obtained by MORALS/CORALS is degenerate. The optimal transformations obtained for the dependent variable and for the second independent variable consist, essentially, of two values, with the largest value on each vector being transformed into a very large value, and all the remaining values being tied together at a very small value. The four corresponding simple correlations are much lower than any others. It is not clear to us why such a degeneracy should occur for these two particular levels of systematic distortion and not for the intervening level (see Figure 1). However, these results should serve as a warning that in cases of extreme nonlinearities and high amounts of error

the MORALS/CORALS algorithm may yield degenerate results. In our experience with real data, however, we have never obtained such degeneracies, suggesting that these extreme Monte Carlo conditions may be uncommon empirically.

3.3 Survey data

The application process for prospective graduate students to a quantitative psychology program can be summarized, from the faculty's point of view, by the variables presented in Table 2. Of the 12 variables presented in this table, 8 are obtained from the 33 graduate student applications and 4 are derived by the faculty, thus the 8 applicant variables may be naturally thought of as independent variables, and the 4 evaluation variables as dependent variables. The eight independent variables concern the degree obtained by the applicant (bachelor's or master's), the applicant's major as an undergraduate (either psychology, mathematics including statistics, or a double math-psych major), the applicant's grade point average, his verbal and mathematical score on the Graduate Record Examination, his score on the Millers Analogies examination, his strongest interest (quantitative psychology or statistics), and his average recommendation rating. The four dependent variables indicate the tentative accept-reject decision made by the faculty at the beginning of the evaluation (including a "defer" decision), the corresponding decision made at the

Table 2

Correlations and regression weights for survey data

	Canonical		Multiple			
	1	2	1	2	3	4
Regression Weights						
Dependent Variables						
Tentative decision	.212					
Final decision	.091					
Suitability order	.970	.996	1.000		1.000	
Applicant action	-.081	-.086		1.000		1.000
Independent variables						
Degree	.137	.204	.214	-.065		
Major	-.136	-.184	-.166	.192		
GPA	.192	.276	.260	-.173		
GRE verbal	-.204	-.295	-.272	1.296		
GRE math	-.281	-.409	-.410	-.957	-.745	.420
Millers Analogies	-.164	-.235	-.232	.187		
Interest	.108	.126	.065	-.348		
Recommendation	-.443	-.636	-.574	.238	-.528	.296
Correlations						
Classical	.925	.923	.921	.593	.842	.297
Optimal	1.000	1.000	1.000	.961	.997	.543
Number of cases removed	7	7	7	7	2	2
Number of iterations	1	2	2	13	15	2

conclusion of the evaluation, the faculty's rank order of all applicants' suitability to the graduate program, and the action taken by the applicants. Note that three of the four dependent variables are weakly equivalent at the ordinal level: The 13 students who ranked most suitable were accepted, and of these the 9 highest were tentatively accepted and the remaining 4 were tentatively deferred. Note also that 8 of those students who were initially deferred were later rejected, and that all of these students were ranked lower than those who were finally accepted.

We performed two canonical regressions on these data, one using all four dependent variables, and one using only suitability and action (due to the relationship between suitability and the other two dependent variables noted above). For each analysis we assumed that all variables were discrete (except suitability), and that the degree, major and interest variables were nominal, with the remaining ordinal. For each analysis we removed all of those students who had missing data (the analyses were also performed with these students left in and with their missing data estimated as stated above, with little difference in results).

The results of each analysis are presented in Table 2. For both analyses the optimal canonical correlation was 1.0 (in either 1 or 2 iterations). Thus it is possible to transform the dependent

and independent variables in a manner which allows a linear combination of one set to be perfectly linearly related to a linear combination of the other set. The pattern of weights is very informative. Among the dependent variables suitability is very heavily weighted, and the remaining receive almost no weight. This is true for both analyses and suggests that a very strong multiple correlation based only on suitability exists (a suggestion which will prove to be correct). The pattern of dependent variable weights is easily interpreted. We should expect the suitability weight to be very high because the suitability rank order is developed by the faculty from the applications summarized by the independent variables. So, if the independent variables summarize all of the information in the application, and if the rank order is consistent with that information, then the regression weight should be very high. The weights on the tentative and final decision variables should be low because their information is already contained in the suitability rank order, as noted above. Finally, the action variable should have a low weight because it reflects many variables not included in the analysis, such as whether or not the applicant was accepted by other graduate programs, how strongly he desired to be in each program, etc. Turning now to the independent variable weights we see that the pattern of weights is the same for the two analyses, with the GRE-math and recommendation variables receiving the heaviest weight. It is reassuring that the pattern of weights is stable between the two analyses.

Due to the perfect relationship derived in both canonical analyses, we decided to perform several multiple regression analyses, using either suitability or action as the dependent variable. The results

of these analyses are also presented in Table 2. (The same measurement assumptions and initial category values were used in these analyses as in the canonical analyses).

We first note that the optimal multiple correlation is still perfect for the analysis using suitability as the dependent variable, but somewhat lower for the action analysis, a result which is consonant with the dependent variable weights in the canonical analyses above. It is also interesting to note that the classical multiple correlation is initially quite strong for suitability but rather weak for action. We next note that the regression weights for the suitability analysis display the same pattern as the weights in the canonical analyses but that the weight pattern for the action analysis has changed. This should be anticipated due to the very large weight for suitability and low weight for action in the canonical analyses.

Finally, we performed one more pair of multiple correlation analyses, each involving only the two independent variables receiving the most weight in the three analyses in which suitability was a dependent variable. This new pair of analyses differed according to the dependent variable, it being either suitability or action. The results are also presented in Table 2, and are as anticipated. The optimal multiple correlation for the suitability analysis is no longer perfect, but is nearly so (.997), and the optimal multiple correlation for the action analysis is very low (.5427). Furthermore, the regression weights for the suitability analysis are essentially as in the previous analyses with this dependent variable.

The final aspect of these analyses to be emphasized is the set of optimal scale values assigned to each nominal variable (see Table 3). (We do not present the scale values for binary variables since they are arbitrarily set to one and zero). The first point to be made is that the scale values are rather stable for all analyses except the multiple correlation analysis of action. This result is in line with previous results. We next note that these variables do not receive very heavy regression weights in any of the analyses, so any interpretation of the variables by themselves does not add much to our understanding of the total process. However, for the three analyses in which suitability was a dependent variable, major received a negative weight, indicating that larger (optimally scaled) major values are associated with smaller (optimally scaled) suitability values. When we note that smaller suitability values indicate greater suitability (the highest rank is 1), then we see that applicants with mathematics backgrounds tend to be rated as most suitable, and those with psychology backgrounds as least suitable, with those having double majors being in between. In a corresponding fashion we can interpret the optimal scale values for interest (which has a positive regression weight in the three analyses in which suitability is involved) as showing that applicants who are most interested in quantitative psychology tend to be judged as most suitable, cognitive psychology as next most suitable, and statistics as least suitable. Finally, the positive regression weight for the binary degree variable, and the coding of this variable (1 for Bachelors, 2 for Masters) indicates that applicants with Bachelors degrees are more suitable.

Table 3

Optimal Scale Values

Variable	Category	Canonical 1	Canonical 2	Multiple 1	Multiple 2	Initial Values
Major	Psychology	1.023	1.016	1.008	1.333	1
	Math/Psych	1.803	1.856	1.921	2.133	2
	Mathematics	3.169	3.129	3.075	-0.554	3
Interest	Quantitative	1.007	.996	.998	.892	1
	Cognitive	1.984	2.010	2.005	2.531	2
	Statistics	3.012	2.992	2.996	2.400	3

Thus, from all of these analyses taken together, we conclude that it is possible to describe the faculty's judgments of the suitability of a prospective graduate student on the basis of only that student's score on the quantitative section of the Graduate Record Examination, and on his strength of recommendation. In addition we conclude that it is not possible to adequately describe the characteristics of those applicants who are to become members of the graduate training program on the basis of the variables summarizing their applications. Furthermore, the faculty tend to prefer students interested in quantitative psychology over those interested in cognitive psychology (and least like those interested in statistics), tend to judge those with a mathematical background as more suitable than those without, and tend to prefer students without masters degrees.

3.3 MDS interpretation

One of the special uses of the MORALS/CORALS approach is to interpret multidimensional scaling (MDS) configurations by projecting external information into the MDS space. Cliff and Young (1968) suggested that multiple correlation techniques could be used to project information into the MDS space in the form of a vector in the space which correlated most highly with the external information. Carroll (1972) suggested that multiple correlation techniques could also project information into an MDS space in the form of a point such that the distances between the projected point and the MDS points correlated most highly with the external information. Both of these suggestions assumed that the external variable is measured at the interval level of measurement, an assumption which is usually untenable, but generally ignored. Carroll and Chang, (1970)

however, have recently extended their proposal to cover variables defined at the ordinal level by a procedure precisely equivalent to the special case of MORALS to be discussed [personal communication].

Hoadley (1975) has obtained a MDS solution for 100 U.S. Senators based on their voting records in the Senate. He also obtained a variety of additional information including ratings of each Senator by liberal, conservative and special interest groups, proportion of votes which supported the Senator's party position on various issues, specific votes on certain key issues, etc. He also computed Guttman scales on several different topics of national concern.

MORALS/CORALS was used with these data in the following manner. Several multiple regressions were performed, all of which assumed that the two dimensions of the MDS space were defined at the interval level of measurement, and that the vector of information being regressed into the MDS space was qualitative. The results of these analyses are presented in Table 4, along with the measurement characteristics of each dependent variable. A geometric representation of one of these analyses is presented in Figure 2. This is a representation of both the MDS space and the regression of the Guttman scale concerning the degree to which each Senator voted in support of agricultural issues. The vector through the space is the direction determined by MORALS/CORALS such that when the points in the space are projected onto the vector, the resulting projections are as strongly correlated with the optimally scaled units as possible. Note that we have presented the optimally scaled units along one side of the vector, and the raw Guttman scale units

Table 4

Correlations and regression weights for MDS interpretation

Variable	Measurement Level	Measurement Process	Correlation Initial	Correlation Final	Weights Initial dimension 1	Weights Initial dimension 2	Weights Final dimension 1	Weights Final dimension 2	# of Iterations
War Powers	binary	d	.4819	.6003	-.476	.086	-.599	.051	1
Agricultural Subs	binary	d	.4426	.5922	-.341	.291	-.456	.389	1
Party	binary	c	.9121	1.0000	.549	.716	.549	.716	(0)
G Agri. Scale	ordinal	d	.7900	.8416	.708	-.367	.779	-.338	2
G Civil Liberty	ordinal	d	.9255	.9333	-.764	+.541	-.765	.552	1
G Gvt. Reg. Business	ordinal	d	.8812	.8918	-.879	-.049	-.889	-.056	1

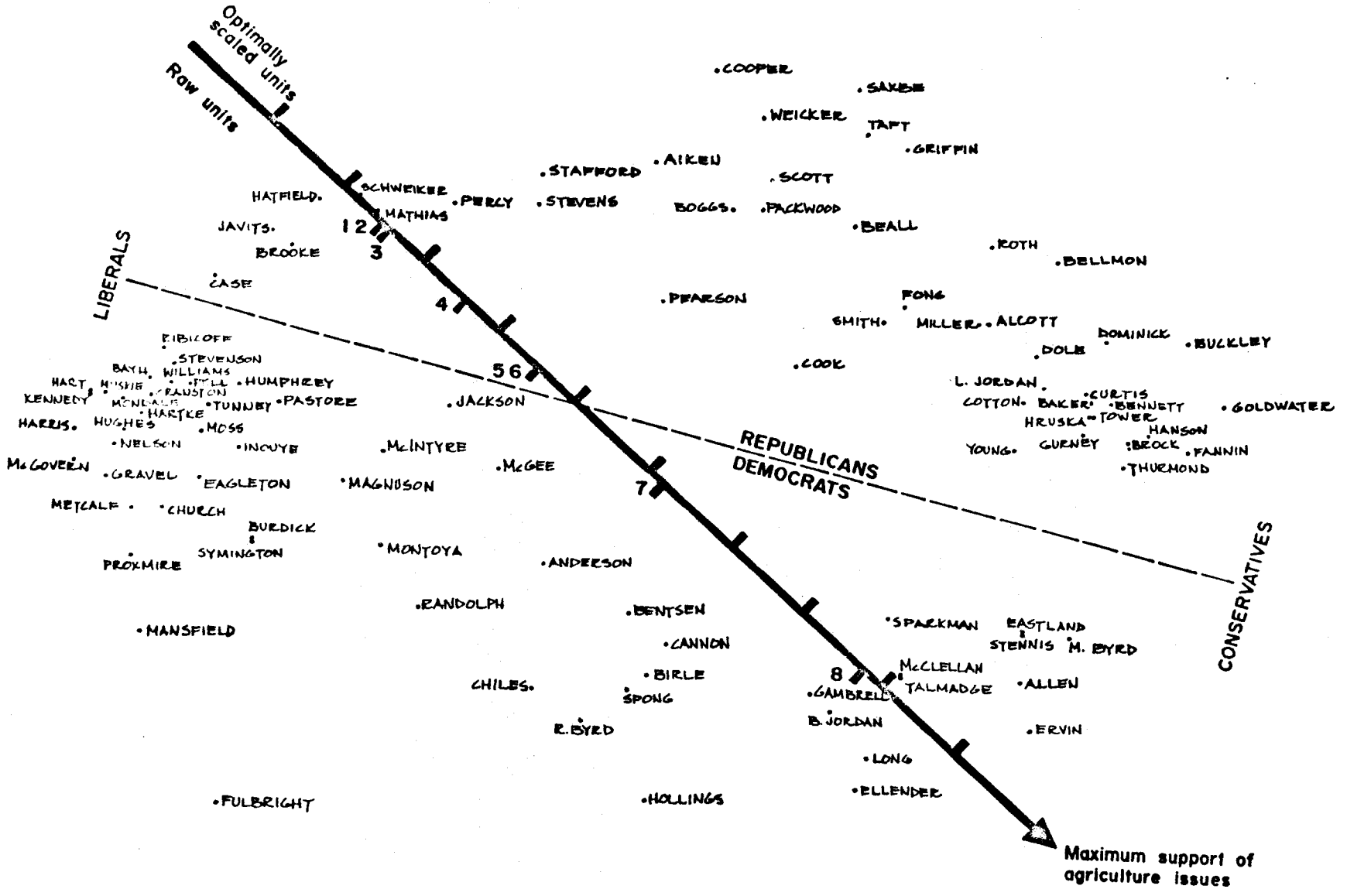

Figure 2

Configuration of U.S. Senators with the optimally regressed agriculture support vector.

along the other side. The relationship between each of these sets of units is the optimal transformation t^{do}. The interpretation of this figure is straight forward: The conservatives (who are in the lower right part of the space) support agricultural issues most strongly, and the liberals least strongly. The strength of support is mostly a function of the liberal-conservative dimension (the horizontal direction) and only a weak function of the party split (the vertical direction). This interpretation corresponds with the weighting scheme presented in Table 4 (dimension 1 is horizontal, dimension 2 vertical).

3.4 Verifying theoretical transformations

One of the special uses of MORALS/CORALS is to empirically determine whether a specific transformation proposed on theoretical grounds is at all like the optimal transformation. If it is, then empirical support is gained for the theoretical argument. Such a situation exists in the area of Psycholinguistics, in which word frequency has served as an independent variable in a large number of studies of verbal learning and verbal behavior (see Carroll and White (1973) for a brief review of these studies). It has been proposed (Carroll, 1970) that a more useful variable is a logarithmic transformation of word frequency, with a specific one, termed the standard frequency index (SFI) being proposed by Carroll. Carroll presented theoretical arguments that SFI ought to be more linearly related to several other variables of interest in these studies than the raw frequency measure. If, when we use MORALS with the appropriate variables, we discover that the optimal transformation of frequency is similar to the SFI transformation, then this lends empirical

support to Carroll's arguments. Thus, MORALS was used to investigate this issue.

In Table 5 we present the results of six MORALS analyses performed with a subset of the data discussed by Carroll and White (1973). In all analyses the dependent variable was the latency with which male and female subjects named each of 93 pictures. This variable was formed, by Carroll and White, by obtaining the reciprocal mean latency over subjects for each word. In all analyses one of the independent variables was the age at which subjects judged they first acquired the meaning of the word (averaged over from 29 to 33 subjects). There were also two additional independent variables which concerned the frequency of occurrence of the word in the English language. One of these variables was obtained from the Thorndike-Lorge frequency count (Thorndike & Lorge, 1944), and the other from the American Heritage Word Frequency Book (Carroll, Davies, & Richman, 1971). Two of the analyses reported in Table 5 included both of these variables, and four included only one of the variables. Furthermore, and of central interest here, for three sets of analyses we used the raw frequencies, and for the other three we used the logarithmically transformed frequencies. For all analyses we assumed that latency and age of acquisition were defined at the interval level of measurement, and that frequency was at the ordinal level. (We also assumed that the measurement process was discrete.) From each analysis we obtained the classical multiple correlation results (i.e., the initial MORALS values) and have presented these in the upper two sections of the table. We also obtained, in each of the six analyses, the MORALS results,

Table 5

Beta coefficients and multiple correlations in prediction of latency data for 93 words

	Analysis 1	2	3
A) Raw Frequency			
Age of Acquisition (mean)	-.763	-.763	-.761
Thorndike-Lorge frequencies	-.015	.020	
American Heritage frequencies	.040		.028
Classical multiple R	.774	.774	.774
B) Log Frequency (SFI)			
Age of Acquisition (mean)	-.476	-.592	-.417
Thorndike-Lorge frequencies	.005	.246	
American Heritage frequencies	.389		.393
Classical multiple R	.816	.791	.816
C) Optimal Frequency			
Age of Acquisition (mean)	-.350	-.537	-.412
Thorndike-Lorge frequencies	.173	.347	
American Heritage frequencies	.440		.507
Optimal multiple R	.859	.815	.852

which are presented in the bottom section of the table. We only present this information for three of the analyses since we obtained identical results regardless of whether we initiated the MORALS process with the raw frequency values or the log frequency values. Thus this is one test of the robustness of the algorithm since the two ways of defining frequency are monotonically equivalent, and should yield identical results when the variables are ordinal.

From the table we can see that the classical multiple correlations based on SFI are an improvement over those based on the raw frequency measure, as stated by Carroll (1973). We can also see that the optimal multiple correlations are a further improvement over the classical multiple correlations, as anticipated. We also note a trend in the beta weights, indicating that as we move from raw frequency through log frequency to optimal frequency the relative weighting of the age of acquisition variable decreases, and the relative weighting of the frequency variables increases, with the American Heritage measures receiving more weight than the Thorndike-Lorge measures. This trend would appear to be in line with Carroll's arguments.

The remaining question is whether the optimal transformations are similar to the logarithmic transformations. The answer to this question can be gained from Figure 3, where we present the optimal transformations of each of the four frequency variables. In these figures the horizontal axes are the raw or log frequency values, and the vertical axes the optimal values. (Note that the optimal transformations of the raw and log variables must be different if MORALS is to obtain the same optimal results. These transformations

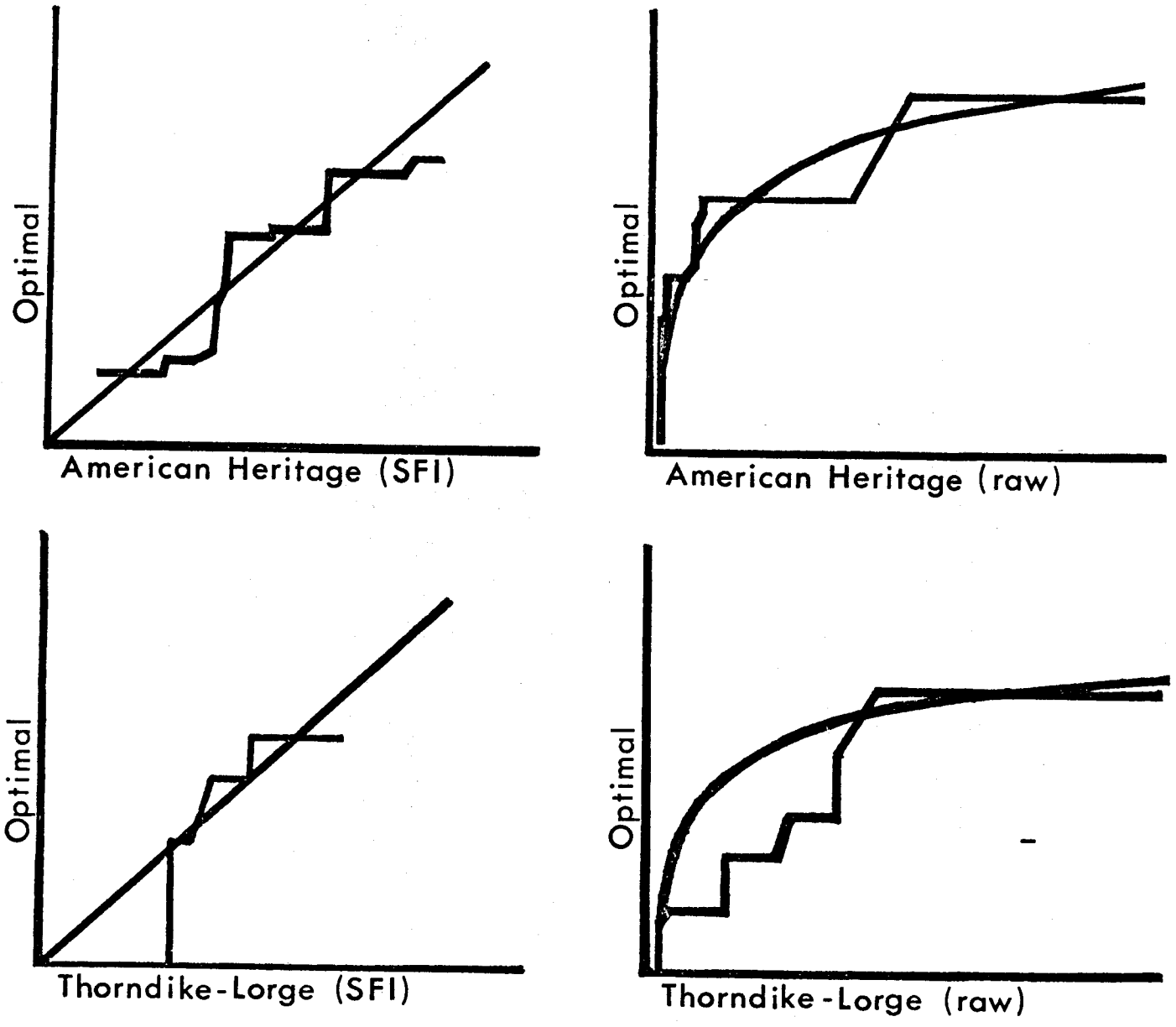

Figure 3

Optimal transformations of the four frequency variables.

differ because the original variables are different, not because the optimally scaled variables are different.) The transformations presented in the figure are from the analyses in which both variables were present (labeled 1 in Table 5), but the essential form of the transformations remained stable over all analyses. It is our conclusion that the optimal transformation of the Thorndike-Lorge variable (indicated by the jagged line) shows a tendency to depart from the SFI transformation (indicated by the smooth line), but that the American-Heritage variable does not. The optimal Thorndike-Lorge variable seems to be transformed in the same manner as the logarithmic transformation, but more extremely, with the smallest frequencies being stretched greatly and the medium and large frequencies being compressed, essentially, into a single value. On the other hand, while there are differences between the optimal and logarithmic American Heritage variables, the differences do not seem to form any clear trend. When we recall that the American Heritage log and optimal variables receive much more weight than the Thorndike-Lorge variables, we reach the conclusion that the theoretical considerations leading to the SFI transformation have been lent empirical support by the results of the MORALS analysis, especially in the case of the American Heritage data.

4.0 Discussion

The results just presented lead us to conclude that the MORALS/CORALS approach to multiple/canonical correlation with a mixture of qualitative and quantitative data is capable of obtaining meaningful results in a variety of situations. Thus our approach not only extends correlation analysis to commonly occurring situations which were not previously covered, but appears to do it in an efficacious

manner.

4.1 Measurement issues.

We have discussed several implications of our approach to measurement elsewhere (Takane, Young & de Leeuw, 1975) so we do not go into them in detail here, but simply present a summary.

The most important point is that within our framework it is possible to obtain empirical information about the measurement level of each variable, specifically within the context posed by the correlation model. All one has to do is to analyze the data several times, making different measurement level assumptions each time. If two (or more) of these analyses yield precisely the same results then the appropriate measurement level is the highest one used for the several otherwise equivalent analyses. It is then possible to conclude that for the chosen model the true measurement level is that highest one, and that this is not simply an assumption of the appropriate level, but an empirically determined level.

As was discussed in our earlier work, the view of measurement implied by the preceding statements is not the common view. We do not hold that measurement level is a characteristic of data in vacuo, rather we hold that the measurement level of a variable is dependent on the interaction of the variable with the model chosen to describe the variable. Our view of measurement differs from the common view in yet another way, and that is that we view measurement level as being a continuous, not a discrete, notion, with certain points on the measurement level continuum being axiomatizable, and many more points being non-axiomatizable. These issues are fully discussed by Takane, Young & de Leeuw (1975).

4.2 Extensions

There are several interesting and potentially useful extensions of the work presented here. As was stated at the outset (Eqs. 1-4), our procedure attempts to optimize a criterion which can be stated, in summation notation, as

$$(25) \qquad C = \sum_{i}^{N} \left(\sum_{a}^{n_a} x^{*}_{ai} - \sum_{b}^{n_b} y^{*}_{bi} \right)^2$$

where N is the number of observation vectors, n_a is the number of variables in set one, n_b is the number of variables in set two, and where we have absorbed the regression weights into the scaling of the optimally scaled variables x^{*}_{ai}. As is emphasized by the summation notation, our procedure is the analog of a canonical correlation procedure since it attempts to optimally scale each variable (within the stated measurement restrictions) such that a linear combination of all of the variables in one set is as much as possible like a linear combination of all of the variables in the other sets.

One possible extension of our work is to modify the order in which the summations are taken in Eq. (25). This can be done in two ways. One of these is to define the optimality criterion

$$(26) \qquad M = \sum_{i}^{N} \sum_{a}^{n_a} \left(x^{*}_{ai} - \sum_{b}^{n_b} y^{*}_{ba} \right)^2$$

where everything is defined as in Eq. (25). A procedure could be developed within the ALS framework which would optimize Eq. (26). Such a procedure should most properly be thought of as an extension of multivariate multiple regression to the qualitative data situation, since in multivariate regression one obtains a linear combination

of the variables in one set such that the combination is as much like each variable in the other set as is possible. The other type of modification to the summations in Eq. (25) involves the optimality criterion

$$(27) \qquad H = \sum_{i}^{N} \sum_{a}^{n_a} \sum_{b}^{n_b} (x^*_{ai} - y^*_{bi})^2 .$$

With this criterion, and with the appropriate algorithm, the variables would be optimally scaled so that each variable in set one is as much as possible like each variable in set two. Such a procedure might be most accurately called homogeneity analysis, since it attempts to make all variables as homogeneous as possible.

A second general class of possible extensions is to those situations in which one of the sets of variables is unobserved. For example, if we are operating with the homogeneity notion in Eq. (27), but wish to obtain a set of latent variables which are as homogeneous as possible, then we could define as optimality criterion

$$(28) \qquad L = \sum_{i}^{N} \sum_{b}^{n_b} (\theta_{bi} - y^*_{bi})^2$$

where θ_{bi} are the latent variables. If we further wish to factor the latent variables into their t principal components,

$$(29) \qquad \theta_{bi} = \sum_{p}^{t} \beta_{bp} \alpha_{pi}$$

then we obtain an analysis for the principal components of qualitative data, since such a procedure would find a matrix of components which are most like the variables in the usual principal components sense (Shepard & Kruskal, 1974).

Another example of this class of extension would be a slightly different modification of the homogeneity criterion, namely

$$K = \sum_{i}^{N} \sum_{b}^{n_b} (\theta_i - y^*_{bi})^2. \tag{30}$$

With this criterion we would attempt to optimally scale the variables to be as much like their mean, since the least squares solution for the single latent variable is the mean of the other variables. This type of homogeneity analysis has been investigated by Lingoes (1973) in his CM-II program.

A third general type of extension of our developments would be in the direction of multidimensional quantification. That is, we might be interested in obtaining the first several canonical variates which jointly optimize the canonical correlation. To do this we could redefine Eq. (25) to be

$$\sum_{q}^{r} c_q = \sum_{q}^{r} \sum_{i}^{N} (\sum_{a}^{n_a} x^*_{aiq} - \sum_{b}^{n_b} y^*_{biq})^2 \tag{31}$$

where we are interested in obtaining r canonical variates. A procedure for optimizing Eq. (29) would be the analog of the usual canonical correlation situation in which several canonical variates are desired. Similar types of extensions could be made to the multivariate correlation situation (Eq. 26), and to the homogeneity situation (Eq. 27).

Finally, it would be possible to extend our development to include multiple sets of variables. While several alternatives are available at this point, one would be able to define an index

$$(32) \qquad Q = \sum_{m>\ell}^{t} \sum_{i}^{t-1} \sum_{a}^{N} \left(\sum_{a}^{n_m} x^{*}_{aim} - \sum_{b}^{n_\ell} x^{*}_{bi\ell} \right)^2 = 2 \sum_{m} \sum_{i} \left(\theta_i - \sum_{a}^{n_m} x^{*}_{aim} \right)$$

where there are t sets of variables. Such a proposal is analogous to the corresponding situation with quantitative variables.

4.3 Reliability and Validity

Due to the great flexibility of the MORALS/CORALS procedure, one may question whether the optimal transformations and linear combinations obtained for a particular set of observations will be replicated by a new set of observations in an otherwise identical empirical situation. That is, does the procedure provide reliable estimates of the optimal transformations and of the linear combinations, or will these estimates fluctuate wildly from sample to sample?

Certainly, in some situations the procedure will yield unreliable estimates, and in others it will yield reliable estimates. The issue centers around the ratio of the number of parameters being estimated to the number of observations, with this ratio being most favorable in the classical regression situation (i.e., when all variables are at the interval level of measurement), and least favorable when all variables are nominal, especially when there are relatively few observations in each category. As has been noted in the companion paper (de Leeuw, Young & Takane, 1975, section 4.1) a necessary condition for a unique solution is that there be at least two observations in at least one category. For the present situation this concept applies to every nominal variable: If at least one nominal variable consists entirely of unique categories (i.e., there is only one observation in each category) then the solution is not uniquely determined, with a very large number of arbitrary

quantifications of the categories yielding correlation coefficients of 1.0. Since the solution is not unique, it is unreliable. Thus, one should avoid nominal variables which have observation categories represented by only one observation. There are other types of situations which also do not yield unique solutions, some of which are discussed in the companion paper.

Generally, one should suspect the reliability of his results in the same situations in which he suspects the reliability of the results of a classical multiple or canonical regression, except that the problems are compounded when the measurement levels of the variables are weakened. Thus, if one has a small number of observations, and a relatively large number of variables several of which are qualitative, then he should suspect the reliability of the results. The survey example given above (section 3.2) is an example in which the reliability of the results are suspect. In this example there are a total of four dependent variables and eight independent variables. All of these twelve variables are qualitative, nine being ordinal and three nominal. Note that there are only 33 cases, seven of which are removed from most of the analyses. Thus, the first canonical analysis involved 12 variables and 26 cases, the second, ten variables and 26 cases. The first two multiple correlation analyses each involved nine variables and 26 cases, and the last two, three variables and 31 cases. It is not surprising that essentially perfect multiple and canonical correlation coefficients are obtained in four of the six analyses. Due to the variable/case ratio one might even suspect the results even if all of the variables were quantitative. We do note, however, that the transformations and

the weight patterns were interpretable, although the interpretation process was highly subjective.

The other two examples would seem to be reliable. The MDS example involves 99 "cases" (Senators) and only three variables, only one of which is qualitative. Similarly, the word frequency example involves 93 "cases" (words) and five variables, only two of which are qualitative. In both instances the qualitative variables are ordinal. Thus it would appear to be the case that these analyses should be nearly as reliable as their classical counterparts. The ease of interpretation in both examples, and the congruence of the interpretation with the investigators' expectations also attest to the reliability of these examples. (Naturally, if reliability is of paramount importance, then any of the standard procedures for empirically determining reliability can be used, including such procedures as split-half analyses and repeated observation designs.)

4.4 Three schools of thought about additivity

There have been three essentially separate traditions centered around the additive model. The oldest and most widely known tradition falls under the rubric Analysis of Variance, with the central focus being on the development of inferential tests of the significance of the model's components. Nearly as old, but much less widely known, is the Optimal Scaling tradition, with the central focus being on the quantification of qualitative variables within the context of the additive model. The most recent tradition is Conjoint Measurement, which focuses on the axiom systems underlying the additive model. Our work is clearly within the optimal scaling tradition, and makes no pretenses about the development of inferential tests

or axiomatic systems. We are simply interested in scaling qualitative variables so that they will be as linearly related as possible.

In the Analysis of Variance tradition the additive model is associated with assumptions about the distributional properties of the observations and/or the errors, additional assumptions about null and alternative hypotheses, and formal procedures for accepting or rejecting the null hypotheses. In terms of the assumptions which must be satisfied for the significance tests to be valid, the Analysis of Variance tradition requires that the error components be independently normally distributed, generally with zero means and equal variances. The formal theory of parametric statistics (and thus of the Analysis of Variance) does not concern either the scale level of the variables, nor the axiomatic system underlying the model. The only critical aspect is the distributional assumption. Moreover, when the null hypotheses are invariant over monotonic transformations of the dependent variable (as is usually the case), the tests can be easily generalized to any other dependent variable which is monotonically equivalent, if the distributional assumptions of the tests are met. Thus there is a tradition within the Analysis of Variance literature of monotonically transforming variables so that the distributional assumptions will be met.

In the Conjoint Measurement view, the additive model is conceptualized as representing the "true" nature of the process which generated the dependent measures. Measurement theory is concerned with a) specifying the exact nature of the conditions under which the dependent measures will be precisely representable by the model; and b) identifying the types of transformations of the dependent

variables which are allowed within the previously identified conditions. In the case of the additive model, conjoint measurement has postulated a set of axioms concerning the structure of the data which are the necessary and sufficient conditions for the additive model to precisely describe the data. Not all of the axioms are empirically testable, but if those that can be tested are satisfied, then there is a monotonic transformation of the dependent variable which allows the variable to be precisely described by the additive model. Thus the central focus of the conjoint measurement tradition is specifying conditions under which monotonic transformations of the dependent variable permit an additive description of the variable.

In the Optimal Scaling tradition the basic goal is to obtain a transformation of the dependent (and independent) variables so that the additive model fits as well as possible, as should be obvious from the body of this paper. This focus is similar to one of the Conjoint Measurement goals, but the difference is that those working in the Conjoint Measurement tradition are interested in specifying the conditions under which the simple additive model will describe the data perfectly, whereas the Optimal Scaling tradition centers on actually providing the best description.

Note that the Conjoint Measurement and Optimal Scaling traditions are similar in that they are both descriptive: Each is concerned with describing the structure of the data with the additive model, one centering on whether a perfect description is possible, and the other one providing the best possible description. The Analysis of Variance tradition, on the other hand, is inferential: It is concerned with making inferences beyond the data on

the basis of the additive model. All three traditions are concerned with data transformations, but the essential difference between the Analysis of Variance tradition and the other two, is that in the former a transformation is desired to improve the adequacy of the inferential process, whereas in the latter a transformation is desired to improve the adequacy of the descriptive process.

Perhaps it should be emphasized that a useful function is performed within each of the traditional approaches to additivity, and that perhaps the best approach is a combination of all three. It seems to the present authors that each approach by itself gives an incomplete picture of the data being analyzed, and that the most complete picture is derived when all three approaches are used. Surely we best understand a particular set of data when we know the degree to which it satisfies the axioms underlying the additive model, when we know the transformations which make the data most additive, and when we know the significance of the components of the additive model.

References

Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T., & Silverman, E. An empirical distribution function for sampling with incomplete information. Annals of Mathematical Statistics, 1955, 26, 641-647.

Bock, R.D. Methods and applications of optimal scaling. Psychometric Laboratory Report #25, University of North Carolina, 1960.

Carroll, J.B. An alternative to Juilland's usage coefficient for lexical frequencies, and a proposal for a Standard Frequency Index (SFI). Computer Studies in the Humanities and Verbal Behavior, 1970, 3, 61-65.

Carroll, J.B., Davies, P., & Richman, B. The American Heritage Word Frequency Book, Boston: Houghton Mifflin, 1971.

Carroll, J.B., & White, M.N. Age-of-acquisition norms for 220 picturable nouns. Journal of Verbal Learning and Verbal Behavior, 1973, 12, 563-576.

Carroll, J.D. Individual differences and multidimensional scaling. In Shepard, R.N., Romney, A.K., & Nerlove, S. Multidimensional Scaling: Theory and Applications in the Behavioral Sciences. Seminar Press, New York, 1972.

Carroll, J.D., & Chang, J.J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. Psychometrika, 1970, 35 283-319.

Cliff, N., & Young, F.W. On the relation between unidimensional judgments and multidimensional scaling. Organizational Behavior and Human Performance, 1968, 3, 207-208.

de Leeuw, J. Nonmetric discriminant analysis. Research Note RN006-68. Leiden, the Netherlands: University of Leiden, Data Theory, 1968.

de Leeuw, J. The linear nonmetric model. Research Note RN003-69. Leiden, the Netherlands: University of Leiden, Data Theory, 1969.

de Leeuw, J. Canonical analysis of categorical data. Doctoral dissertation, University of Leiden, the Netherlands, 1973.

de Leeuw, J. Normalized cone regression. Leiden, the Netherlands: University of Leiden, Data Theory, mimeographed paper, 1975.

de Leeuw, J., Young, F.W., & Takane, Y. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. Psychometric Laboratory Report # 140. Chapel Hill: L.L. Thurstone Psychometric Laboratory, 1975.

Fisher, R. Statistical Methods for Research Workers (10th ed.). Edinburgh: Oliver and Boyd, 1938.

Hayashi, C. On the quantification of qualitative data from the mathematico:statistical point of view. Annals of the Institute of Statistical Mathematics, 1950, 2, 35-47.

Hotelling, H. The most predictable criterion. Journal of Educational Psychology, 1935, 26, 139-142.

Hoadley, J. Spatial analysis of Senate voting patterns. Master's thesis, University of North Carolina, 1974.

Kruskal, J.B. Nonmetric multidimensional scaling. Psychometrika, 1964, 29, 1-27, 115-129.

Lingoes, J.C. The Guttman-Lingoes Nonmetric Program Series. Ann Arbor, Mich.: Mathesis Press, 1973.

Shepard, R.N., & Kruskal, J.B. A nonmetric variety of linear factor analysis. Psychometrika, 1974, 39, 123-157.

Srinivasan, V. Linear programming computational procedures for ordinal regression. Unpublished manuscript, Graduate School of Business, Stanford University, 1974.

Takane, Y., Young, F.W., & de Leeuw, J. Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. Psychometric Laboratory Report (in preparation) L.L. Thurstone Psychometric Laboratory, Chapel Hill, N.C., 1975.

Thorndike, E.L., & Lorge, I. The Teacher's Wordbook of 30,000 Words. New York: Teachers College Press, 1944.

Wold, H., & Lytkens, E. (Eds.) Nonlinear iterative partial least squares (NIPALS) estimation procedures (group report). Bulletin of the International Statistical Institute, 1969, 43, 29-51.

Young, F. W. A model for polynomial conjoint analysis algorithms. In Shepard, R.N., Romney, A.K., & Nerlove, S. Multidimensional Scaling, Vol. 1. New York: Seminar Press, 1972.

Young, F.W. Methods for describing ordinal data with cardinal models. Journal of Mathematical Psychology, 1975, 12 (in press).