NONMETRIC INDIVIDUAL DIFFERENCES MULTIDIMENSIONAL SCALING: AN ALTERNATING LEAST SQUARES METHOD WITH OPTIMAL SCALING FEATURES

YOSHIO TAKANE AND FORREST W. YOUNG

UNIVERSITY OF NORTH CAROLINA

JAN DE LEEUW

UNIVERSITY OF LEIDEN

A new procedure is discussed which fits either the weighted or simple Euclidian model to data that may (a) be defined at either the nominal, ordinal, interval or ratio levels of measurement; (b) have missing observations; (c) be symmetric or asymmetric; (d) be conditional or unconditional; (e) be replicated or unreplicated; and (f) be continuous or discrete. Various special cases of the procedure include the most commonly used individual differences multidimensional scaling models, the familiar nonmetric multidimensional scaling model, and several other previously undiscussed variants.

The procedure optimizes the fit of the model directly to the data (not to scalar products determined from the data) by an alternating least squares procedure which is convergent, very quick, and relatively free from local minimum problems.

The procedure is evaluated via both Monte Carlo and empirical data. It is found to be robust in the face of measurement error, capable of recovering the true underlying configuration in the Monte Carlo situation, and capable of obtaining structures equivalent to those obtained by other less general procedures in the empirical situation.

Key words: Euclidian model, INDSCAL, measurement, similarities, data analysis, similarities data, quantification, successive block algorithm.

1. Purpose and Motivation

One of the most vigorous areas of endeavor in recent multi-dimensional scaling research concerns the representation of individual differences. The weighted Euclidian model is currently the most widely used individual

differences model of the various ones that have been proposed. One of the
main attractions of this model undoubtedly relates to the strict isolation of
information common to all individuals from information unique to each
individual. The idea of representing communality among sets of observations
by a single multidimensional Euclidean space, while representing the unique-
ness of each individual by differential weights attached to the dimensions
of the space is an ingeneous idea particularly conductive to simple and
straightforward interpretation. The fact that the dimensions of the space
are unrotatable makes the model even more attractive.

   The weighted Euclidian model is certainly not the most general individual
differences model proposed within the multidimensional scaling framework
[Tucker, 1972], nor is it appropriate to all types of individual differences
[McGee, 1968]. Furthermore, the most successful implementation of the
model [Carroll & Chang, 1970] is severely limited in terms of the types of
data to which the model can be applied, particularly in light of recent interest
in nonmetric multidimensional scaling [Kruskal, 1964].

   It is the purpose of this paper to propose and evaluate a new procedure
for fitting the weighted Euclidian model to data that are much less severely
restricted than those appropriate to the Carroll-Chang procedure. Our
procedure is appropriate to data that may have missing observations, that
may be defined at the nominal, ordinal, interval or ratio measurement levels,
that may be discrete or continuous, and that may or may not be asymmetric,
conditional or replicated. Furthermore, our procedure is able, without
further complications, to fit the simple unweighted Euclidian model. Thus
several individual differences models [Carroll & Chang, 1970; McGee, 1968;
Young, 1975] as well as models not including individual differences notions
[Kruskal, 1964; Torgerson, 1952] and other previously undiscussed variants
can be realized within one common framework.

   The weighted Euclidian model and the associated procedures for fitting
the model to empirical data were proposed by several people at about the
same time [Horan, 1969; Bloxom, Note 1; Carroll & Chang, 1970]. The most
successful procedure and the most complete proposal is that of Carroll and
Chang. Their INDSCAL (*in*dividual *d*ifferences *scal*ing) procedure is formally
an n-way generalization of Eckart and Young's [1936] two-way canonical
decomposition which Carroll and Chang call the CANDECOMP procedure.
This procedure is performed, after an initial conversion of observed dissimi-
larities to product moments, by alternately obtaining least squares estimates
of the individual differences weights W (for fixed estimates of the stimulus
configuration X), and then obtaining least squares estimates of X given W.
This procedure belongs to a class of numerical procedures termed alternating
least squares (ALS) procedures by de Leeuw, Young and Takane [1976],
which have the desirable property of being necessarily convergent. That is,
it is never possible for an ALS procedure to obtain an iteration which worsens

the function it is designed to optimize. On every iteration the function must be improved due to the conditional least squares properties of each phase of an ALS procedure. More will be said on this later.

The Carroll-Chang CANDECOMP procedure has two consequences which are relevant to the present discussion. First, the minimization criterion (called STRAIN, by Carroll) is defined in terms of the product moments computed from the raw data, not in terms of the raw data themselves. Thus INDSCAL does not optimize the fit between the weighted Euclidian model and the data, strictly speaking, but rather the fit between a vector product model and a transformation of the data. Second, due to the operation which converts dissimilarities into scalar products (which involves addition, etc.), the procedure is metric.

Bloxom [Note 1] proposed a gradient procedure to optimize STRAIN which is also a metric procedure. Unfortunately, due to the nature of gradient procedures, the convergence properties of the Carroll-Chang ALS-type procedure are lost. This may account for the reported [Carroll & Chang, 1970] inferiority of Bloxom's procedure in terms of speed of convergence relative to the INDSCAL procedure. Perhaps for this reason Bloxom [1974] proposed another procedure based on the equivalence of the problem as posed in the STRAIN framework to the analysis of covariance structures proposed by Jöreskog [1970]. The performance of this proposal has yet to be investigated.

Schönemann [1972] presents an elegant algebraic solution for the weighted Euclidian model. However, since the logic of his developments is not oriented towards optimizing a well defined quantity, it cannot be applied to real data with the expectation of unqualified success, as Schönemann notes. This means that the procedure has little practical significance to the data analyst. His idea, however, has been extended by de Leeuw [Note 5] to obtain a rational initial start to be used for more robust procedures for fitting the weighted Euclidian model. We will go into this topic further in later portions of this paper.

All of the procedures discussed up to this point place very stringent requirements on the data. Specifically, they all require that the data be symmetric, have no missing observations, be unreplicated and unconditional, and be defined at least at the interval level of measurement. Several procedures which relax some or all of these restrictions have been proposed and investigated, with varying degrees of success.

Carroll and Chang's first nonmetric procedure, mentioned briefly in their original paper [1970] and called NINDSCAL (nonmetric INDSCAL) is a two-phase procedure which uses the metric CANDECOMP procedure in the first phase (iteratively until convergence) and Kruskal's [1964] least squares monotonic regression in the second phase. These two phases are iteratively applied. It is important to note that the first phase minimizes STRAIN (which is defined on scalar products as discussed above), whereas

the second phase minimizes Kruskal's STRESS, which is defined on the raw data. Since two different functions are involved, NINDSCAL has no assurance of convergence on a stable point, and eventually either oscilates or diverges after a few iterations. Furthermore, the procedure is very inefficient, and of the several data restrictions noted relaxes only the measurement level requirements.

For these reasons, Carroll and Chang have recently [Note 3] proposed another nonmetric procedure to minimize STRAIN that uses an ALS method after initial estimates of W and X are obtained by an improved CANDECOMP procedure. This approach, which involves STRAIN in all phases of each iteration, is the first stable procedure for nonmetric multidimensional scaling which involves the weighted Euclidian model. It has the highly desirable consequence of relaxing all of the data restrictions noted above. However, the procedure is within the STRAIN framework, and thus does not directly optimize the fit between the distance model and the raw data, but rather between the scalar products computed from an optimal monotonic transformation of the raw data and the scalar products computed from the coordinates. Of the various procedures reviewed here, this is the soundest (at least theoretically), although its efficiency is yet to be reported.

A third nonmetric procedure for fitting the weighted Euclidian model has been tried by the second author of this paper. This procedure uses a gradient technique to simultaneously improve estimates of W and X by using the derivatives of the STRESS loss function. While this procedure (a) uses one loss function throughout the entire procedure, and (b) optimizes the fit to the data directly, it has been found to be highly susceptible to the exact nature of the starting point. A careful choice of the initial orientation of X is required. Although this difficulty could be remedied by using de Leeuw's [Note 5] initial rotation procedure (as is done in the work to be reported here), it appears to be the case that the procedure still suffers from the use of the gradient procedure.

Finally, a gradient procedure has been proposed by Yates [Note 17] for nonmetrically fitting the weighted Euclidian model. This procedure is in neither the STRESS or STRAIN framework; rather, it attempts to minimize the proportion of variance in the model which is due to incorrectly ordered pairs of distances (relative to the order of the dissimilarities). This goal has been adopted by several authors in the context of the unweighted Euclidian model [Guttman, Note 9; de Leeuw, Note 6; Johnson, 1973], and has been fully discussed by de Leeuw [1975] and Young [1975]. While this procedure has the advantage of optimizing a relationship defined directly in terms of the raw data and subjects the data to none of the restrictions mentioned above, it suffers from mixing together two different optimizing functions, as shown by de Leeuw [1975] and discussed by Young [1975].

In this paper we present a new nonmetric procedure for fitting the

weighted Euclidian model which a) is in the STRESS framework; b) uses the ALS approach; and c) removes all of the data restrictions mentioned above.

2. The Problem

The problem we solve in this paper is that of obtaining a robust and efficient procedure for nonmetric individual differences multidimensional scaling. In this section we discuss the most important aspects of the problem, namely the individual differences models, the types of data, and the optimization criterion utilized in our work.

Individual Differences Models

As emphasized in the previous section, we use the weighted Euclidian model to represent individual differences. This model is

$$(1) \qquad d_{ijk}{}^2 = \sum_{a=1}^{t} w_{ia}(x_{ja} - x_{ka})^2, \qquad w_{ia} \geq 0,$$

as is well known (the non-negativity restriction is optional). However, as was briefly mentioned in the preceding section, we also treat the (unweighted) Euclidian model within our framework. This model is equivalent to (1) when all $w_{ia} = 1$, and can also be viewed as an individual differences model in certain circumstances. We will discuss the full variety of models subsumed by (1) in Section 5 of this paper.

Types of Data

Previous authors of multidimensional scaling papers [Shepard, 1962; Kruskal, 1964; Guttman, 1968; Carroll & Chang, 1970] have emphasized a dichotomy of measurement levels which they termed metric and nonmetric. When placed in the context of Stevens's [1951] measurement theory, it is clear that these terms correspond to three of the four measurement levels delineated by Stevens, namely ordinal (nonmetric) and interval or ratio (metric). The developments presented here, on the other hand, extend multidimensional scaling to data defined at all four of Stevens's levels, including the nominal level. Furthermore, we also distinguish two types of measurement processes (discrete and continuous) and three types of conditionality (unconditional, matrix-conditional, and row-conditional). While we discuss these notions here as though they form "types of data," this is a pedagogical simplification of our philosophical position, as will be discussed in Section 5.

The general nature of the problem faced by an analysis procedure explicitly designed for data having such a wide variety of measurement characteristics is best viewed in the light shed by Fisher's notion of optimal scaling [Fisher, 1946]. Fisher's objective in proposing optimal scaling was to scale the observations so that (a) they would fit the model as well as possible in a least squares sense, and (b) the measurement characteristics of the observations would

be strictly maintained. Fisher's optimal scaling notion is one of the corner-stones of our own work.

Let us define the squared observations O, the optimally scaled squared observations D^*, and the squared distances D. (The optimally scaled squared observations are commonly referred to as the disparities in the MDS context. We sometimes refer to them as the estimates, since they are least squares estimates of the squared distances.) Each of these symbols represents a collection of matrices. That is, O is a collection of all matrices O_i for all individuals i from whom we have obtained observations o_{ijk} about stimulus pairs (j, k). Correspondingly, D^* is the collection of matrices D_i^* with elements d_{ijk}^{*2}, and D is the collection of all matrices D with elements d_{ijk}^2 defined by (1).

With these definitions we can formally represent the optimal scaling problem as a transformation problem, as follows. We wish to obtain a transformation t of the raw observations which generates the optimally scaled observations d_{ijk}^*; i.e.,

$$(2) \qquad\qquad t[o_{ijk}] = [d_{ijk}^*],$$

where the precise definition of t is a function of the measurement level, process, and conditionality, and is such that a least squares relationship exists between d_{ijk}^* and d_{ijk} given that the measurement characteristics are strictly maintained. In the remainder of this section we discuss in detail the measurement restrictions which must be maintained. In a later section we present the corresponding least squares methods for obtaining the transformations.

To fully understand the several levels, process, and conditionality 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. Stated most 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 three types of measurement restrictions (level, process, and conditionality restrictions) concern three different aspects of the observation categories. The process restrictions concern the relationships among all the observations *within* a single category; the level restrictions concern the relationships among all the observations *between* different categories; and the conditionality restrictions concern the possibility of sets of categories. We will first take up the process restrictions, then the level restrictions, and finally the conditionality 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 that it is continuous. 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. These process restrictions are related to the "primary-secondary" distinction discussed by Kruskal [1964], and to the "weak-strong" distinction discussed by Guttman [1968]. In the discrete 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. In the continuous 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

$$(3) \qquad t^d : (o_{ijk} \sim o_{mno}) \rightarrow (d_{ijk}{}^* = d_{mno}{}^*),$$

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

$$(4) \qquad t^c : (o_{ijk} \sim o_{mno}) \rightarrow \begin{cases} (d_{ijk}{}^- = d_{mno}{}^- \leq d_{ijk}{}^* \leq d_{ijk}{}^+ = d_{mno}{}^+), \\ (d_{ijk}{}^- = d_{mno}{}^- \leq d_{mno}{}^* \leq d_{ijk}{}^+ = d_{mno}{}^+), \end{cases}$$

where $d_{ijk}{}^-$ and $d_{ijk}{}^+$ are the lower and upper bounds of the 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. Denoting this, however, 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 not to be equal.

We now turn to the second set of restraints on the several measurement transformations t, the level restraints. With these restraints we determine the nature of the allowable transformations t 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 Stevens's four measurement levels.

For nominal variables, we introduce no level restraints as the characteristics of nominal variables are completely specified by the previously mentioned process restraints.

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. That is,

$$(5) \qquad t^{o} : (o_{ijk} < o_{mno}) \rightarrow (d_{ijk}{}^{*} \leq d_{mno}{}^{*}),$$

where the superscript on t^{o} indicates the order restriction, and where $<$ 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 or ratio) variables, we require that the real numbers assigned to the observations be linearly related to the observations. That is,

$$(6) \qquad t^{l} : d_{ijk}{}^{*} = \delta_{o} + \delta_{1}o_{ijk} ,$$

where $\delta_{o} = 0$ for ratio variables. When necessary we denote the interval transformation as t^{i} and the ratio transformation as t^{r}. More generally, we may require that the assigned numbers be related to the observations by a polynomial of known degree:

$$(7) \qquad t^{p} : d_{ijk}{}^{*} = \sum_{q=0}^{p} \delta_{q}o_{ijk}{}^{q},$$

(where the summation starts at 1 for ratio variables). 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.

Finally, we turn to the third type of measurement restrictions, those concerning the conditionality of the observations. As has been emphasized

by Coombs [1964], it may be that the measurement characteristics of the observations are conditional on some aspect of the experimental situation in such a way that some observations cannot be meaningfully compared with other observations. For example, if several subjects in a paired comparison similarity experiment are required to judge the similarity of all pairs of stimuli, we are usually unwilling to say that one subject's judgment of 7 (on a similarity scale of 1 through 9, for example) can be said to represent more similarity than another subject's judgment of 6. We just are not sure that the subjects are using the response scale in identical ways. In fact, we are pretty sure that they do not use the scale identically, so we say that the measurements are conditional on the subject. More generally, we refer to this type of conditionality as matrix-conditionality, since all observations within a matrix are comparable, but not those observations between matrices. It is also possible to have row-conditional observations, as discussed by Coombs [1964, Ch. 17], and unconditional observations. (Note that Coombs' unconditional case corresponds with our matrix-conditional case.)

Formally, we state that the domain of the measurement transformation t is dependent on the type of conditionality. For unconditional data the domain is the entire set of observations and the transformation is denoted t. For matrix-conditional data the domain is a single matrix of data and the transformation is denoted t_i . Finally, for row-conditional data the domain is a single row of a single matrix, and the transformation is denoted t_{ij} . The previous discussion of measurement level and process were implicitly in terms of unconditional data, and all of the definitions of level and process must be modified appropriately. We do not explicate these modifications here as they are lengthy and obvious. Of course other patterns of conditionality are possible, though unlikely. It may also sometimes be the case that different measurement levels or processes may be associated with conditionality. We do not go into these generalizations in this paper, although they have been discussed by Young [1973] and Kruskal, Young and Seery [Note 14].

Optimization Criteria

Most of the procedures for fitting the weighted Euclidian model which we discussed in the first section were in the STRAIN framework. That is, they were designed to optimize a suitably normalized version of the function

$$(8) \qquad \zeta^2(X, W, P^*) = \sum_{i=1}^{N} \text{tr} \, (P_i^* - XW_iX')'(P_i^* - XW_iX'),$$

where P^* is the collection of P_i^* for $i = 1, \cdots , N$, where W_i is a diagonal matrix of weights for subject i, and where P_i^* is the matrix of pseudoscalar products derived from subject i's dissimilarities under either metric or nonmetric assumptions. It should be emphasized that the pseudoscalar products P^* are determined from the optimally scaled data D^* according to the pro-

cedure first suggested by Eckart and Young [1936]. This procedure involves multiplying each element $d_{ijk}*^2$ by $-1/2$ and then, using these $-1/2d_{ijk}*^2$, subtracting the row and column mean from and adding the matrix mean to each $-1/2d_{ijk}*^2$. This process, known as doubling centering, yields the elements $p_{ijk}*$ of $P*$. The matrix P_i* is called a matrix of pseudoscalar products because if r real principal components of P_i* are obtained, and if these components are arranged column-wise in a matrix F_i, then $P_i = F_iF_i'$ is the rank r matrix which is a least squares fit to P_i*. Commonly, both P_i and P_i* are referred to as scalar product or product moment matrices, but sometimes, for the sake of clarity, we refer to P_i* as "pseudo" scalar products, and P_i as scalar products.

Equation (8), STRAIN, is a least squares criterion defined between the scalar products derived from the data and the scalar products derived from the model. Although the optimization of STRAIN is very straightforward when the data are metric, it is rather complicated when they are nonmetric. Two fundamentally different optimization procedures have been proposed. The more satisfactory of these approaches, proposed by Carroll and Chang [Note 3], assumes that the observed dissimilarities must be monotonic with a set of values from which the scalar products P_i* are computed. That is, it is required that

$$(9) \qquad\qquad t^o[o_{ijk}] = [d_{ijk}*],$$

so that P_i* may be computed from D_i* in a way which optimizes STRAIN. While the measurement aspects of this approach are sound, the optimization problem is very complex, and the efficiency and robustness of the procedure is yet to be documented. The other, less satisfactory approach, taken by Levinsohn and Young [1974], involves computing a matrix of scalar products P_i directly from the raw observations at the outset of the analysis. The procedure then optimizes STRAIN under the assumption that P_i is non-metric. That is, this procedure requires that

$$(10) \qquad\qquad t^o[p_{ijk}] = [p_{ijk}*].$$

Certainly the measurement aspects of this approach are confusing since the data must be assumed to be metric in order to derive the scalar products which are themselves assumed to be nonmetric. It might be pointed out, however, that this approach is by far the simplest computationally, and has the desirable property of requiring much less storage than any of the other procedures discussed in this paper. This procedure, then, is particularly suited to small computers.

Due to the complexity of the first procedure and the measurement characteristics of the second, we are inclined to adopt a criterion which is more consistent with the STRESS framework. More precisely, we define a least squares criterion on the squared distances, namely

$$(11) \qquad \phi^2(X, W, D^*) = \sum_i \sum_j \sum_k^{i-1} (d_{ijk}^{*2} - d_{ijk}^2)^2,$$

where d_{ijk}^{*2} is an element of D_i^*, where d_{ijk}^* is defined by (1), and where (11) is subject to suitable normalization conditions. Since (11) is in the STRESS framework, but differs in that it is defined on squared distances d_{ijk}^2 and squared estimates d_{ijk}^{*2}, we refer to the formula as SSTRESS. (Note that d_{ijk}^{*2} is the least squares estimate of d_{ijk}^2, not the square of the least squares estimate of d_{ijk}.) Hayashi [1974] and Obenchain [Note 15] have developed multidimensional scaling procedures within the SSTRESS framework, and Young [Note 18] has discussed the index.

While SSTRESS and STRESS are not strictly equivalent, the monotonic restriction

$$t^o[o_{ijk}^2] = [d_{ijk}^{*2}],$$

defined on o_{ijk}^2 and d_{ijk}^{*2} is precisely equivalent to the monotonic restriction defined on o_{ijk} and d_{ijk}^*. While this precise equivalence also follows with the nominal and ratio levels of measurement, it does not follow with the interval level of measurement, where a linear relationship between o_{ijk} and d_{ijk}^* implies a nonlinear relationship between o_{ijk}^2 and d_{ijk}^{*2}. We will further investigate this inconvenience later on, but at the moment it suffices to say that this difficulty can be surmounted. This allows us to state that the measurement restrictions

$$(12) \qquad t[o_{ijk}] = [d_{ijk}^*],$$

and

$$(12') \qquad t[o_{ijk}^2] = [d_{ijk}^{*2}],$$

are equivalent over the four measurement levels.

We do not mean to imply that SSTRESS is in every way equivalent to STRESS, of course. One important difference is that large values of d_{ijk} and d_{ijk}^* receive more emphasis with SSTRESS than STRESS. A simple example will make this clear. Suppose we have the following two cases:

(A) $\qquad\qquad d_{ijk} = 2, \qquad d_{ijk}^* = 1,$

(B) $\qquad\qquad d_{ijk} = 5, \qquad d_{ijk}^* = 6.$

If we use STRESS the relative contribution of these discrepancies is equal, but if we use SSTRESS we have a ratio of 3 to 11, which is quite different from equality. This effect is more marked when we compare the case

(C) $\qquad\qquad d_{ijk} = 5, \qquad d_{ijk}^* = 4,$

with Case (B). In Case (C) we have squared discrepancies of 9 if evaluated by (11). So even if we have the same d_{ijk} and the difference is equal when

STRESS is used, the direction of the difference differentially contributes to SSTRESS. A simple algebraic manipulation clarifies the point even further. Define

$$d_{ijk}{}^* = d_{ijk} + e_{ijk},$$

where e_{ijk} may be positive or negative. With STRESS the amount that the discrepancy between d_{ijk} and $d_{ijk}{}^*$ contributes is simply $e_{ijk}{}^2$. However, with SSTRESS we have

$$[d_{ijk}{}^2 - d_{ijk}{}^{*2}]^2 = [d_{ijk}{}^2 - (d_{ijk} + e_{ijk})^2]^2$$

$$= e_{ijk}{}^2[e_{ijk} + 2d_{ijk}]^2,$$

so that not only the absolute magnitude of e_{ijk} but also the sign of e_{ijk} (d_{ijk} is always non-negative) and the magnitude of d_{ijk} are related to the overall evaluation of fit. The relation, although algebraically tractable, is not straightforward and not entirely illuminating. We cannot compare the absolute magnitude of fit because the normalization factors in the two formulas may be different.

There is, of course, no *a priori* reason for choosing one or the other of the two formulas. The important point is that the adoption of the SSTRESS formula is perfectly compatible with the measurement level restrictions mentioned above (just as is STRESS), whereas the STRAIN formula is not. Our basic reason for choosing SSTRESS over STRESS is, simply, algorithmic convenience. As you may have noticed, the individual differences weights W (Eq. 1) are linear with respect to the squared distances, but not with respect to the distances themselves. This greatly simplifies the estimation procedure since the least squares estimates of W can be obtained by a series of elementary matrix operations when SSTRESS is adopted as the optimization criterion.

3. The ALSCAL Algorithm

In this section we present in detail an *alternating least squares* algorithm for individual differences *scaling* (ALSCAL).

The alternating least squares (ALS) method is a general approach to parameter estimation which involves subdividing the parameters into several subsets, and then obtaining least squares estimates for one of the parameter subsets under the assumption that all remaining parameters are in fact known constants. The estimation is then alternately repeated for first one subset and then another until all subsets have been so estimated. This entire process is then iterated until convergence (which is assured) is obtained.

With this general definition of ALS one can find its beginnings in the work of Yates [1933] and Horst [Note 11], and follow its development through many researchers, culminating in the NILES/NIPALS work of Wold and

associates [Wold & Lyttkens, 1969]. Generally ALS has been used in the metric situation where one is concerned only with estimation of the model parameters. The extension of ALS to the nonmetric situation in which the procedure is used to estimate data parameters (i.e., to optimally scale the data) as well as model parameters was first made by Torgerson, in the initial configuration routine of the TORSCA algorithm for nonmetric multidimensional scaling [Young & Torgerson, 1967]. Since then ALS has been used by Roskam [Note 16] in the nonmetric principal components situation, Young [1972] for initial values in the polynomial conjoint scaling situation, de Leeuw [Note 4] for the canonical analysis of categorical data, de Leeuw, Young and Takane for nonmetric ANOVA [1976] and Young, de Leeuw and Takane for nonmetric multiple and canonical regression [1976]. The most recent nonmetric results directly motivated the present work, which extends the ALS approach to quadratic models.

The ALSCAL algorithm involves two major phases and two minor phases. The first major phase involves obtaining the least squares estimates of the optimally scaled observations D^* under the assumption that the configuration X and the weights W are constants. That is, we solve the conditional least squares problem which minimizes SSTRESS (11) under the condition that X and W are not variables. Notationally, we indicate this as $\text{MIN}_{D^*}[\phi^2(D^* \mid X, W)]$. The second major phase involves two separate minimization subphases, the first solving the problem $\text{MIN}_W[\phi^2(W \mid X, D^*)]$ and the second the problem $\text{MIN}_X[\phi^2(X \mid W, D^*)]$. The two minor phases are initialization and termination phases. The flow we have chosen is as follows.

0. *Initialization phase*

Compute the initial values of X and W directly from O using a modification of Schönemann's algebraic solution.

1. *Optimal scaling phase*

1.1 Calculate the squared weighted Euclidian distances D using X and W.

1.2 Obtain the optimally scaled (least squares estimated) disparities D^* from the distances D, the observations O, and the relevant measurement restrictions. Use the de Leeuw, Young and Takane [1976] method.

1.3 Normalize appropriately.

2. *Termination phase*

Determine whether the rate of improvement of SSTRESS is sufficiently low to warrant termination. If so, print results and stop. If not, go to the next step.

3. *Model estimation phase*

    3.1 Calculate the new least squares estimates of the weights W from the old X and the new D^* (from step 1.3) by regression techniques.

    3.2 Impose nonnegativity constraints on W, if necessary, by an ALS technique developed here.

    3.3 Calculate the new least squares estimate of the configuration X from the new weights just calculated in steps 3.1 and 3.2 and the D^* computed in step 1.3, by using Gill and Murray's modification of the Newton-Raphson procedure.

    3.4 Return to step 1.1 for another iteration.

Finally, a comment should be made about the ensuing discussion, which is limited to the weighted Euclidian model as applied to symmetric data with no missing elements. These limitations are only made to simplify the discussion. The unweighted Euclidian model may be fit to the data by simply skipping the weight estimation phase (which implicitly fixes the weights to unity). Asymmetric data may be easily handled by changing summation ranges and matrix orders. Missing data may be treated by excluding all missing elements from the optimization criterion, with estimates of the missing data being generated from the model parameters obtained at the conclusion of the analysis. All of these options have been included in the ALSCAL program, and, as will be demonstrated in Section 4, have been extensively evaluated.

Initialization Phase

    The initialization procedure discussed in this section is very similar to the work presented by Schönemann [1972] in which he obtained an algebraic solution to (8) for the error-free ratio measurement level case.

    Let us suppose that there are N scalar product matrices P_i (one for each of the N subjects i) of order n (there are n stimuli) which satisfy

$$(13) \qquad P_i = XW_iX',$$

where the symbols are defined as in (8). (Recall that W_i is a diagonal matrix of weights for subject i, whereas W is a rectangular matrix of weights for all subjects.) The problem is to recover X and W_i from the P_i, under the assumption that X is of full column rank, and that the diagonal elements of W_i are strictly positive for at least one subject. For any nonsingular diagonal matrix T of order t (there are t dimensions), we have

$$(14) \qquad P_i = XT(T^{-1}W_iT^{-1})TX',$$

and consequently must make some restriction on the size of the W_i for identification purposes. Thus, we define

$$D_. = \frac{1}{N} \sum_{i=1}^{N} W_i ,$$

and assume that $D_. = I$, implying that $P_. = X\dot{X}'$ (where $P_.$ is the average P_i). Solutions to this particular equation are determined up to a rotation. We select an arbitrary one of them, for example by using t steps of a Cholesky process or by using the t dominant eigenvalues and vectors of $P_.$. Call this arbitrary solution Y. It follows that

$$X = YK,$$

where K is a rotation matrix. We also know that

$$(X'X)^{-1}X'P_iX(X'X)^{-1} = W_i ,$$

should be diagonal for each i. It follows that we should select our rotation K in such a way that

(15) $$K'(Y'Y)^{-1}Y'P_iY(Y'Y)^{-1}K = W_i ,$$

is diagonal for each i. (Note that $K'K = KK' = I$, and that $K^{-1} = K'$).
      Let

(16) $$C_i = (Y'Y)^{-1}Y'P_iY(Y'Y)^{-1}.$$

It is the case that any linear combination of the N matrices C_i (with different roots) can be used to find the rotation K. Assume that such a linear combination e is possible. We then compute the (unique) set of eigenvectors of

(17) $$C_e = \sum_{i=1}^{N} e_iC_i ,$$

to find K and compute W_i from (15). Thus we have obtained the configuration X, and the weights W. It follows from the assumption we have made that the solution is unique (up to permutations of the dimensions). Note that the assumption that there is a linear combination e is, essentially, equivalent to the assumption that the weights for at least one subject i are all different.

      The preceding developments, which closely follow those presented by Schönemann [1972], are only appropriate to error-free data due to the relationship defined by (13). In the fallible case in which the relationship is only approximately true we need to make two choices. First, we need to define $P_.$, and second, we need to define e. The first problem is quite easily solved by simply double centering the elements of each data matrix O_i with elements o_{ijk}^2 (and dividing by -2) to obtain a matrix P_i of scalar products for each subject. We then average over subjects to obtain $P_.$, which can be decomposed into

(18) $$P_. = YY',$$

to obtain Y, the arbitrarily oriented configuration which best reproduces the averaged scalar products. Note that a) if the data are asymmetric, we average $o_{ijk}{}^2$ and $o_{ikj}{}^2$ for all k and j within each matrix i before double centering; b) if the data contain missing elements, each element is estimated as being equal to the subject's mean judgment; and c) the conditionality of the data is ignored.

The second problem, that of defining the best orientation of the configuration and the associated weights, is solved by obtaining a rotation matrix K which simultaneously diagonalizes the matrices C_i as much as possible. The method suggested by de Leeuw and Pruzansky [Note 7] is used.

Since the procedure just outlined assumes that the data are metric, it is possible to obtain negative weights, especially when the metric assumption is radically violated. (Note that our definition of the weighted Euclidian model includes the requirement that all weights be non-negative.) When negative weights are observed we use the following admittedly arbitrary procedure: We simply add the absolute value of the largest negative weight to all weights, thus ensuring that all weights are non-negative. We then calculate the distances (1) and disparities (as explained in the next section), replace the raw data with the disparities and repeat the procedure outlined above. We are not certain of the theoretical consequences of this procedure although in all cases we have tested the results are satisfactory.

Optimal Scaling Phase

In the optimal scaling phase we wish to optimally scale the squared observations O to obtain the disparities D^* which a) meet the selected measurement restrictions, and b) are least squares estimates of the squared distances D, given the measurement restrictions. We call this the optimal scaling phase because it obtains a scaling of the raw observations that is optimal in the Fisher [1946] sense of optimal scaling. That is, it maximizes the correlation between observations and model while respecting the measurement characteristics of the observations. In this phase we assume that only the optimal scaling variables D^* are free to vary; the stimulus configuration X and the subject weights W are held constant. Thus we solve the conditional least squares problem $\text{MIN}_{D^*}[\phi^2(D^* \mid X, W)]$.

Compute distances. The first step in the optimal scaling phase is to compute the D_i from the current X and W by (1).

Optimal scaling. The second step in the optimal scaling phase is to actually perform the optimal scaling. As we will see, for most of the data types discussed in Section 2 the optimal scaling procedure is quite familiar (linear or monotone regression), although some of the types result in rather novel procedures. However, all of the various types of optimal scaling transformation can be defined as a linear transformation of the squared distances. That is,

(20) $$d_{ijk}^{*2} = l(d_{ijk}^{2}),$$

where l indicates a linear transformation paralleling the measurement restrictions used to define t earlier. Furthermore, l defines d_{ijk}^{*2} so that SSTRESS (11) is minimized for fixed values of W and X, in a least squares sense.

We will not discuss the specific features of these transformations here since a detailed account is already presented in an earlier paper [Young, de Leeuw & Takane, 1976]. Instead, we present a simplified characterization of l using matrix notation. Since we are regressing d_{ijk}^{2} onto o_{ijk}^{2} in the least squares sense under the various measurement restrictions mentioned above, l may be represented by a projection operator of the form

(21) $$l : E = Z(Z'Z)^{-1}Z',$$

where Z is, in general, a matrix of vectors defining the space onto which the vector of d_{ijk}^{2} is regressed.

For the ratio transformation t^r, Z is simply the vector O of squared observations. For the interval transformation t^i, Z reduces to the ratio case after the appropriate additive constant is estimated. In both these cases the least squares estimates may be obtained by well-known regression techniques. In the ordinal and nominal cases Z is defined as a matrix of dummy variables indicating the distances which must be tied to satisfy the measurement restrictions. For the continuous-ordinal transformation t^{co}, the elements to be tied involve order violations, whereas for the discrete-ordinal transformation t^{do}, the elements to be tied also involve observations which are categorically equivalent. Kruskal's least squares monotonic transformation [1964] defines t^{co} when the primary approach to ties is chosen, and defines t^{do} when the secondary approach is used. For the discrete-nominal case the matrix Z indicates that distances which correspond to categorically equivalent observations are to be tied. The obvious least squares estimates in this case simply involve category means. Finally, for the continuous-nominal case the matrix Z indicates those distances which fall outside of the desired interval. In this case the least squares estimates are the interval boundaries for those distances which are in violation, and the distances themselves for those which are not in violation. We use de Leeuw, Young and Takane's [1976] pseudo-ordinal procedure to determine the optimal boundaries.

Note that for some transformations Z is known before the analysis is made, and in other cases it is not. Specifically, for all discrete transformations except the discrete-ordinal transformation Z is known a priori, and for the remainder Z is only known after the analysis is made. Furthermore, in these cases Z varies from iteration to iteration depending on the nature of the distances. To be precise, we should indicate that Z (and therefore E) is a function of d; but the notation $Z(d)$ or $E(d)$ would soon become very cumbersome, and is thus suppressed.

The important thought at this point, however, is that for all four measurement levels, and for both measurement processes, we can represent the optimal scaling as a projection operator of the form shown by (21). This means that if we define a column vector d containing the $Nn(n - 1)/2$ elements $d_{ijk}{}^2$ and another column vector d^* containing the corresponding elements $d_{ijk}{}^{*2}$, then we can make the important observation that

$$(22) \qquad\qquad d^* = Ed.$$

Furthermore, this equation, which is implicitly in terms of unconditional data, can be easily extended to conditional data. For matrix-conditional data we define Z_i for each individual separately and then construct a block-diagonal supermatrix Z with the Z_i's on the diagonal. For row-conditional data we define Z_{ij} for every row of every individual's data matrix and then construct the block-diagonal supermatrix Z with these Z_{ij}'s on the diagonal. In both cases E remains defined as before. Thus the projection operator notion and (22) apply for all three types of conditionality. Note that the various rows or matrices of conditional data may be defined with any mixture of measurement characteristics, as there is nothing requiring them to all be defined identically. Also, any other pattern of conditionality is acceptable.

The chief importance of (22) is that we can now easily express SSTRESS entirely in matrix notation, and entirely in terms of the distances. If we define $\tilde{E} = I - E$, then SSTRESS (11) can be rewritten as

$$(23) \qquad\qquad \phi^2(X, W, D^*) = d'\tilde{E}d.$$

In a parallel manner we can rewrite the normalized SSTRESS formula as

$$(24) \qquad\qquad \phi'^2(X, W, D^*) = d'\tilde{E}d/d'd$$
$$= d'\tilde{E}(d'd)^{-1}\tilde{E}d.$$

Note that in this form SSTRESS involves only the distances and not the disparities, a point which has been discussed at length by Young [1975a].

The final issue to be raised in this section is the procedure for estimating the additive constant when the data are defined at the interval measurement level. (A similar problem has been solved by Messick and Abelson, 1956.) The problem is as follows. When we assume that the observations are defined at the interval level, then

$$(25) \qquad\qquad d_{ijk}{}^* = a(o_{ijk}) + b,$$

for some unknown constants a and b. If we were optimizing STRESS, then the estimation problem would be a simple regression problem involving the distances d_{ijk} and the observations o_{ijk}. However, the situation is complicated by the fact that we are actually optimizing SSTRESS. Instead of the simple linear relationship above, we are actually faced with the quadratic relationship

(26) $$d_{ijk}^{*2} = a^2(o_{ijk})^2 + 2ab(o_{ijk}) + b^2.$$

This is clearly different from the simple regression of d_{ijk}^2 on o_{ijk}^2, which is implied by a linear relationship between d_{ijk}^2 and o_{ijk}^2 (unless $b = 0$ as in the ratio case).

While it is possible to directly solve (26), it is much simpler to redefine the problem as

(27) $$d_{ijk}^{*2} = \alpha + \beta(o_{ijk}) + \gamma(o_{ijk})^2,$$

for which we wish to obtain the best estimates of α, β, and γ, under the constraint that

(28) $$\beta^2 = 4\alpha\gamma.$$

We now introduce three definitions. First, we define the parameter vector $\chi' = [\alpha, \beta, \gamma]$. Second, we define an $N[n(n-1)/2]$ by 3 matrix of second degree polynomials of the observations (unities in Column one, observations in Column two, and squared observations in Column 3). We denote this matrix O (note that this is not the same O as used in other sections of the paper). Finally, we define a column vector d having the $N[n(n-1)/2]$ elements d_{ijk}^2 arranged in the same manner as the o_{ijk} in O.

These definitions allow us to express SSTRESS in the interval measurement situation as

(29) $$\phi^2(\chi, \lambda \mid O, d) = (d - O\chi)'(d - O\chi) + \lambda(\beta^2 - 4\alpha\gamma),$$

which we seek to minimize by solving for χ and λ (the Lagrangian multiplier).

The least squares estimate for the constrainted parameters is

(30) $$\hat{\chi} = (O'O)^{-1}O'd + \lambda q.$$

To solve for the Lagrangian multiplier, we define

(31) $$(O'O)^{-1}g = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix},$$

where $g' = [-2\gamma, \beta, -2\alpha]$ is the derivatives of (28). Then we must solve

(32) $$(\hat{\beta} + \lambda q_2)^2 = 4(\hat{\alpha} + \lambda q_1)(\hat{\gamma} + \lambda q_3).$$

We select the best of the two solutions (i.e., the one which minimizes SSTRESS) by evaluating the set of $\hat{\chi}$ corresponding to each root.

Normalize. The third and final step in the optimal scaling phase is to normalize the solution. There are two separate considerations at this juncture, one concerning the normalization of the parameters (the configuration, weights and optimal scale values) and the other the normalization

of the loss function. While the normalization of the loss function must be performed on every iteration and must be performed in a specific manner in order to avoid certain kinds of degenerate solutions, the normalization of the parameters need only be done prior to printing the final solution. In this way the numbers displayed to the user are in some standard units that enable comparison with other solutions the user may have.

Carroll and Chang [1970] have discussed the relevant issues concerning normalizing the parameters. As they pointed out, two of the three aspects of the problem represented by (1) and (11) (the data, the weights, and the configuration) must be normalized, with the remaining aspect being left unnormalized. While the choice and the actual details of the normalization are arbitrary, we choose to continue the conventions adopted by Carroll and Chang. Specifically, the configuration is normalized so that the mean projection on each dimension is zero and the variance of the projections on each dimension is unity. However, whereas Carroll and Chang normalize the data, we must normalize the optimally scaled data. After all, we cannot normalize qualitative data, whereas we can normalize the optimally scaled data even when the data themselves are qualitative, since the optimally scaled data are always quantitative. It is interesting to note that there is a subtle reason which disallows normalizing the distances D instead of the optimally scaled data D^*, even though it has been argued by Kruskal and Carroll [1969] and Young [1972] that the choice is arbitrary. While the arguments of Kruskal and Carroll and Young apply for a single matrix of unconditional data, or even for several matrices of either matrix conditional or unconditional data, they do not apply to row conditional data. The problem is that the distances for a particular subject are all jointly determined up to a single multiplicative constant. They are all defined on a single measurement scale at the ratio level of measurement. Thus, in the case of row conditional data, it is not possible to adjust the distances in one row of the matrix by one multiplicative constant, and the distances in another row by another constant. Since we are not allowed to apply different multiplicative transformations to each row of the distance matrix, we are forced (at least in the row conditional case) to normalize the optimally scaled data (which may legitimately be subjected to different transformations as discussed in Section 2). Thus, for each partition (row, matrix, or whatever) of optimally scaled data, we normalize so that the sum of squares of the squared optimally scaled data is a constant.

There is one final consideration in the normalization of the parameters. One of the conventions adopted by Carroll and Chang [1970] was to normalize the "pseudo" scalar products so that their sums of squares was constant. They chose to do this (instead of normalizing the sums of squares of the raw data) for a number of reasons, one of which is highly relevant here. They showed that if one were to set the total sums of squares of an individual's

pseudoscalar products equal to one, then the individual's sum of squared weights could be interpreted as indicating, roughly, the proportion of variance in the pseudoscalar products which is accounted for by the model. (The word "roughly" can be eliminated from the previous sentence when it is true that $X'X = I$, which is not usually the case.) There is, unfortunately, no corresponding procedure which allows for the sum of squared weights to be interpreted as indicating the proportion of variance accounted for in the optimally scaled data. Thus, when the analysis is completed, we perform one final normalization which involves computing the pseudoscalar products from the optimally scaled data, and then setting each individual's sum of squared pseudoscalar products equal to unity. This allows for the interpretation that an individual's sum of squared weights roughly indicates the proportion of variance in his pseudoscalar products which is accounted for by the solution. This is the same interpretation as afforded by Carroll and Chang's procedure. Note, however, that the conditionality of the data and the resulting differences in normalization have certain implications for interpreting the weights. These implications are discussed in Section 5.

We now turn to the second normalization consideration, that of normalizing the loss function. Although we actually perform the optimization relative to the unnormalized loss function stated in (11), we can indirectly optimize a normalized function if suitable steps are taken, as has been discussed by de Leeuw, Young and Takane [1976]. This characteristic is very convenient, since we do not have to deal directly with the normalized function (which is the ratio of two quadratic forms) whose partial derivatives are considerably more complicated than those of the unnormalized function. We only gain this simplicity, however, if we normalize the function relative to the optimally scaled data. No gain is made if we normalize relative to the distances. This is important to note since normalization is commonly relative to the distances, and since the arguments of Kruskal and Carroll [1969] and Young [1972] would again lead one to suspect that the choice is arbitrary. The reasoning given in the previous paragraph which led us to conclude that the choice is not arbitrary is the same reasoning which leads to the same conclusion here.

Thus, we normalize the loss function relative to the optimally scaled data D^*. There are two more considerations, however. One is a point emphasized by Kruskal and Carroll [1969] and Roskam [Note 16] which leads to the conclusion that the normalization must be within partitions (to use the terminology of Young, 1973). That is, for unconditional data, where all of the data form a single partition, we compute a single normalized SSTRESS:

$$\phi_u^2 = \frac{\sum_i \sum_j \sum_k (d_{ijk}^{*2} - d_{ijk}^2)^2}{\sum_i \sum_j \sum_k d_{ijk}^{*4}} ,$$

whereas for matrix conditional data, where each matrix forms a separate partition, we must compute a normalized SSTRESS value for each matrix, and then obtain their average:

$$\phi_m^2 = \frac{1}{N} \sum_k^N \left[\frac{\sum_i \sum_j (d_{ijk}^{*2} - d_{ijk}^2)^2}{\sum_i \sum_j d_{ijk}^{*4}} \right].$$

Finally, for row conditional data, where each row is a partition, we compute a normalized SSTRESS for each row and then average:

$$\phi_r^2 = \frac{1}{Nn} \sum_k^N \sum_i^n \left[\frac{\sum_j (d_{ijk}^{*2} - d_{ijk}^2)^2}{\sum_j d_{ijk}^{*4}} \right].$$

(Note that we use the sum of squares of the *squared* optimally scaled data, since the loss is defined on squared distances and squared optimally scaled data. This is the reason for the appearance of the fourth power in the denominator.) We can summarize the above formulas in a single formula by introducing a new symbol δ_{ijkl} which has a value of 1 if d_{ijk}^* is in partition l and a value of zero otherwise. We can then write the above formulas as

$$(33) \quad \phi'^2 = \sum_l \left[\frac{1}{\sum_i \sum_j \sum_k \delta_{ijkl}} \right] \left[\frac{\sum_i \sum_j \sum_k (d_{ijk}^{*2} - d_{ijk}^2)^2 \delta_{ijkl}}{\sum_i \sum_j \sum_k d_{ijk}^{*4} \delta_{ijkl}} \right],$$

which is the complete expression for the normalized SSTRESS function being optimized by ALSCAL.

The remaining point to be considered is whether we should actually normalize by the sum of squares of the squared optimally scaled data, as in (33), or by the variance of the squared optimally scaled data. This question, which has also been addressed by Kruskal and Carroll [1969] and Roskam [Note 16], is probably best answered by saying that for row conditional data we should probably use the variance, whereas for the other types of data we should use the sums of squares. In this paper, however, we use only (33) given above. We do plan on incorporating both types of formulas into the ALSCAL program, however.

Finally, if we are solving the unnormalized problem represented by (11) (as we are), and we actually wish to optimize the normalized problem represented by (33) (as we do), then it can be shown that all we have to do is solve the unnormalized problem and adjust the length of the vector of squared optimally scaled data by multiplying all of its elements by the ratio of the sums of squares of the squared distances to the sums of cross products of the squared distances and squared optimally scaled observations:

$$(34) \qquad \frac{\sum_i \sum_j \sum_k d_{ijk}^4 \delta_{ijk}}{\sum_i \sum_j \sum_k d_{ijk}^{*2} d_{ijk}^2 \delta_{ijk}}.$$

Termination Phase

The termination phase is extremely simple. We must only determine the value of SSTRESS on the current iteration (11) and compare this value with the previously determined value. If the amount of improvement is less than some arbitrary criterion, then we terminate; if not we continue. The simplicity of this phase is due to one of the characteristics of an ALS procedure, namely that an ALS iteration never worsens the value of SSTRESS (a proof of this characteristic may be found in de Leeuw, Young and Takane, 1976).

Model Estimation Phase

In the model estimation phase we solve two conditional least squares problems successively. The first subphase solves the conditional least squares problem $MIN_W[\phi^2(W \mid X, D^*)]$, whereas the second subphase solves the problem $MIN_X[\phi^2(X \mid W, D^*)]$. In this section we discuss both of these problems.

Compute weights. To estimate W we obtain the partial derivatives of (11) with respect to the elements of W and set the derivatives to zero. This system of homogeneous equations is then solved with respect to W. To simplify the derivation we define an order $n(n-1)/2$ by t matrix Y, where the columns of Y contain all interpoint distances as projected onto each dimension (i.e., each element of column a of Y is $(x_{ia} - x_{ja})^2$, the dimension-wise squared difference between stimuli i and j). We also define an order N by $n(n-1)/2$ matrix D^*, whose rows contain the $n(n-1)/2$ optimally scaled observations for each individual, with the elements arranged to correspond with Y. (This D^* contains the same information as the D^* used in earlier parts of this paper, but organized differently. In this section we refer to this organization of the information when we use the symbol D^*.) These definitions allow us to write SSTRESS as

$$(35) \qquad \phi^2(Y, W, D^*) = tr(D^* - WY')'(D^* - WY'),$$

from which we see that the least squares estimates of W are

$$(36) \qquad W = D^*Y(Y'Y)^{-1}.$$

Nonnegativity weight constraint. There is one difficulty in using the regression approach just outlined for obtaining W: The non-negativity constraints placed on the weights (1) may be violated. Thus we now discuss a method of incorporating this constraint (or any other linear inequality constraint) which is strictly within the ALS framework.

An observation basic to the procedure to be presented is that the estimation process presented in (36) is independent for each individual. That is, the values estimated for the weights for one individual do not affect the estimated values of the weights for any other individual. This can be seen from the fact that SSTRESS (35) can be decomposed into a summation of

separate components, each of which is a function of only a single subject. Since the weights for one subject are independent of those for the others, we can impose a non-negativity constraint on subjects with negative weights without having to modify the weights for other subjects. Note, however, that the weights for a given subject are not independent of each other, which means that we cannot simply set a subject's negative weights to zero and leave his positive weights unchanged. If we do this we destroy the least squares properties of the weight estimates.

Our solution to this problem is as follows. First, we obtain the unconstrained least squares estimates of W by (36). We use these estimates for those subjects with non-negative weight vectors. For the other subjects we set one of the negative weights to zero, which is the constrained least squares estimate under the condition that all the other weights are constant. Then we re-estimate the value of another weight under the assumption that all other weights are constant. The conditional least squares estimate for a single weight is

$$(37) \qquad w_{ia} = (d_i^* - \sum_{b \neq a} w_{ib} y_b)' y_a / (y_a' y_a),$$

where y_a is the a'th column of matrix Y (35) which contains the squares of the interpoint distances as projected onto the a'th dimension. If this unconstrained conditional least squares estimate is negative, we set it to zero. We then repeat this process for each dimension until all weights for the subject are non-negative.

Compute coordinates. The second subphase of the model estimation phase is to determine the stimulus coordinates X. This subphase is somewhat more complicated than the weight estimation subphase, since the partial derivatives of SSTRESS with respect to the elements of X are not linear in the x_{ia}'s. Rather, SSTRESS is quartic in the x_{ia}'s, so the derivatives are a system of cubic equations. There are several ways of solving such a system. We first review some of the possibilities, and then present the method we have adopted.

Perhaps the most elegant solution, at least for a theoretical point of view, would be to analytically solve the system of $m = n^*t$ simultaneous cubic equations for the m unknowns. This has been suggested by Oberchain [Note 15]. It is possible to do this by either Euclid's or Kronecker's elimination method [Bôcher, 1907], in which the system of m simultaneous polynomial equations is eventually reduced to a single polynomial equation in one unknown and $m - 1$ linear simultaneous equations in $m - 1$ unknowns. The problem is then reduced to finding the numerical solutions to a simple polynomial equation and, after substitution of the solution into the remaining linear equations, finding the solution to a system of $m - 1$ linear equations [Wilf, 1960]. The method is particularly favorable in our situation since we have only to solve cubic equations, and there is an analytic solution for a cubic

equation with one unknown. While this approach has theoretical beauty, it is impractical due to the number of equations in our case (as many as 500 or 600).

The opposite extreme is to solve for only a single coordinate x_{ia} at a time, with a total of m such solutions on each iteration. That is, we could use the analytic solution to a cubic equation with one unknown to obtain the conditional least squares estimates for a single coordinate, under the assumption that all other coordinates (and of course all the W and D^*) are fixed. The previous estimate for this coordinate is then immediately replaced with the new estimate. Note that after m such estimations we have obtained new estimates for all of the coordinates, but that these are not the same as those obtained by the simultaneous method discussed in the previous paragraph. This is the case despite the fact that the two procedures will eventually converge on the same estimates. For any given iteration the simultaneous method achieves the most improvement in fit, but takes the most time to do it.

Of course we are not limited to only these two choices; the quickest method probably lies somewhere in between the two extremes. That is, it may be best to estimate a block of x_{ia}'s simultaneously, making sure that the number of coordinates being simultaneously estimated is not so large that it slows down the entire process to the point where it cancels the benefits derived from simultaneous estimation.

Optimizing the efficiency of our algorithm is a difficult yet crucial problem. After several trials and errors we have found a method which appears to be more efficient than any other currently available algorithm (some sketchy evidence on this point will be presented later). We apply a modified Newton-Raphson method to obtain a new set of conditional least squares estimates for all of the coordinates of a single point simultaneously, successively solving for each point in turn. Thus we estimate x_{ia} $(a = 1, \cdots, t)$ simultaneously for a specific j and successively for each stimulus j $(j = 1, \cdots, n)$. This is the approach taken by Yates [1972].

The Newton-Raphson method is well-known, of course, but our application of it is unique. We use it to obtain conditional least squares estimates which solve the problem $\mathrm{MIN}_{x_j} [\phi(x_j \mid x_k, W, D^*), (j \neq k)]$. Thus our approach is to place the Newton-Raphson method within the ALS framework to solve a system of cubic equations in t unknowns. This demonstrates again the flexibility of the ALS approach. The use of Newton-Raphson in conjunction with ALS is particularly attractive in the present context because the function being optimized is smooth and the evaluation of the function requires very few computations. Thus the approach should be quick and robust, as indeed it is.

We actually use a recent modification of the Newton-Raphson procedure developed by Gill and Murray [Note 8] which ensures the positive

definiteness of the Hessian at the current point. This guarantees that we are proceeding in a downhill direction. Since the Hessian is always positive semi-definite at a minimum, it is desirable to ensure that it is so during the entire estimation process. If it is not "sufficiently" so (in a complex sense discussed at length by Gill and Murray), deliberately chosen values are added to the diagonal to force it to be positive definite. This avoides convergence to a maximum or to some other stationary point that is not a minimum.

We now provide the first and second derivatives of SSTRESS with respect to x_{ia} for fixed j and $a = 1, \cdots, t$. To simplify the derivation, we note that

$$(38) \quad d_{ijk}^{*2} - d_{ijk}^2 = d_{ijk}^{*2} - \sum_a^t w_{ia}(x_{ia} - x_{ka})^2,$$

$$= d_{ijk}^{*2} - \sum_a^t w_{ia}x_{ia}^2 + 2 \sum_a^t w_{ia}x_{ia}x_{ka} - \sum_a^t w_{ia}x_{ka}^2,$$

and we introduce several definitions. First, we collect the terms which do not involve x_{ia} (the fixed terms) and define them to be

$$(39) \qquad\qquad h_{ijk} = d_{ijk}^{*2} - \sum_a^t w_{ia}x_{ka}^2.$$

We organize these terms into a vector h_i which contains all h_{ijk} for fixed j and for $k \neq j$ (this vector has $N(n - 1)$ elements). We also define a super-matrix $G = [G_1, G_2]$, with $N(n - 1)$ rows and $2t$ columns. The two sub-matrices are defined as follows:

$$(40) \qquad G_1 = -2 \begin{bmatrix} w_{11}x_{11} & \cdots & w_{1t}x_{1t} \\ \cdot & & \cdot \\ w_{11}x_{k1} & \cdots & w_{1t}x_{kt} \\ \cdot & & \cdot \\ w_{11}x_{n1} & \cdots & w_{1t}x_{nt} \\ \vdots & & \vdots \\ w_{N1}x_{n1} & \cdots & w_{Nt}x_{nt} \end{bmatrix}, \qquad k \neq j,$$

and

$$(41) \qquad G_2 = \begin{bmatrix} w_{11}u & \cdots & w_{1t}u \\ \vdots & & \vdots \\ w_{N1}u & \cdots & w_{Nt}u \end{bmatrix},$$

where u is an $n - 1$ component column vector of unities. We also define a $2t$ component supervector α_j, consisting of a vector x_j of coordinates of

stimulus j on t dimensions and a vector whose elements are the squares of the elements in x_j . Since it is possible to express the squared elements as the product of a diagonal matrix and a vector, we further define X_j to be an order t diagonal matrix with the coordinates of stimulus j on its diagonal (do not confuse this with the entire matrix of coordinates denoted X). Then

$$(42) \qquad \alpha_j = \begin{bmatrix} x_j \\ X_j x_j \end{bmatrix}.$$

We can now define SSTRESS as

$$(43) \qquad \phi^2(X, W, D^*) = \frac{1}{2} \sum_i^n (h_i - G\alpha_i)'(h_i - G\alpha_i).$$

(The 1/2 is present since the summation is over all $N \cdot n^2$ elements d_{ijk}^{*2}, whereas in previous definitions of SSTRESS the summation was over only the lower triangular portion of each matrix.)

The gradient vector (first derivatives of SSTRESS with respect to x_j) can be expressed as

$$(44) \qquad g = -[I, 2X_j]G'h_i + [I, 2X_j]G'G\alpha_i$$

$$= [G_1'G_1 + 2X_jG_2'G_1 + G_1'G_2X_j + 2X_jG_2'G_2X_j]x_j$$

$$- [G_1h_i + 2X_jG_2h_i].$$

The off-diagonal elements of the Hessian (matrix of second order partial derivatives) are (for $a \neq b$)

$$(45) \qquad h_{ab} = [e_a', 2x_{ja}e_a']G'G \begin{bmatrix} e_b \\ x_{jb}e_b \end{bmatrix},$$

where e_a is a vector with unity in the a'th position and zeros elsewhere. The a'th diagonal element of the Hessian is

$$(46) \quad h_{aa} = -[o, 2e_a']G'h_i + [e_a', 2x_{ja}e_a']G'G \begin{bmatrix} e_a \\ 2x_{ja}e_a \end{bmatrix} + [o', 2e']G'G \begin{bmatrix} x_j \\ X_j x_j \end{bmatrix}.$$

We use the gradient and Hessian with Gill and Murray's [1974] procedure for the Newton-Raphson method. With this procedure one obtains the l'th estimate of x_j , which we denote $x_j^{(l)}$, according to

$$(47) \qquad x_j^{(l)} = x_j^{(l-1)} - \theta \tilde{H}^{(l-1)-1} g^{(l-1)},$$

where θ is a stepsize determined to ensure that $\phi^{2(l)} < \phi^{2(l-1)}$, where $\tilde{H} = H$ when H is positive definite, and where $\tilde{H} = H + F$ for F, a diagonal matrix with positive diagonal values when H is not positive definite. The matrix F

is determined according to Gill and Murray's developments. While it is the case that SSTRESS must be evaluated several times in determining the estimate of x, each point's coordinate vector, it is a very simple and quick evaluation since the optimal scaling D^* is fixed during the evaluation. Thus, we do not have to perform this time-consuming operation, which is one of the nice features of the ALS approach. If we were using the more standard gradient approach we would have to perform the optimal scaling for each evaluation of SSTRESS, and the algorithm would be very slow. (This may account for the inefficiency of Yates' [Note 17] procedure which performs the optimal scaling after each point's coordinates are estimated.)

Once we have minimized SSTRESS relative to a single point, we repeat the procedure for another point, until all points have been subjected to the process. This defines a single iteration. The entire process is repeated until convergence is obtained. Note that once a point's coordinates have been estimated, the old coordinates are immediately discarded and the new estimates are inserted before the next point's coordinates are estimated. This prompt replacement is mandatory since each suboptimization is not independent from the others.

There is one minor theoretical problem with the procedure just proposed. The function being minimized (43) is a quartic function; therefore its gradient (44) is a system of cubic equations. This system of cubic equations has, at most, 2^t minima, of which not more than $2^t - 1$ are local minima, and 1 is the global minimum (these assertions will be supported in Section 5). Thus, in one dimension there may be one local minimum in addition to the global minimum, in two dimensions there may be as many as three local minima, in three dimensions up to seven local minima, etc. The procedure we have proposed converges on one of the minima without ensuring that it is the global minimum. While numerical analysis results indicate that we will most often converge on the optimal minimum (especially if we have a good initial estimate, as we do, and if the optimal minimum has a rather better value than the local minima, a situation which may or may not exist here), we will at least occasionally converge on non-optimal minima. An alternative procedure which is free of this problem will be proposed in Section 5. Regardless, we are of the opinion that this theoretical difficulty with the proposed procedure will have little practical effect, an opinion supported by the results presented in the next section. When we recall that the present part of the estimate process seeks the optimal location of a single point, we see that there are many self-correcting opportunities built into the overall estimation process. This may be the reason that the theoretical difficulty has little practical effect. Finally, we note that with previous nonmetric procedures the local minimum problem was so complicated that it defied analysis. We believe that our procedure, while not totally free from local minimum problems, is clearly superior to previously proposed procedures in this regard.

4. Examples

In this section we present examples of the use of ALSCAL to demonstrate its efficacy. The first examples utilize the weighted Euclidian model, and the last the unweighted model. For the weighted model we first perform a small Monte Carlo study which allows us to compare the structures obtained by ALSCAL with the true structures which were used to generate the artificial data. We further evaluate the performance of ALSCAL in the weighted Euclidian case by comparing the structures obtained by ALSCAL with those obtained by INDSCAL when both are used to analyze the same real (not artificial) data. For the unweighted model we evaluate ALSCAL by comparing the structures it obtains for sets of real data with those obtained by other investigators using the standard MDS algorithms for applying the unweighted model. Finally, we evaluate the ability of ALSCAL to analyze nominal data by comparing the structure obtained from a set of data which has been previously analyzed under the assumption that the measurement level is ordinal. It is not possible to compare these results with other algorithms designed to multidimensionally scale nominal data since no such algorithms have been proposed previously.

We believe that the reader will conclude, from the evaluations outlined in the previous paragraph, that ALSCAL is very robust in all the situations for which it was designed.

Monte Carlo Study

The general outline of the Monte Carlo study is as follows. First, we generate an arbitrary "true" configuration and "true" weights, which together we call the "true" structure. We then determine the dissimilarities by computing distances (according to the weighted Euclidian distance formula) and introducing either random or systematic error, or both. We then submit these errorful dissimilarities to ALSCAL to obtain the "derived" structure (stimulus configuration and weights). Finally, we compare the derived structure with the true structure in order to evaluate how robust ALSCAL is to random and systematic error.

Actually, the purpose of the experiment is twofold. First, it should be the case that analysis of dissimilarities which contain no random error but which are systematically distorted monotonically should, if we assume that the data are ordinal, produce a derived structure which is identical to the true structure no matter how severely we distort the true distances. Furthermore it is anticipated that if we analyze these same systematically distorted distances while inappropriately assuming that the data are interval, then a systematic bias should be found in the derived structure. Of course, the degree of bias should be a function of the degree of distortion.

The second purpose of the Monte Carlo study is to determine the robustness of ALSCAL in the face of random error. Ideally, ALSCAL should be able to recover the true structure when there is a moderate degree of random error no matter what measurement assumptions we make about the data (at least when there is not much systematic error). Note that this point relates not only to the ALSCAL algorithm, but also to the weighted Euclidian model itself. To the authors' knowledge there has been no Monte Carlo study which evaluated the effect of error (either random or systematic) on the recovery of the true structure, and which attempted to evaluate the goodness of recovery to such aspects of the model as the number of points or subjects, the number of true and recovered dimensions, the amount of error, etc. (Note that Jones and Waddington [Note 13] have investigated the effect of subjects who use only a subset of the dimensions.) Our study is by no means a complete or exhaustive study of these variables. Nonetheless, we believe that such a study needs to be done and that ours may be viewed as a precursor to such comprehensive studies.

We hypothesized the "true" structure shown in Tables 1-a and 1-b. We chose a small two-dimensional structure for ease of presentation, with the actual numbers arbitrarily assigned.

We emphasize that our results are not independent of this particular structure, particularly with respect to the number of stimuli (which is rather small compared to most empirical studies using this model), the number of subjects (which is also on the small side), the number of dimensions, and the actual structure. The configuration of stimuli is shown in Figure 1 by the black circular dots (the lines connecting the dots differentiate the true configuration from several other configurations also presented in this figure). The "true" subject weights are given by the black circular dots in Figure 2. Note that these weights, which are equally spaced along a straight line, indicate that the subjects are "moderately" heterogeneous in terms of their relative weighting of the dimensions. This situation, in our experience and in the experience of Carroll (personal communication), is optimal for obtaining a robust and meaningful analysis with INDSCAL. Note also that subjects generally attach relatively more importance to Dimension I than to Dimension II.

Weighted Euclidian distances were calculated from these stimulus coordinates and individual weights. While computing these distances random error was introduced. It is debatable when and where the error component should be added (i.e., to the distances, to the coordinates, or to the weights; before or after the systematic monotonic distortion; etc.). We arbitrarily chose to follow the procedure of Young [1970] in which independent random normal error is added to the stimulus coordinates. Such error is generated anew for each pair of stimuli. Thus $d_{ijk}{}^2$, under the lth degree of error perturbation, is generated by

Table 1-a Hypothesized stimulus configuration

Stimulus	Dimension I	Dimension II
1	1.37198	1.36082
2	0.77174	1.36082
3	0.77174	-1.49691
4	-1.02899	0.40824
5	-1.62923	-0.54433
6	-0.42874	-0.54433
7	0.17149	-0.54433

Table 1-b Hypothesized weight configuration

Subject	Dimension I	Dimension II
1	0.40917	0.01805
2	0.36371	0.03610
3	0.31824	0.05415
4	0.27278	0.07220
5	0.22731	0.09025
6	0.18185	0.10831
7	0.13639	0.12636
8	0.09092	0.14441
9	0.04546	0.16246

Table 1-c Values of $\gamma^{(\ell)}$

Error Level ℓ	Dimension I	II
1	0	0
2	.180	.286
3	.600	.953

$$(48) \qquad d_{ijk}^{(l)} = \left[\sum_{a=1}^{t} w_{ka}(x_{ia} - x_{ja} + (z_{ijka})\gamma_a^{(l)})^2 \right]^{1/2},$$

where $z_{ijka} = z_{ija} - z_{ika}$, where $z_{ija} \sim N(0, 1)(i, j = 1, \cdots, 7)$, $(a = 1, 2)$ and where $\gamma_a^{(l)}$ is a parameter specifying the variability of the errors. Note that $d_{ijk}^{(l)2}$ does not follow the noncentral chi-squares distribution (as it does in Young, 1970) since the variability is different across dimensions ($\gamma_a^{(l)}$ depends on dimensions and moreover, dimensions are differentially weighted). Note also that the same z_{ijka}'s are used for different error levels. The values of $\gamma_a^{(l)}$ are shown in Table 1-c. Since z_{ija} and z_{ika} are independent, the variance
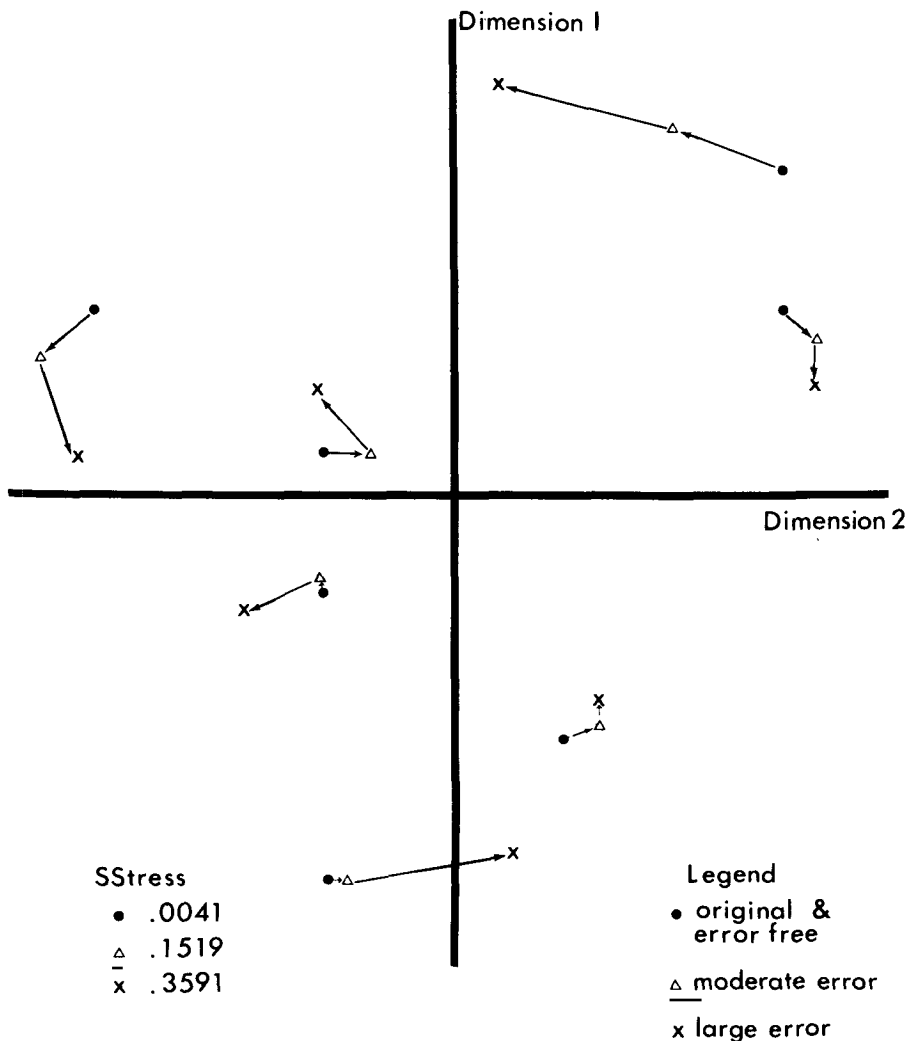
FIGURE 1
Monte Carlo study: Effects on the stimulus configuration when the data are assumed
to be ordinal.

of $(z_{ijka})\gamma_a^{(l)}$ is $2(\gamma_a^{(l)})$. Note that the stimulus configuration x_{ia}'s are standard-
ized so that they have unit variances for both dimensions. We refer to the
case when $l = 1$ as the error free case, $l = 2$ as the moderate error case, and
$l = 3$ as the large error case. (For $l = 3$ the error variance is .720 for Dimension
I and 1.816 for Dimension II, which is much larger than the error variance
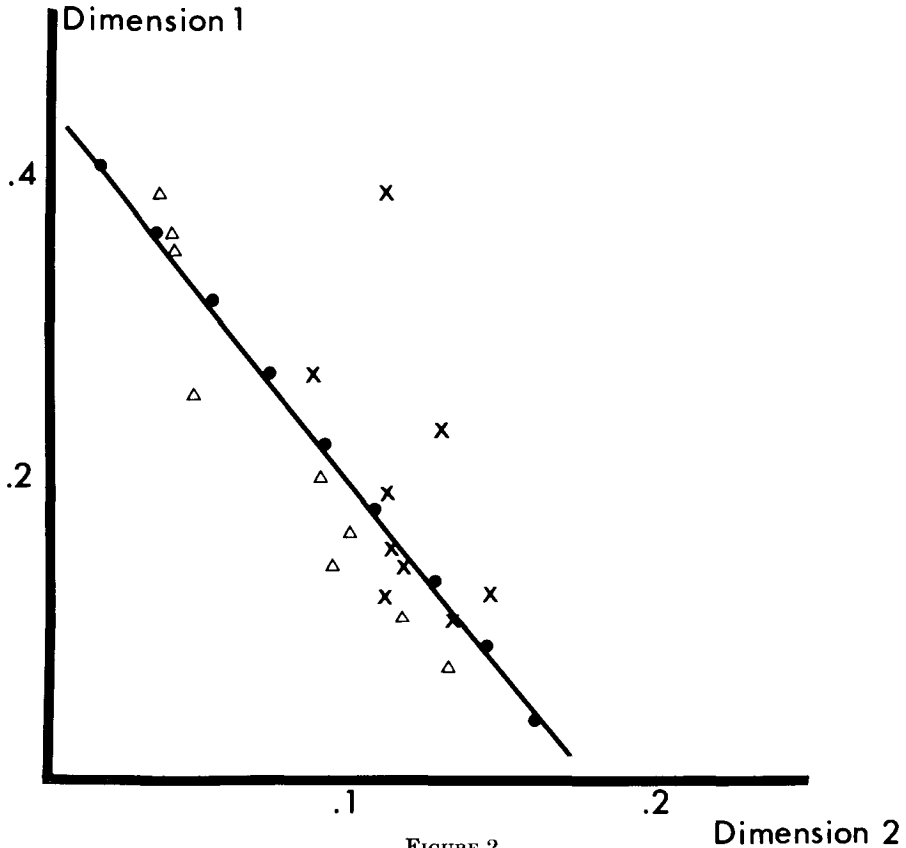used in most Monte Carlo studies.)

FIGURE 2
Monte Carlo study: Effects on the weight space when the data are assumed to be ordinal.

Next we introduced systematic monotonic error by either squaring the randomly perturbed distances in (45), or by raising these distances to the fourth power. Thus we have three levels of systematic error: No distortion (the error perturbed distances themselves), moderate distortion (the squared perturbed distances), and high distortion (the perturbed distances raised to the fourth power).

Finally, these systematically and randomly distorted distances served as the dissimilarities input to ALSCAL for analysis. The derived structures are displayed in Figures 1 (the stimulus configuration) and 2 (the weights). First of all, the algorithm perfectly recovered the true structure form the error-free dissimilarities. The structure, which is indistinguishable from the true structure, is presented in these figures as the black circular dots. The structures resulting from the moderate and high degrees of systematic (monotonic) distortion when there was no random error in the data are also

indistinguishable from the true structure when the assumption is (correctly) made that the data are measured at the ordinal level. Thus the dots in Figures 1 and 2 represent four structures: The true structure and the structures derived by ALSCAL for three levels of monotonic distortion when there is no random error in the data and when the data are assumed to be ordinal. We will discuss what happens when these data are assumed to be metric in a moment.

Figures 1 and 2 also display the structures derived by ALSCAL when there is moderate random error (the triangles) and when there is large random error (the squares). Note that there is, once again, no discernible effect for systematic distortion when the data are assumed to be ordinal, with all three levels producing identical structures. The effect of random error shows up in these figures in a very interesting and somewhat surprising way. As the level of error increases, the actual structure of the stimulus configuration (as evidenced by the interpoint distances) is relatively unaffected. However, the entire configuration changes from the true orientation towards an orientation which is more nearly like the principal components fo the group space (i.e., the variance on the first dimension is increasing and that on the second decreasing, a change which is reflected in the overall magnitude of the weights). This effect is most pronounced for the highest amount of error. However, considering in particular that the same z_{ijka} is added across different error levels, we refrain from definitive comments at this stage of investigation.

The weights, on the other hand, simply show a nonsystematic deterioration as the amount of error increases. Although the relatively heavier weighting on Dimension I is preserved, the order of individual subjects along the dimensions of the weight space is destroyed, let alone the ratio of an individual's weights to each other. Note also that the weights on the second dimension (which suffers from relatively more random error) tend toward their mean as the error increases. These results appear to the authors to be very provocative and worthy of systematic study. However, since the main intent of this paper is not to perform a systematic investigation, we will not dwell on the matter any further, although we will examine a possible cause in the discussion section. Finally, let us emphasize that these results are identical for all levels of systematic monotonic distortion when the data are assumed to be ordinal. This shows that the theoretical invariance of the results over monotonic distortion is also an empirical invariance.

This is not to say that systematic monotonic distortion has no effect when we (incorrectly) assume that the data are metric. The fact that does can be seen in Figures 3 and 4. These figures show the effects of assuming that the data are ratio when there is systematic monotonic distortion. The results are shown separately for each level of random error since there is a substantial interaction between the effect of systematic and random error in this case. Thus we have Figures 3a, 3b, and 3c for the stimulus configurations obtained

by ALSCAL for the three levels of random error, and Figures 4a, 4b, and 4c for the corresponding subject weights.

Figures 3a and 4a present the results from data with no random error. In these figures there are three points plotted for each stimulus and individual, one for the "true" and "no distortion" configurations (which are identical), one for the "moderate distortion" configuration, and one for the "high distortion" configuration. The effect of monotonic distortion of the data is small (though obvious) upon the stimulus configuration (Figure 3a); the general configural relations among the stimuli remain intact (though modified). When we recall that there was no effect of systematic monotonic distortion when the ordinal measurement assumption was made, and when we compare those results with the present results, we see that appropriate measurement assumptions can in fact improve the descriptive quality of the weighted Euclidian model. Note that the effect of systematic error on the configuration is random (there is no discernible pattern of point displacement). There is, however, a systematic effect of systematic error, but it is now contained in the weight space (Figure 4a). There seem to be two general tendencies. First, as the distortion increases the weights tend to show less variance on Dimension II. Second, as distortion increases the configuration of weights becomes slightly concave upward (in contrast to the true linear, equal-spaced weight configuration). We find it very difficult to rationalize these effects.

We now turn to the worst possible case, that involving systematic monotonic error when the wrong measurement level is assumed and when there is random error as well. The results are presented in Figures 3b and 3c (stimuli) and 4b and 4c (weights). Each of these figures contains four plotting symbols for each stimulus (or weight), one for the true value and one for the observed values under the three levels of systematic error (the "no distortion" and "true" values no longer coincide due to the presence of random error). As opposed to previous results there seems to be very little systematic effect of both kinds of error combined, except to say that increasing error yields further deterioration of both stimulus and weights spaces. It appears to be the case (though we may be stretching it a bit) that the effects are more pronounced on Dimension II than on Dimension I. Specifically, the variance of a point's projection on Dimension II is larger than on Dimension I in Figures 3b and 3c (and even perhaps in 3a), which indicates that a point is more poorly determined on Dimension II and on Dimension I. Correspondingly, we see in the weight space that the variance of weights decreases faster on Dimension II than on Dimension I as error increases. This suggests that our hypothetical subjects are becoming less differentiated by Dimension II more quickly than by Dimension I.

This small and admittedly very incomplete Monte Carlo study tells us several important things. First, ALSCAL recovers a known configuration when there is no error, for ordinal measurement assumptions as well as
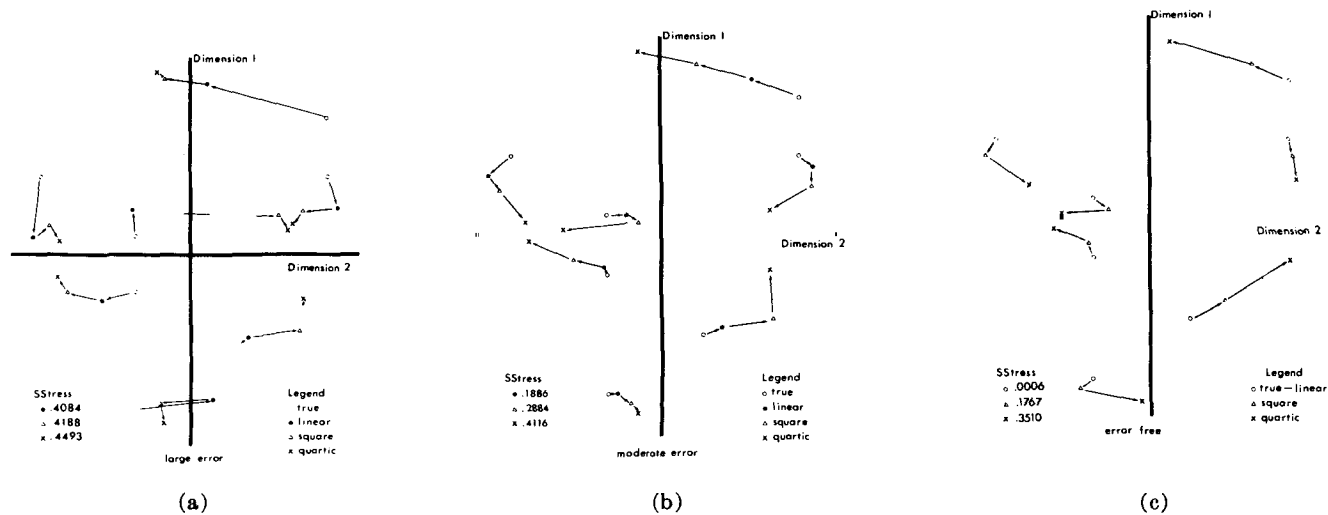
(a)

(b)

(c)

FIGURE 3

Monte Carlo study: Effects on the stimulus configuration when the data are assumed to be ratio.
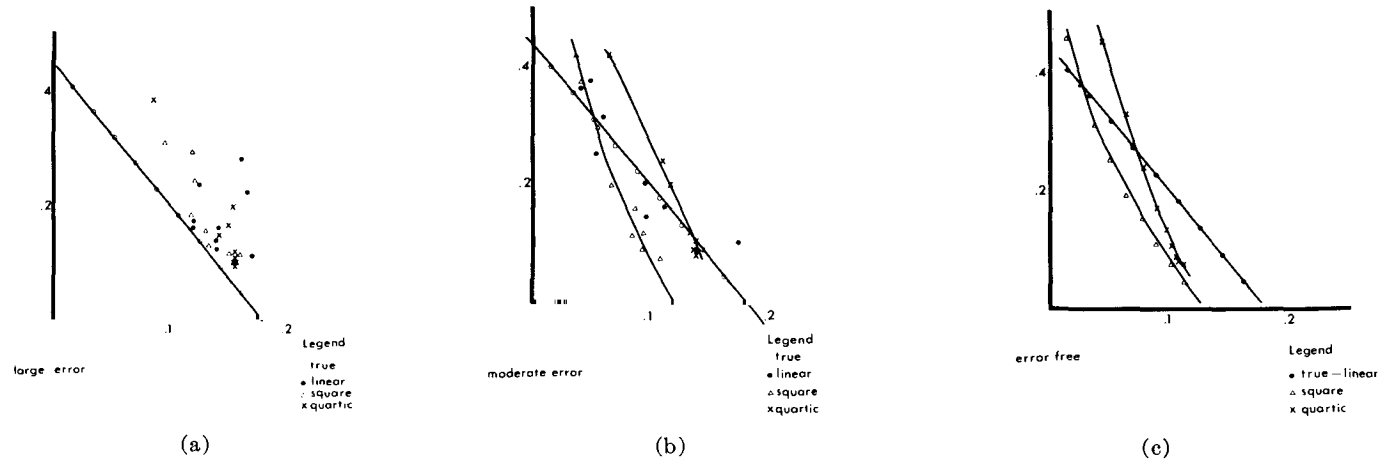
(a) (b) (c)

FIGURE 4
Monte Carlo study: Effects on the weight space when the data are assumed to be ratio.

interval. Second, ALSCAL is robust in the face of monotonic transformations of ordinal data. Third, the recovery of the structure of the stimulus configuration in the face of large amounts of random error remains surprisingly accurate when the appropriate (or weaker) measurement assumption is made. Fourth, the weight structure is degraded by the presence of random error. And fifth, the combination of monotonic and random error is totally detrimental when the measurement level is assumed to be ratio.

Real Data and the Weighted Model

We now investigate the behavior of ALSCAL with real data appropriate to the weighted Euclidian model. We choose data which have been previously analyzed so that we can compare our results with those already published. Specifically, we employ data gathered by Jones and Young [1972] who successfully employed the weighted model to describe the social structure of a small, intact, and naturally occurring task-oriented group (the staff, students, and faculty of a university-based teaching and research laboratory). They used Carroll and Chang's INDSCAL algorithm to obtain three dimensions which, with the help of additional data and analysis, they interpreted as representing the status, political persuasion, and professional (task) interests of the members of the group. They were able to interpret detailed characteristics of both the stimulus and weight spaces with great success.

When we analyzed these data with ALSCAL under the assumption that they were measured at the ordinal level, we obtained a solution whose stimulus was essentially identical to that obtained by Jones and Young (who used the ratio assumption). However, the ALSCAL weight structure was more homogeneous than the one found by Jones and Young. When these data were reanalyzed under the ratio assumption, the stimulus configuration was essentially unchanged, but the weights were more heterogeneous. In both cases the weight structure was interpretable in a manner similar to the Jones and Young interpretation, even though it was not identical. Although the homogeneity of the weights is partly a function of measurement level, more homogeneous weights can be expected with ALSCAL than with the INDSCAL method, as will be discussed in Section 5. Finally, we note that these analyses assumed the data were matrix-conditional, which is, implicitly, the assumption made by Jones and Young in their use of INDSCAL. When the analysis is performed with the unconditional assumption, the results are quite different, and not easily interpretable.

The second set of real data analyzed with the weighted Euclidian model was collected by Jacobowitz [Note 12; Young, 1975b]. These data are particularly suited to our purposes since they are row-conditional data, and since there have been no previously developed algorithms for applying the individual differences model to such data. (There are, however, several algorithms for fitting the simple Euclidian model to conditional data.)

The stimuli forming the basis of these data are fifteen names of body parts. Each subject was presented with a single one of these fifteen stimuli and was asked to rank order the remaining fourteen stimuli in terms of their similarity to the fifteenth (called the standard stimulus). Another stimulus was then selected to be the standard and the process was repeated. The subject was required to produce fifteen such conditional rank orders, each a rank order of fourteen stimuli with respect to their similarity to the fifteenth. (The study also involved three other sets of stimuli . . . kinship terms, color terms, and "have" verbs . . . which we do not cover.)

There were fifteen subjects at each of four age levels, the ages being 6 year olds, 8 year olds, 10 year olds and adults. In our analysis we included only the youngest and oldest groups since if there are any reliable individual differences (which Jacobowitz found by analyzing each age group separately with the Euclidian model), they should most certainly appear between the two most extreme age groups.

ALSCAL obtained three dimensions which were similar to those obtained in Jacobowitz's previous analyses with the simple Euclidian model (see Figures 5 and 6). Dimension I (vertical) is interpreted as face terms vs. limbs (both upper and lower) with "body" in between. Dimension II (horizontal) contrasts upper limbs with lower limbs, with face terms and "body" in between. Dimension III (front-to-back) represents "body" vs. everything else (or more precisely, whole vs. parts hierarchy). In Figure 6 we present the associated weight configurations in which the weights for adults and children are indicated by different symbols. (Zero weight on all dimensions is at the lower back corner of the cube; the further away from this corner, the heavier the weight.) We observe a clear distinction between the two groups of subjects; the groups are almost perfectly separated. Every child puts more weight on Dimension II (horizontal) than each adult, whereas adults are nearly always better represented by the combination of Dimensions I and III. In light of the previous interpretation this indicates that 6 year old children are relatively homogeneous (they uniformly emphasize the second dimension), whereas adults are more heterogeneous (they split between Dimensions I and III). No adults evaluate Dimension II highly, but three of them are inclined to emphasize Dimension III rather than I. We do not have any further evidence, however, concerning which factors distinguish Dimension III adults from Dimension I adults (who are in the majority). The clear distinction between younger children and adults in their way of evaluating dimensions of body parts seems very interesting and of empirical importance.

Real Data and the Unweighted Model

The evidence supporting the robustness of ALSCAL in the unweighted case is clear and abundant. We have analyzed Funk et al.'s ethnic data [1975], McGuire's size confusion data [Shepard, 1958, p. 511], Ekman's color data
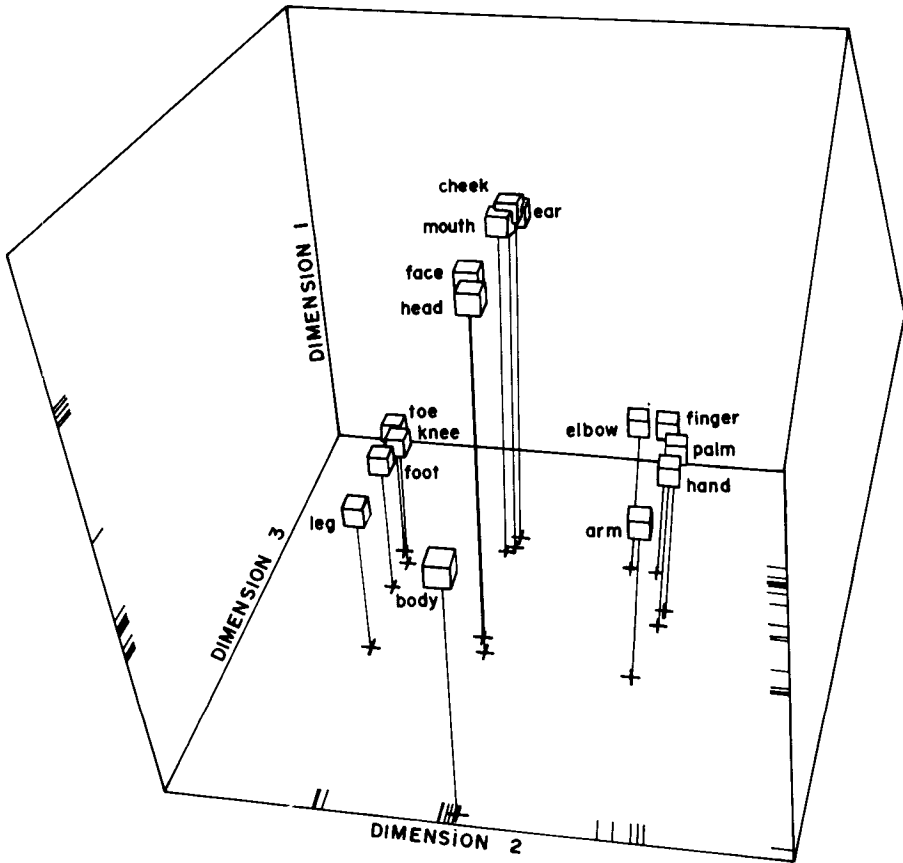
FIGURE 5
Jacobowitz Body parts data: Three-dimensional stimulus space.

[1954, p. 468], Miller and Nicely's sound data [1953], Peterson and Barney's vowel data [1952], Green and Rao's breakfast menus data [1972], and Hayashi's rice data [1974] among others. In all cases the obtained stimulus configuration was virtually indistinguishable from the published results, even though the latter were obtained by a variety of MDS algorithms.

We do not present any of the above results in detail. Instead we present some of the results we obtained under measurement assumptions weaker than those made by the above authors. Hayashi [1974] analyzed the dissimilarity of various rice strains by his recently proposed MDS method, which makes the assumption that the dissimilarities are defined at the ordinal level. We reanalyzed his data with ALSCAL under the assumption that they are defined

at only the nominal level, a particularly weak assumption in this case since there are only four observation categories. Our results are in close agreement with Hayashi's (see Figure 7). Note that the (nominal) observation categories are ordered, at the conclusion of the analysis, in the fashion assumed by Hayashi. We obtained these results from (ordinally incorrect) initial category values which were generated randomly, as well as from the (ordinally correct)
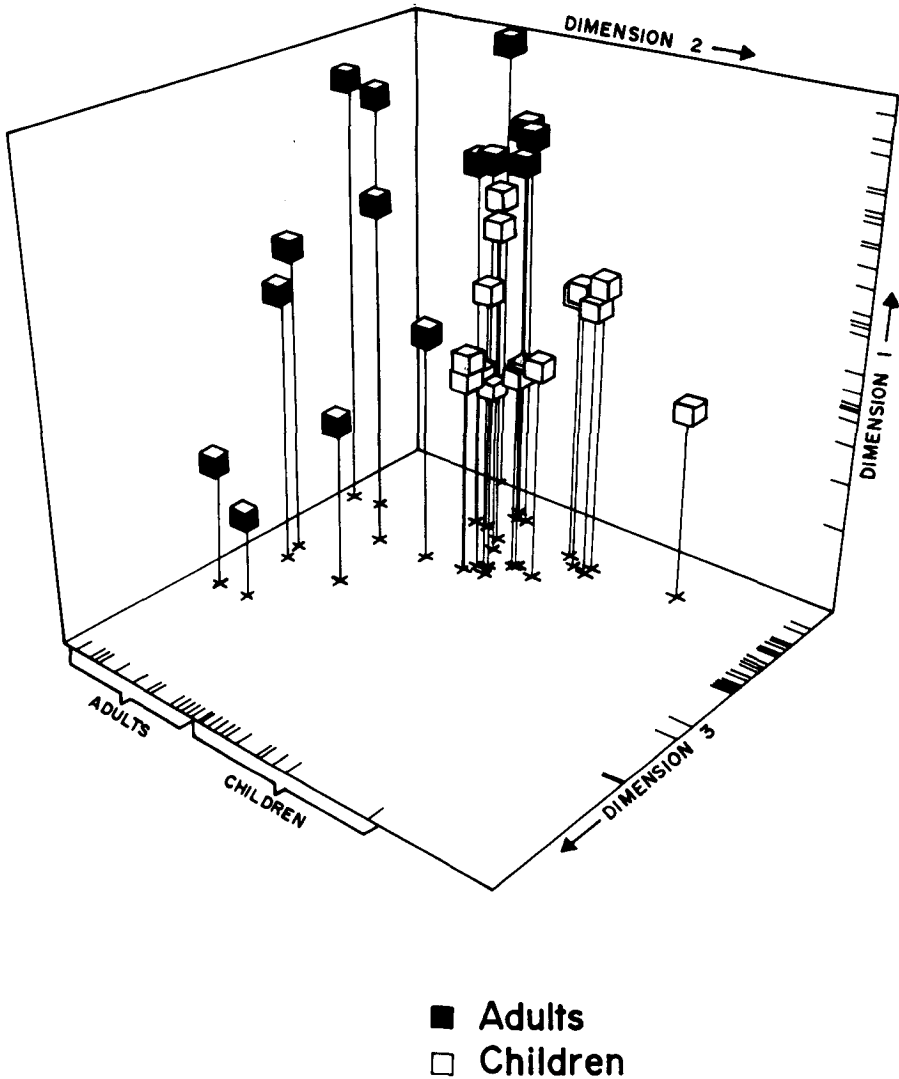

■ Adults
□ Children

FIGURE 6
Jacobowitz Body parts data: Three-dimensional weight space (C for children, A for adults)
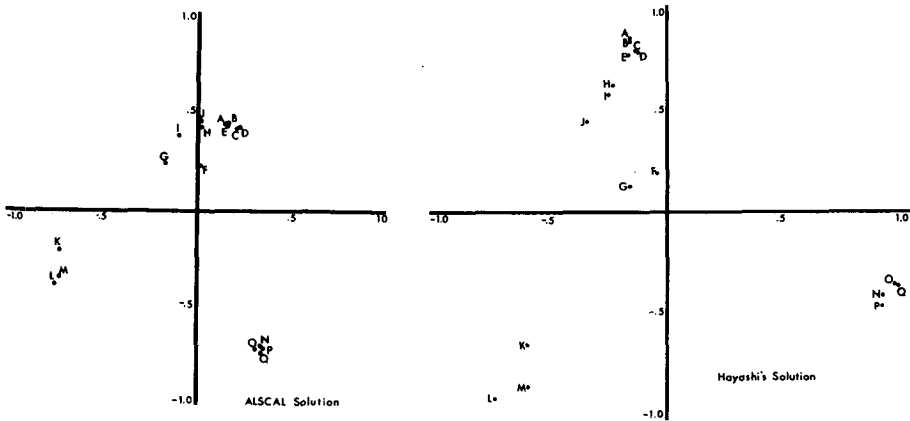
FIGURE 7
Hayashi's rice-strain data: Stimulus space (unweighted model).

values used by Hayashi. Thus, we see that ALSCAL converges to the same solution independently of the initial category values, though the number of iterations required to reach convergence is of course larger for the random values.

We reanalyzed Ekman's color data using the nominal measurement assumption, and collapsing the number of observation categories to nine by combining them. We analyzed the collapsed observation categories under both ordinal and nominal assumptions and in both cases obtained essentially the same color circle as Ekman (see Figure 8 where the numbers indicate color wavelength). This is in spite of the fact that the data are similarities (not dissimilarities), which means that the order of the values assigned to the observation categories is the reverse of the order of the desired distances. For the ordinal assumption, the user informs ALSCAL to compensate for the reversal, of course. However, for the nominal assumption the initial category values, being equal to the raw observations, are in the worst possible order relative to the desired distances. They are worse than randomly generated values. Even so, ALSCAL is able to overcome this very poor initialization and obtain the desired configuration (at the cost of a number of iterations). Finally, it should be pointed out that the quantification of the category values which was obtained was essentially the same for both measurement assumptions, implying that the ordinal assumption is appropriate.

For the unweighted Euclidian model we conclude that a) ALSCAL reveals the same stimulus structure as other algorithms; b) ALSCAL is able to obtain identical solutions under the nominal measurement assumption as under the ordinal assumption when the stronger assumption is appropriate; and c) the obtained stimulus structure is uneffected by choice of initial category values when the nominal assumption is used.

x   raw data
△   collapsed nominal
•   collapsed ordinal
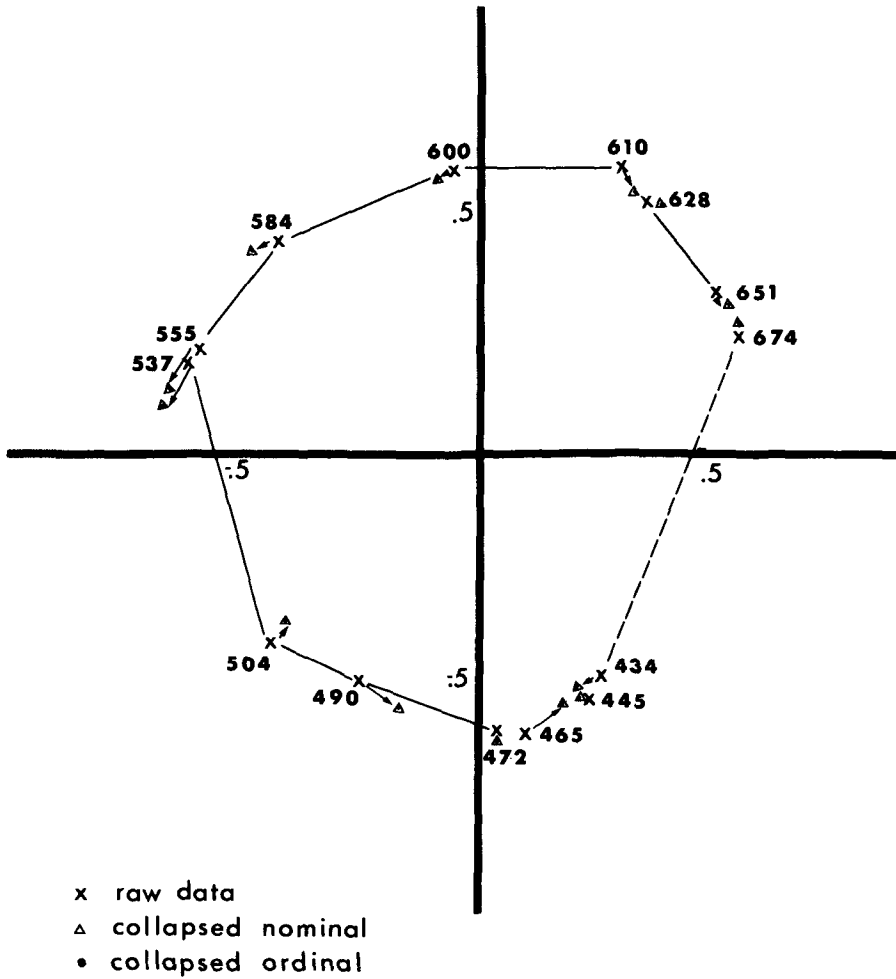
FIGURE 8
Ekman's color data: Stimulus space (unweighted model).

5. Discussion

Having completed the presentation and evaluation of the model and method, we now turn to a discussion of some related issues.

Interpretation of X and W

We will not dwell at length on the interpretation of X and W since Carroll and Chang [1970] have already done so. The interpretation of X is in every way identical to the interpretation given in the earlier work (X represents the stimuli as points in an unrotatable Euclidian space with dimensions

of unit length). However, there are three subtle differences in the interpretation of W, although its general nature is unchanged (W represents the subjects as vectors whose direction indicates the relative importance of each dimension to each subject).

The first difference in the interpretation of W is that with unconditional data it is permissible to make direct intersubject weight comparisons, whereas for conditional data (of either type) and for Carroll and Chang's method (which is tacitly matrix-conditional), inter-subject comparisons can only be made indirectly via within-subject weight ratios (a point often overlooked with the earlier procedure, by the way). For example, if Subject A has weights of .80 and .60 on the two dimensions of a configuration, then for any type of data we may say that Subject A places $1.33 = .80/.60$ more weight on Dimension one than on Dimension two. Similarly, if Subject B has weights of .40 and .60 on the same two dimensions, then we may say that he places .67 as much weight on Dimension one as he does on two. Such within-subject comparisons are straightforward. However, with between-subject comparisons we must be careful, as it is only for unconditional data that we can make the simple statement that Subject A finds Dimension one twice as relevant as Subject B does, and that they both find Dimension two to be equally relevant. For conditional data, on the other hand, we must say that Subject A emphasizes Dimension one relative to Dimension two twice as heavily as Subject B does (since $1.33/.67 = 2$). It is the case, however, for all types of data that the magnitude of the weights (the length of the weight vector, say) indicates in a general way the degree to which the subject's data are represented by the solution obtained by ALSCAL. We discuss this topic next.

The second difference is in the interpretation of the length of the weight vectors. The general interpretation is the same for both procedures; it is said that they loosely represent the goodness of fit of the model to the data obtained from the individual subject. More specifically, for both procedures it can be said that the length of the weight vector (sum of squared weights) roughly represents the proportion of variance accounted for in the subject's scalar products. The difference is that in the Carroll and Chang approach this "variance accounted for" is being optimized, whereas in our procedure it is not. As was noted by Carroll and Chang, the "variance accounted for" notion is only precisely true when the configuration is exactly orthonormal ($X'X = I$). When the configuration is only approximately orthonormal, as is the usual case, then this interpretation of the length of the weight vector is only roughly true. Note carefully that the weight vector length does not represent the proportion of variance (or of anything else) accounted for in the subject's judgments.

The third difference in weight interpretation is in the meaning of a vector of zero weights when the data are assumed to be at the ratio level of measurement. In the Carroll and Chang situation, zero weights for a subject means

that the model of his judgments consists of a scalar product matrix which is entirely zero. In our situation, however, the same subject would have a distance matrix which was entirely zero. Now in the Carroll and Chang approach the model's zero scalar products matrix is fit to a set of (pseudo)-scalar products (those computed from the data) which have a zero mean. Thus the mean of the two matrices is the same. However, in our approach the zero distance matrix is fit (in the ratio case) to the raw data. The data do not have a zero mean, and thus the means of the data and distances are not the same. Therefore, in our approach a vector of zero weights is going to contribute relatively more to the apparent lack of fit than in the Carroll and Chang approach. In a practical sense this means that for ratio data, zero weights are less likely to occur with our approach, and that the weight structure obtained with our approach may be similar to the Carroll and Chang weights, but certainly not identical (except in certain unlikely situations). In particular, we expect that our weights should tend to be more nearly homogeneous than those obtained from the Carroll and Chang procedure, when the data are assumed to be measured at the ratio level. Thus, we caution the user to be careful in opting the ratio assumption, due to the effect of the weights. (Carroll, in a personal communication, also warns of the use of the ratio assumption with his procedure, but for different reasons.)

Individual Differences

   As was briefly mentioned in the introductory section, there are several different multidimensional scaling models realizable within the ALSCAL framework. The models are obtained by combining either the weighted or unweighted Euclidian model with one of the three types of conditionality, and with either one or more than one subject (several of the combinations are either impossible or nonsensical). We discuss the meaningful models briefly in this section.

   While most of the models can be collectively referred to as individual differences models, there are two distinct types of non-individual differences models. One of these is the standard unweighted Euclidian model applied to a single matrix of data (i.e., when $N = 1$). Clearly this is not an individual differences model since there is but one individual. The other non-individual differences model is obtained when one analyzes several matrices of data with the unweighted Euclidian model under the assumption that the data are unconditional. While it might appear that this is an individual differences model (since there are several matrices), the reasons that we view it as a non-individual differences model will become clear after the discussion of individual differences in the next few paragraphs.

   There are three psychologically distinct individual difference models realizable within the ALSCAL framework. Individual differences can be allowed only in the response process (i.e., response bias), only in the judg-

mental process (including perceptual and cognitive processes), or in both the response and judgmental processes. It should be clear by now that individual differences in judgmental processes are reflected by the weights of the weighted Euclidian model. Thus we must choose this model if we are interested in allowing for the Horan [1969] and Carroll and Chang [1970] type of individual differences. It may not be so clear, however, that by assuming the data are conditional we are implicitly allowing for individual response bias differences, the type allowed for by McGee's [1968] developments. Thus, if the data are measured at the ordinal level, each individual is allowed to have his own unique monotonic response transformation, while if the data are interval, each individual has a unique linear response transformation. Note that this type of individual differences results from either type of conditionality, since for row-conditional data each individual has a unique *set* of response transformations, while for matrix-conditional data each has a single unique transformation. However, if we make the assumption that the data are unconditional, then we are assuming that all individuals have identical response biases, and thus tacitly assuming that there are no individual differences in this regard.

Thus, we can allow for two types of individual differences via either the model weights or the data conditionality. Obviously, we can permit both types of individual differences to occur by simply applying the weighted Euclidian model to conditional data. But what happens if we apply the unweighted model to unconditional data? Then we have the second type of non-individual differences model discussed above. This model allows for replicated data, but assumes that the replications arise from subjects with identical judgmental and response processes.

Oblique Axes and Individual Rotations

Several weighted models have been proposed which are more general than the one discussed here. Among these are IDIOSCAL, a model which allows for individual differences in the orientation of axes [Carroll & Chang, Note 2]; PARAFAC, a model which permits individuals to have weighted oblique dimensions [Harshman, Note 10]; and an extension of Tucker's three-mode factor analysis [Tucker, 1966; Levin, 1965] to multidimensional scaling [Tucker, 1972]. All of these models have been proposed in the scalar products framework. Thus they optimize the STRAIN index (8) with the definition of the weights matrix changed in different ways for the different models.

As has been discussed by Carroll and Chang [Note 2], the distance version of these models (as well as the models covered by our previous developments) are all special cases of the following distance model:

$$(49) \qquad d_{ijk}{}^2 = \sum_{a=1}^{t} \sum_{b=1}^{t} (x_{ia} - x_{ja}) r_{kab} (x_{ib} - x_{jb}),$$

or, in matrix notation,

$$(50) \qquad d_{ijk}{}^2 = (x_i - x_j)'R_k(x_i - x_j).$$

Here x_i is a column vector of coordinates for point i, and R_k is a square symmetric matrix of inter-dimension relations for subject k. The relationship of this model to the one treated by ALSCAL is that ALSCAL restricts R_k to be a diagonal matrix.

The other models mentioned at the beginning of this section are obtained as follows. For Carroll and Chang's model we obtain the spectral decomposition of R_k .

$$(51) \qquad R_k = U_k W_k U_k',$$

where U_k is orthogonal and W_k is diagonal; where U_k can be interpreted as a subject's orthogonal rotation of the original coordinates X to a new orientation; and where his weights W_k are applied to the rotated configuration. Thus this model allows for individual differences in the orientation of axes as well as the types of individual differences discussed in the preceding section. (Note that the orientation of X is not unique.)

For Harshman's model we decompose R_k so that

$$(52) \qquad R_k = W_k C W_k',$$

where W_k is diagonal, and C is square symmetric with unit diagonals; where C is interpreted as a matrix of cosines of angles between oblique dimensions; and where W_k is a subject's weights on the obliquely transformed dimensions. Thus this model allows for the same types of individual differences as discussed in the previous section, but makes the fundamentally different assumption that the axes which are being weighted are oblique transformations of the stimulus space X (whose orientation is uniquely determined). Note that all subjects weight the *same* oblique dimensions.

For Tucker's model we decompose R_k so that

$$(53) \qquad R_k = W_k C_k W_k',$$

where the matrices have the same nature as in Harshman's model, with the essential difference that each subject k has his own oblique transformation C_k, as indicated by the subscript. Thus Tucker permits a type of individual difference not covered by the previous models, namely that each individual has his own personal oblique transformation of the coordinate space, as well as his own weighting of the dimensions of his space. (Note that the orientation of X is not unique here.) This decomposition of R_k is the most general of those presented in this section, including all of the models previously discussed in this paper.

ALSCAL can be easily extended to cover any of the models treated in this section by modifying the weight estimation phase (Section 3). The modification is to redefine the matrix Y so that all pairs of dimensions are

present as well as all pairs of points. Thus, if we define Y to be an order $n(n - 1)/2$ by $t(t + 1)/2$ matrix with general element $(x_{ia} - x_{ja})(x_{ib} - x_{jb})$, and then apply (36) we would obtain least squares estimates of the R_k, which can then be decomposed in the desired way.

Minkowski Spaces

One of the limitations of the work presented here is that it only encompasses Euclidian coordinate spaces and does not include other Minkowski spaces. Such a generalization, which is very simple with the standard gradient approach [Kruskal, 1964; Lingoes, 1973; Young, 1972], is annoyingly difficult in the ALS approach. In fact, the extension is impossible within the ALS framework unless we adopt a different optimization criterion. If we defined lSTRESS on the l-power weighted Minkowski distances, i.e.,

$$(54) \qquad d_{ijk}{}^l = \sum_{a=1}^{t} w_{ia} |x_{ja} - x_{ka}|^l, \qquad w_{ia} \le 0, \qquad 1 \le l \le \infty,$$

so that lSTRESS would become

$$(55) \qquad \phi^2(X, W, D^*) = \sum_i \sum_j \sum_k^{j-1} (d_{ijk}*^l - d_{ijk}{}^l)^2,$$

then with some rather minor modifications in Section 3 we could extend our developments to other Minkowski spaces. However, this proposal is not entirely meaningful, especially when the value of l is at all large. It is interesting to note, though, that for City Block ($l = 1$) lSTRESS is identical to STRESS. Thus it would be both simple and meaningful to extend ALSCAL to include City Block space. Such an extension would also be rather useful since City Block space is probably the most commonly used non-Euclidian coordinate space in applications of multidimensional scaling to social science data. However, this extension might not be robust due to the well-known frequency of local minima in City Block space.

Weighted Unfolding Models

A further limitation of the work discussed in this paper is that it does not apply to Coombs's unfolding model [1964]. This model can easily be incorporated into nonmetric multidimensional scaling programs as has been discussed by Young [1972], and carried out by Kruskal, Young and Seery [Note 14], Lingoes and Roskam [1973] and Young [1973]. It would be possible to incorporate the unfolding model within the ALSCAL framework by using the notions discussed by these authors. If this were done, then the full model (incorporating the oblique axes and individual rotations notions as well as the Minkowski notions discussed in Section 5) would be

$$(56) \qquad d_{ijk}{}^l = \sum_{a=1}^{t} \sum_{b=1}^{t} |x_{ia} - y_{ja}|^{l/2} r_{kab} |x_{ib} - y_{jb}|^{l/2},$$

where the new terms y_{ia} and y_{ib} are coordinates of the "ideal point," in Coombs's terminology. We obtain Coombs's unfolding model by assuming that the space is an unweighted Euclidian space (i.e., $l = 2$, and the matrices R_k are identity matrices). When these restrictions are placed on (56) we obtain Coombs's method for modeling individual differences, as has been discussed extensively by Coombs. We also obtain a way of modeling any second set of objects as points in the Euclidian space, as has been discussed by Young [1972] among others. For example, if we have a matrix of judgments concerning the degree to which each of several American Ethnic subgroups can be described by each of several adjectives, then we can use the unfolding model to develop a "joint" Euclidian space containing points representing both the Ethnic groups and the adjectives. Precisely this approach was taken by Funk, et al. [in press]. If the ALSCAL procedure were extended to incorporate the model implied by (56), then we could also perform an individual differences unfolding analysis of the type of data gathered by Funk, et al. Whereas Funk, et al. had to average the data they gathered from 50 subjects to obtain a single matrix to be "unfolded," it would be possible to avoid this averaging procedure and adopt a model incorporating the Carroll-Chang type of individual differences weights (or indeed, any of the other types of weights discussed in Section 5). This approach would be particularly ideal for the Funk, et al. study, since it is reasonable to assume that the particular Ethnic subgroup to which the judge belongs affects his judgments of the degree to which certain adjectives describe certain Ethnic groups. We issue a note of caution, however, as there are some well-known degeneracies which occur with the unfolding model, and we might expect that this extension of the ALSCAL procedure would not be any more robust to these problems than preceding procedures.

Measurement

Within our framework one can obtain empirical information about the measurement level of his raw data, at least within the context set by the MDS model. All that has to be done 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 equivalent analyses. He can then conclude that within the MDS situation 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.

The reasoning behind these statements is as follows. If a set of raw data is analyzed twice, and if the only difference in the two sets of analysis options is the assumed measurement level, and if the obtained results (X, W, D^*, and SSTRESS) are identical for both analyses, then the lower measurement level (which places relatively weak restrictions on the optimal scaling) is yielding

exactly the same transformation as the higher measurement level (which enforces stricter conditions). That is, if the two analyses involve nominal and ordinal assumptions and yield identical results, then in the nominal case the transformation actually satisfies the ordinal requirements. When this occurs it is appropriate to conclude that the data are in fact measured at least at the higher of the two levels of measurement when these data are analyzed with the chosen model.

Note that the view of measurement implied by the preceding statements is not the common view. We do not adopt the commonly held position that measurement level is a characteristic of data *in vacuo*. Rather, it is our view that the measurement level of a particular set of data is dependent on the interaction of that data with the model chosen to describe the data. When a set of data is analyzed by some model, the method of analysis necessitates assuming that certain types of data transformations are permissible. These transformations, and the operations they entail, imply that a certain level of measurement has been assumed to exist in the data. If one can vary the types of allowable transformations, and only perform operations on the data which are commensurate with the transformations, then one can determine how well the data "measure up", as it were, to the requirements of each measurement level. This is the approach taken here. Note ,however, that this cannot be done outside of the context created by the chosen model, as should be clear. It may be that a set of data is monotonically (but not linearly) related to the distances of an MDS model, but it would not be correct to conclude that they are ordinal for it may be the case that they are linearly related to some other model.

It may appear to be the case that the argument is purely academic, and that the situation will never arise in practice. After all, we are requiring that the results of the several analyses be exactly equivalent. However, the situation actually occurred in one of the examples given above. For the Hayashi [1974] data the nominal and ordinal results were precisely identical, allowing us to conclude that the raw data that we analyzed were at least ordinal in the MDS context. We do believe ,though, that our requirement of strict equivalence is overly stringent, and we would prefer to develop a test to indicate how well a particular set of data approximates a particular measurement level. We have not yet done this, however.

Our view of measurement differs from the common view in one more fundamental way. As was implied by the end of the previous paragraph, we do not view measurement as being at one of a set of discrete levels. Our view is that measurement level is a continuous notion, not a discrete one. While it is obviously the case that only certain discrete points on the measurement level continuum can be axiomatized, it is not our understanding that these are the only measurement levels. The intermediate measurement levels between the various axiomatizable points represent levels of measurement

which approximate, to a greater or lesser degree, the next higher axiomatized level. Thus, if we analyzed a single set of data under nominal, ordinal and interval assumptions, and we discovered that the results were identical for the nominal and ordinal cases, and "very similar" in the interval case, then we would conclude that the measurement level of the data when analyzed by the chosen model is somewhere between the ordinal and interval points, and perhaps nearer the interval point. The most critical feature of the analysis for deciding how nearly one approximates a particular measurement level is the investigation of the nature of the optimal scaling D^*. In the example just given, to conclude that the results were "very similar" in the interval case, we would have to go back to the ordinal case and determine how far the optimally scaled data (D^*) deviate from linearity. Formally, we might obtain the Pearson correlation between D^* and the set of data O, as a descriptive indication of deviation from linearity (note that this is obtained for the ordinal level analysis for which the Spearmen rank order correlation between D^* and O is perfect). While this is an adequate descriptive device, clearly we cannot use it for formally testing a measurement level hypothesis, which is what we would most like to do.

This notion of a measurement continuum is involved in another important aspect of our situation. It is commonly stated that nonmetric procedures quantify qualitative data. Indeed, one of the main reasons for the popularity of nonmetric procedures is this magical conversion of measurement level. Strictly speaking, such a conversion of measurement level only occurs, in our view, when the quantitative model perfectly describes the qualitative data. Thus in our situation it is necessary to obtain a zero SSTRESS value in order to precisely quantify qualitative data. The degree to which SSTRESS is not zero indicates the degree to which we were unable to quantify our data with the MDS model. Rephrased in the terms used in the preceding paragraph, the SSTRESS value indicates how far along the measurement level continuum we have moved from the assumed measurement level towards the ratio measurement level (which is the level of the MDS model). Perhaps a more useful index of quantification would be the Pearson correlation between the optimally transformed data D^* and the distances D. Note that if the same set of data is analyzed under several different measurement level assumptions, then the SSTRESS (and quantification correlation) will be best for the weakest assumption, indicating, as it should, that we have moved further along the measurement continuum. However, this is not due to the fact that we have reached a higher degree of quantification, but to the fact that we assumed a lower degree of qualification, as it were.

Finally, these two uses of the measurement continuum, and the two descriptive correlation indices proposed, are perfectly commensurate with each other. For the Hayashi [1974] data analyzed in the previous section, a Spearman rank order correlation performed between O and D^* for the nominal

analysis would be unity, indicating that the data are actually ordinal when analyzed by the chosen model. The Pearson correlation between D^* and D would be the same for the two analyses (as in the SSTRESS), meaning that no more quantification was possible under the nominal assumption that under the ordinal assumption. This implys that the data are ordinal. Finally, the Pearson correlation between D^* and D is not unity (nor is the SSTRESS zero), indicating that the data are not perfectly consistent with the model, and therefore that the model has not been able to perfectly quantify the data. Please keep in mind that we only use the correlations descriptively, and that the main weakness of our proposal to use such indices to investigate measurement level is that we have no formal methods for deciding when a goodness of fit measure is significant.

Efficiency

The last topic we take up is the efficiency of ALSCAL, both in terms of speed and memory requirements. The memory requirements of ALSCAL are most easily discussed, so we take them up first. As compared with the metric INDSCAL, only about one-half of the amount of data may be accommodated in the same amount of space. This follows from the fact that with a nonmetric program, one must store both the original data and the optimally scaled data, whereas with a metric program, one only needs to store the data. Thus twice the core storage is required with a nonmetric program. In most other regards ALSCAL and INDSCAL are comparable in terms of storage requirements. Of course, the storage requirements of ALSCAL are roughly comparable to those for other nonmetric MDS programs, with the added storage for subject weights being balanced by the lack of a gradient matrix.

Turning now to the speed of ALSCAL, we first discuss the manner in which the speed is a function of various aspects of the analysis situation. Note that there are four separate computational sub-problems: a) solving for initial values; b) obtaining the optimal scaling transformations; c) computing the weights; and d) determining the configuration. Of these four problems all except the weight problem are adversely effected by increasing the number of points. On the other hand, if the number of subjects is increased only two phases, the optimal scaling and weight phases, will be slower. If we increase the number of dimensions then all phases should be slower except the optimal scaling phase which will be uneffected (except in the ordinal case where increasing dimensionality will improve efficiency, due to the likelihood that the order will be more nearly correct). Finally, the ordinal measurement level should take noticeably longer than any of the other levels, due to the sorting. In Table 2 we present the times required to analyze the Jones and Young [1972] data as a function of dimensionality and measurement levels. These times are CPU time only, with no $I/0$ time included. We have set the convergence criterion to a value of $\delta = .001$ where δ is the improvement in

TABLE 2

CPU time/number of iterations required for program convergence.
(The CPU units are arbitrary.)

Dimensionality	Measurement Level			
	nominal	ordinal	interval	ratio
1	5.2/4	22.1/4	6.4/4	6.0/4
2	7.8/4	16.7/4	8.2/4	6.4/3
3	13.4/5	21.6/5	11.7/4	11.9/4

SSTRESS from one iteration to the next. (Note we use ϕ, that is, the square root of [11]). We also present the number of iterations to convergence.

Evaluating an algorithm's speed relative to another algorithm is a difficult problem, as has been stressed by Spence [1972] and Lingoes and Roskam [1973]. Here the main source of difficulty is the fact that ALSCAL optimizes a different function than any of the other routines, so it is difficult to ensure that the various programs are obtaining equally precise solutions. We follow the lead of Spence and simply use the default termination values associated with each program. While this does not get around the precision problem, it does at least correspond to the likely state of affairs in the real world. In Table 3 we present the CPU times required to analyze the Hayashi [1974] data in two dimensions by ALSCAL, KYST and POLYCON, and the CPU times required to analyze the Jones and Young [1972] data in three dimensions by ALSCAL and INDSCAL. (KYST and POLYCON were both optimizing Kruskal's second STRESS formula whereas ALSCAL was not, which accounts for the larger stress value obtained from ALSCAL.) We also present the value of Kruskal's first STRESS formula for comparison. (Note that none of the programs optimized this formula but perhaps STRESS 2 is closer to STRESS 1 than SSTRESS 1.) Finally, we have also presented the last improvement in the function being optimized as a rough precision indicator. We believe it is fair to conclude that ALSCAL is more efficient in terms of computation time than other currently available programs.

We must admit that the relative speed of ALSCAL is a fortuitous rather than an anticipated result. Perhaps the speed of ALSCAL is related to a fact recently reported in the numerical analysis literature. There is a class of algorithms, called nonlinear block successive overrelaxation algorithms [Hageman & Porsching, 1975] which are very closely related to ALS algorithms, and which are currently quite popular among numerical analysts. These algorithms are like an ALS procedure in that they divide the estimation problem into a series of conditional estimation problems (successive blocks), each of which has an analytic solution. These algorithms differ from an ALS

TABLE 3
Comparison of the efficiency of several scaling algorithms.

Program	CPU time	Itera- tions	STRESS 2	STRESS 1	Improve- ment	Data
ALSCAL (nominal)	6.3	6	.476	.251	.0001	Hayashi
ALSCAL (ordinal)	5.7	6	.476	.251	.0001	Hayashi
KYST	15.1	16	.429	.211	.0001	Hayashi
POLYCON	56.8	25	.455	.225	.0001	Hayashi
ALSCAL (ratio)	11.9	4	–	.302	.0003	Jones & Young
INDSCAL (ratio)	63.4[a]	17	–	–	.0098	Jones & Young

[a]Another run with a different random start took 73.5 CPU seconds.

procedure in that they do not go precisely to the minimum in each subproblem, but overstep the minimum. The overstepping is referred to as overrelaxation. It has been found that these procedures are fastest when the several sub-problems involve approximately the same number of parameters. This condition holds, roughly, in ALSCAL. It has been found with these procedures that overrelaxation improves the efficiency of the algorithm. Thus we may be able to further improve the efficiency of ALSCAL by this technique. (We are currently looking into this possibility.)

Finally, it should be noted that the order in which the three conditional minimization problems are solved is not very critical in terms of the parameter values eventually obtained at convergence. Nor indeed does it appear that the initialization procedure is very critical in this regard, although other procedures may evidence more frequent incidents of local minima solutions (which are seldom, if ever, obtained with the initialization used here). Further-more, no matter how frequently we solve one of the subproblems relative to another one (within reasonable limits, of course), we eventually obtain the same estimates. Thus, the ALS approach is somewhat arbitrary in these terms. However, it is the case that the speed of convergence is heavily affected, and our particular choice of flow was strongly related to this concern. From our experience with ALS procedures, it seems that the most efficient procedure is the one in which each subproblem is solved the same number of times in an iteration. Thus, it is usually more efficient to solve each subproblem once per

iteration than to solve for X, say, three times and the other aspects once. This experience is probably closely related to the numerical analysis result reported in the previous paragraph.

Local Minima

As was mentioned at the end of the section on computing coordinates (Section 3), a theoretical difficulty was found to exist in the proposed method. (This difficulty was first pointed out to us by Robert F. Baker, and was independently noted by J. Douglas Carroll, to both of whom we owe our gratitude.) In this section we discuss the issue fully, and propose a revised procedure which is free from local minima problems. It should be emphasized at this point that the theoretical problem appears to have no practical consequences, as should be clear from the evaluation presented in Section 4. Furthermore, the procedure to be proposed, while having a certain theoretical beauty, is of unknown efficiency, and may prove to be less rapid than the procedure evaluated above.

Recall that the function being minimized (43) is a quartic function, and that its gradient (44) is a system of simultaneous cubic equations. Let us simplify the situation by considering the form of the quartic equation when there is only a single variable; i.e., for a single coordinate instead of for all t coordinates of a point. Since SSTRESS can never have a value less than zero, the quartic equation (which is SSTRESS) must always have a positive value. This implies that it has at most two minima and one maximum (is "W" shaped) and cannot have two maxima and one minimum (cannot be "M" shaped). (We must say "at most" because it will not always be the case that these minima and maxima will all be distinct.) If we now consider the case of t coordinates varying simultaneously (as was done in Section 3), we will come to the conclusion that there are at most a total of 3^t roots to the system of simultaneous cubic equations forming the gradient of the function. Geometrically, these 3^t roots correspond to a single maximum, 2^t minima (of which at least one is the overall minimum), and $3^t - 2^t - 1$ saddle points. The Gill and Murray variation of the Newton-Raphson procedure which we use to locate the minimum is designed to avoid all saddle points (and to avoid the maximum, of course), but it cannot distinguish between the 2^t minima. It simply goes to the "nearest" one. This is the theoretical problem with our procedure. In fact, in a technical sense, the procedure we have proposed should not be called an alternating least squares procedure, since we are not assured of always finding the conditional minimum in each phase, which is the defining characteristic of such an ALS procedure.

If we consider exactly what happens to each coordinate, one at a time, we can develop a modification of the proposed algorithm which is truly an alternating least squares procedure. These are two basic types of situations which can occur in the solution of a single cubic equation in one variable.

Type I: The cubic equation has three real roots.
Type Ia: All three roots are distinct.
Type Ib: Two of the roots are equal.
Type Ic: All three roots are equal.
Type II: The cubic equation has one real root and two imaginary roots.

(It is not possible for a cubic equation to have two real roots and one imaginary root.) When all three roots are real and distinct (Type Ia), the cubic equation can be written (in general form) $(x - a)(x - b)(x - c) = 0$, where a, b and c are the roots and we can assume, for the sake of simplicity, that $a < b < c$. In this case the minima of the quartic function will necessarily correspond to roots a and c, i.e., to the smallest and largest roots (recall the W shape of the quartic). When either $a = b < c$, or $a < b = c$ (Type Ib), the cubic equation can be written as either $(x - c)(x - b)^2 = 0$ or as $(x - a)(x - b)^2 = 0$. It can be shown in this case that the distinct root corresponds to the minimum, and that the non-distinct roots correspond to a stationary point which is neither a minimum nor a maximum, but a level (plateau) point. In the case where $a = b = c$ (Type Ic), the cubic equation can be written $(x - a)^3 = 0$, and all three real roots correspond to the single minimum of the (degenerate) quartic (which also has no other stationary points). Finally, let us consider the case of one real and two imaginary roots (Type II). In this case the cubic equation can be written $(x - a)(x^2 + 2dx + e) = 0$, where $(d^2 - e) < 0$. (Note that when $d^2 - e = 0$ we have cases Ib or Ic, and when $(d^2 - e) > 0$ we have case Ia.) In this case the quartic function has only one minimum and no other stationary points, as in Type Ic. (The difference between Types Ic and II is that in the former the quartic is symmetric, whereas in the latter it is not.) Note that it is not possible to have a quartic which has one minimum and two plateau points since the first derivative of such a function is not cubic.

Based on the above analysis of the various types of situations which can occur in the solution of a cubic equation, we notice that we need either two or zero evaluations of the objective function to locate the global minimum. For the Type Ia situation we need to evaluate only the smallest and largest root, since the middle root corresponds to the maximum. For the Type Ib situation we do not need to evaluate the function at all since the distinct root will necessarily correspond to the minimum. Similarly, for the Type Ic and Type II situations we also do not need to evaluate the function since there is only one minimum (for Type Ic we can select any root, and for Type II we select the positive root).

If we wish to have a truely ALS procedure (which is in itself a debatable point), we can, then, switch to a procedure which estimates coordinates one at a time. In fact, this would probably not increase the computation time since the computations in (44), (45) and (46) (which lead up to [47]) are more

time consuming that the computations involved in (47) or in the procedure suggested here. We plan on investigating the efficiency of this new approach in the near future.

It sounds as though the approach just discussed has no local minimum problems, and is thus the first procedure for nonmetric multidimensional scaling to guarantee obtaining the global minimum. Unfortunately, this is not quite the case. While it is true that *once the process is started* there would be no local minimum problems, it is obviously the case that the final solution point is dependent on the initial solution point. Thus, it is more accurate to state that the solution obtained is the conditional global minimum, since it is conditional on the values used to start the entire process, but is in every other way free of local minimum problems. One now understands why such effort is spent on obtaining a good starting point.

6. *Conclusions*

We conclude that ALSCAL is the first viable algorithm for nonmetric individual differences multidimensional scaling.

ALSCAL is robust. As has been shown, ALSCAL can recover the true underlying structure in the Monte Carlo situation, at least when the measurement assumptions are appropriate and when there is not too much error introduced into the data. Furthermore, ALSCAL obtains the same structure as that obtained by other algorithms in those special cases for which algorithms have been previously developed.

ALSCAL is flexible. Most of the currently popular individual differences models and the widely used simple Euclidian model fall within the ALSCAL framework, thus ALSCAL is flexible with regard to the models which can be fitted to the data. Furthermore, ALSCAL is flexible with regard to the data since essentially all of the commonly discussed types of data (and some types not previously discussed) fall within ALSCAL's province.

ALSCAL is rapid. While there are difficulties associated with evaluating the rapidity of one algorithm relative to another, we tentatively conclude that ALSCAL is more rapid than previously developed algorithms.

The viability of ALSCAL leads us to feel very encouraged about the two keystones of our work, namely alternating least squares, and optimal scaling. Our previous work [de Leeuw, Young & Takane, 1976; Young, de Leeuw & Takane, 1976] has shown that these two keystones yield viable results with linear models. The current work extends this viability to quadratic models.

Note that the viability of our research is not bought without cost. Perhaps the main cost is that a separate, highly specific algorithm must be constructed for each class of models, thus eliminating the possibility of developing one very general algorithm for all situations.

An indirect cost associated with our work is that the alternating least squares approach to solving least squares problems, namely dividing the

problem into a series of simple subproblems, is only as simple as the simplest subproblem. In our previous work with linear models each of the subproblems was very simple. However, with the current work one of the subproblems, that of obtaining the best coordinate values, was not very simple, and the resulting algorithm is rather complex. Note that the derivation of the solution to a subproblem, which must be strictly least squares, may sometimes be difficult, as was the case here.

However, we believe that the costs of our approach are outweighed by the benefits. We are confident that the alternating least squares and optimal scaling keystones will provide a viable approach to other models in addition to the linear and quadratic ones investigated so far.

REFERENCE NOTES

1. Bloxom, B. *Individual differences in multidimensional scaling* (Research Bulletin 68-45). Princeton, N. J.: Educational Testing Service, 1968.
2. Carroll, J. D. & Chang, J. J. *IDIOSCAL (Individual Differences in Orientation Scaling)*. Paper presented at the Spring meeting of the Psychometric Society, Princeton, N. J., April, 1972.
3. Carroll, J. D. & Chang, J. J. *Some methodological advances in INDSCAL*. Paper presented at the Spring meeting of the Psychometric Society, Stanford, California, April, 1974.
4. de Leeuw, J. *Canonical discriminant analysis of relational data* (Research Bulletin RB004-75). Leiden, The Netherlands: Datatheorie, University of Leiden, 1975.
5. de Leeuw, J. *An initial estimate for INDSCAL*. Unpublished note, 1974.
6. de Leeuw, J. *The positive orthant method for nonmetric multidimensional scaling* (Research Note RN 001-70). Leiden, TheN etherlands: Datatheorie, University of Leiden, 1970.
7. de Leeuw, J. & Pruzansky, S. *A new computational method to fit the weighted Euclidean model (SUMSCAL)*. Unpublished notes, Bell Laboratories, 1975.
8. Gill, P. E. & Murray, W. *Two methods for the solution of linearly constrained and unconstrained optimization problems* (NPL Report NAC 25). Teddington, England: National Physics Laboratory, November, 1972.
9. Guttman, L. *Smallest space analysis by the absolute value principle*. Paper presented at the symposium on "Theory and practice of measurement" at the Nineteenth International Congress of Psychology, London, 1969.
10. Harshman, R. A. *Foundations of the PARAFAC procedure: Models and conditions for an explanatory multi-modal factor analysis* (Working Papers in Phonetics No. 16). Los Angeles: University of California, 1970.
11. Horst, P. *The prediction of personal adjustment* (Bulletin 48). New York: The Social Science Research Council, 1941.
12. Jacobowitz, D. *The acquisition of semantic structures*. Unpublished doctoral dissertation, University of North Carolina, 1975.
13. Jones, L. E. & Wadington, J. *Sensitivity of INDSCAL to simulated individual differences in dimension usage patterns and judgmental error*. Paper delivered at the Spring meeting of the Psychometric Society, Chicago, April, 1973.
14. Kruskal, J. B., Young, F. W., & Seery, J. B. *How to use KYST, a very flexible program to do multidimensional scaling and unfolding*. Unpublished manuscript, Bell Laboratories, 1973.
15. Obenchain, R. *Squared distance scaling as an alternative to principal components analysis*. Unpublished notes, Bell Laboratories, 1971.

16. Roskam, E. E. *Data theory and algorithms for nonmetric scaling* (parts 1 and 2). Unpublished manuscript, Catholic University, Nijmegen, The Netherlands, 1969.
17. Yates, A. *Nonmetric individual-differences multidimensional scaling with balanced least squares monotone regression.* Paper presented at the Spring meeting of Psychometric Society, Princeton, N. J., April, 1972.
18. Young, F. W. *Polynomial conjoint analysis: Some second order partial derivatives* (L. L. Thurstone Psychometric Laboratory Report, No. 108). Chapel Hill, North Carolina: The L. L. Thurstone Psychometric Laboratory, July, 197?.

REFERENCES

Bloxom, B. An alternative method of fitting a model of individual differences in multidimenlonal scaling. *Psychometrika*, 1974, *39*, 365–367.

Bôcher, M. *Introduction to higher algebra.* New York: MacMillan, 1907.

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*, 238–319.

Coombs, C. H. *A theory of data.* New York: Wiley, 1964.

de Leeuw, J. *Canonical analysis of categorical data.* Leiden, the Netherlands: University of Leiden, 1973.

de Leeuw, J., Young, F. W. & Takane, Y. Additive structure in qualitative data: An alternating least squares method with opitmal scaling features. *Psychometrika*, 1976, *41*, 471–503.

Eckart, C. & Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1936, *3*, 211–218.

Ekman, G. Dimensions of color vision. *Journal of Psychology*, 1954, *38*, 467–474.

Fisher, R. A. *Statistical methods for research workers* (10th ed.). Edinburgh: Oliver and Boyd, 1946.

Funk, S., Horowitz, A., Lipshitz, R. & Young, F. W. The perceived structure of American ethnic groups: The use of multidimensional scaling in stereotype research. *Sociometry*, in press.

Green, P. E. & Rao, V. R. *Applied multidimensional scaling: A comparison of approaches and algorithms.* New York: Holt, Rinehart and Winston, 1972.

Guttman, L. A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, 1968, *33*, 469–506.

Hageman, L. A. & Prosching, T. A. Aspects of nonlinear block successive overrelaxation. *SIAM Journal of Numerical Analysis*, 1975, *12*, 316–335.

Hayashi, C. Minimum dimension analysis. *Behaviormetrika*, 1974, *1*, 1–24.

Horan, C. B. Multidimensional scaling: Combining observations when individuals have different perceptual structures. *Psychometrika*, 1969, *34*, 139–165.

Johnson, R. M. Pairwise nonmetric multidimensional scaling. *Psychometrika*, 1973, *38*, 11–18.

Johnson. S. C. Hierarchical clustering schemes. *Psychometrika*, 1967, *32*, 241–254.

Jones, L. E. & Young, F. W. The structure of a social environment: A longitudinal individual differences scaling of an intact group. *Journal of Personality and Social Psychology*, 1972, *24*, 108–121.

Jöreskog, K. A general method for analysis of covariance structures. *Biometrika*, 1970, *57*, 239–251.

Kruakal, J. B. Nonmetric multidimensional scaling. *Psychometrika*, 1964, *29*, 1–27; 115–129.

Kruskal, J. B. & Carroll, J. D. Geometric models and badness-of-fit functions. In P. R. Krishnaiah, *Multivariate analysis* (Vol. 2), New York: Academic Press, Inc., 1969.

Lawson, C. L. & Hanson, R. J. *Solving least squares problems*. Englewood Cliffs, N. J.: Prentice-Hall, 1974.

Levin, J. Three-mode factor analysis. *Psychological Bulletin*, 1965, *64*, 442–452.

Levinsohn, J. R. & Young, F. W. Two special-purpose programs that perform nonmetric multidimensional scaling. *Journal of Marketing Research*, 1974, *11*, 315–316.

Lingoes, J. C. *The Guttman–Lingoes nonmetric program series*. Ann Arbor, Michigan: Mathesis Press, 1973.

Lingoes, J. C. & Roskam, E. E. A mathematical and empirical analysis of two multidimensional scaling algorithms. *Psychometrika Monograph Supplement*, 1973, *38* (4, Pt. 2).

McGee, V. C. Multidimensional scaling of *n* sets of similarity measures: A nonmetric individual differences approach. *Multivariate Behavioral Research*, 1968, *3*, 233–248.

Messick, S. J. & Abelson, R. P. The additive constant problem in multidimensional scaling. *Psychometrika*, 1956, *21*, 1–15.

Miller, G. A. & Nicely, P. E. An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 1953, *27*, 338–352.

Peterson, G. E. & Barney, H. L. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 1952, *24*, 175–184.

Schönemann, P. H. An algebraic solution for a class of subjective metric models. *Psychometrika*, 1972, *37*, 441–451.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. *I. Psychometrika*, 1962, *27*, 125–140. (a)

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 1962, *27*, 219–246. (b)

Shepard, R. N. Stimulus and response generalization: Tests of a model relating generalization of distance in psychological space. *Journal of Experimental Psychology*, 1958, *55*, 509–523.

Spence, I. A Monte Carlo evaluation of three nonmetric multidimensional scaling algorithms. *Psychometrika*, 1972, *37*, 461–486.

Stevens, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951.

Stoer, J. On the numerical solution of constrained least-squares problems. *SIAM Journal of Numerical Analysis*, 1971, *8*, 382–411.

Torgerson, W. S. Multidimensional scaling: I. Theory and method. *Psychometrika*, 1952, *17*, 401–419.

Tucker, L. R. Relations between multidimensional scaling and three-mode factor analysis. *Psychometrika*, 1972, *37*, 3–27.

Tucker, L. R. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 1966, *31*, 279–311.

Wilf, H. S. The numerical solution of polynomial equations. In A. Ralston & W. S. Wilf (Eds.), *Mathematical methods of digital computers* (Vol. 1). New York: Wiley, 1960.

Wold, H. & Lyttkens, E. Nonlinear iterative partial least squares (NIPALS) estimation procedures. *Bulletin ISI*, 1969, *43*, 29–47.

Yates, F. The analysis of replicated experiments when the field results are incomplete. *The Empire Journal of Experimental Agriculture*, 1933, *1*, 129–142.

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

Young, F. W. Nonmetric multidimensional scaling: Recovery of metric information. *Psychometrika*, 1970, *35*, 455–474.

Young, F. W. POLYCON: A program for multidimensionally scaling one-, two-, or three-way data in additive, difference, or multiplicative spaces. *Behavioral Science*, 1973, *18*, 152–155.

Young, F. W. Methods for describing ordinal data with cardinal models. *Journal of Mathematical Psychology*, 1975, *12*, 416–436. (a)

Young, F. W. Scaling replicated conditional rank-order data. *Sociological Methodology*, 1975, *12*, 129–170. (b)

Young, F. W., de Leeuw, J., & Takane, Y. Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 1976, *41*, 505–529.

Young, F. W. & Torgerson, W. S. TORSCA, a Fortran-IV program for nonmetric multidimensional scaling. *Behavioral Science*, 1968, *13*, 343–344.