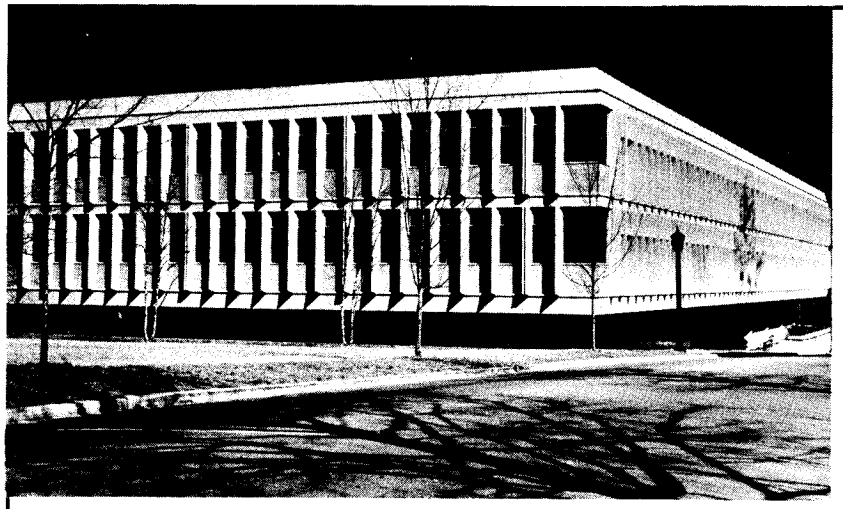


49

Nonmetric Individual Differences Multidimensional Scaling: An Alternating Least Squares Method with Optimal Scaling Features

Yoshio Takane

Forrest W. Young

and

Jan de Leeuw

December, 1975

Report Number 147

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

CHAPEL HILL, N. C. 27514

Abstract

A new procedure is discussed which fits either the weighted or simple Euclidian model to data which 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 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, non-oscilatory, quick, and relatively free from local minimum problems.

The procedure is evaluated via both monte carlo and empirical data with the conclusion being that it is 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.

Current addresses:

Jan de Leeuw, Datatheorie, Central Rekeninstituut, Wassenaarseweg 80, Leiden, The Netherlands.

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

This project was supported in part by Research Grant No. MH10006 and Research Grant No. MH26504, awarded by the National Institute of Mental Health, DHEW. We wish to thank Robert F. Baker, J. Douglas Carroll and Amnon Rapoport for comments on an earlier draft of this paper. Portions of the research reported here were presented to the spring meeting of the Psychometric Society, 1975.

1.0 Purpose and motivation

One of the most vigorous areas of endeavor in recent multidimensional scaling research concerns the representation of individual differences, with the weighted Euclidian model currently being the most widely used individual differences model of the various ones which 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 Euclidian space, while representing the uniqueness of each individual by differential weights attached to the dimensions of the space is an ingeneous idea particularly conducive to simple and straightforward interpretation. Furthermore, 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 and 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 which are much less severely restricted than those appropriate

to the Carroll-Chang procedure. Our procedure is appropriate to data which may have missing observations, which may be defined at the nominal, ordinal, interval or ratio measurement levels, which may be discrete or continuous, and which 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 and Chang (1970), McGee (1968), and 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 initial proposal of the weighted Euclidian model and the associated procedures for fitting the model to empirical data were made by several people at about the same time (Horan, 1969; Bloxom, 1968; Carroll and Chang, 1970), with the most successful procedure and the most complete proposal being that of Carroll and Chang. Their INDSCAL (individual differences scaling) 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 dissimilarities 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 (1975a), which

have the desireable property that they are 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 (1968) 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 and 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 is 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 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 (1974) 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 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 (1974) proposed another nonmetric procedure to minimize STRAIN which uses an ALS method after initial estimates of $\underline{W}$ and $\underline{X}$ 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, and 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, at least theoretically, the soundest, 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 $\underline{W}$ and $\underline{X}$ 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, with a careful choice

of the initial orientation of $\underline{X}$ being required. Although this difficulty could be remedied by using de Leeuw's (1974) 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 (1972) for nonmetrically fitting the weighted Euclidian model which is in neither the STRESS or STRAIN framework, but which 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, 1969; de Leeuw, 1970; 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.

2.0 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.

2.1 Individual differences models

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

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

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 Eq. (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 Eq. (1) in section 5.2 of the paper.

2.2 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' (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' 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).

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 cornerstones of our own work.

Let us define the squared observations $\underline{O}$, the optimally scaled squared observations $\underline{D}^*$, and the squared distances $\underline{D}$. (The optimally scaled squared observations are commonly referred to as the disparities in the MDS context, and 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, $\underline{O}$ is a collection of all matrices $\underline{O}_i$ for all individuals i from whom we have obtained observations o_{ijk} about stimulus pairs (j,k). Correspondingly, $\underline{D}^*$ is the collection of matrices $\underline{D}^*_i$ with elements d^{*2}_{ijk}, and $\underline{D}$ is the collection of all matrices $\underline{D}$ with elements d^2_{ijk} defined by Eq. (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}

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

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 of the paper 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. Put 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, and process 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. In the former case the discrete nature of the process is reflected by the fact that we choose a single (discrete) number to represent all observations in the category; whereas in the latter case, the continuity of the process is

reflected by the fact that we choose real numbers from a closed (continuous) interval of real numbers. Formally, we define the two restrictions as follows: the discrete restriction is

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

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

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

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, but to denote this would involve a somewhat more complicated notational system. Note also that for all observations in a particular category the corresponding rescaled observations are required to fall in the interval but need not be equal.

We now turn to the 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' 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:

(5) $\quad t^{o}: \quad (o_{ijk} < o_{mno}) \rightarrow (d^{*}_{ijk} \leq d^{*}_{mno})$

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:

(6) $\quad t^{\ell}: \quad d^{*}_{ijk} = \delta_{o} + \delta_{1} o_{ijk}$

(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) $\quad t^{p}: \quad d^{*}_{ijk} = \sum_{q=0}^{p} \delta_{q} o^{q}_{ijk}$

(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, it is usually the case that we are unwilling to say that one subject's judgment of 7 (on a similarity scale of 1 through 9, let's say) 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 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, although we do not explicate these modifications as they are lengthly 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 & Seery (1973).

2.3 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

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

where $\underline{P}^*$ is the collection of $\underline{P}_i^*$ for i=1,...,N, where $\underline{W}_i$ is a diagonal matrix of weights for subject i, and where $\underline{P}_i^*$ is the matrix of scalar products derived from subject i's dissimilarities under either metric or nonmetric assumptions.

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 straight-forward 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 (1974) assumes that the observed dissimilarities must be monotonic with a set of values from which the scalar products $\underline{P}_i^*$ are computed. That is, it is required that

$$(9) \qquad t^{\circ}[o_{ijk}] = [d^*_{ijk}],$$

so that $\underline{P}_i^*$ may be computed from $\underline{D}_i^*$ 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 $\underline{P}_i$ directly from the raw observations at the outset of the analysis. The procedure then optimizes STRAIN under the assumption that $\underline{P}_i$ is nonmetric. That is, this procedure requires that

$$(10) \qquad t^{\circ}[p_{ijk}] = [p^*_{ijk}] \quad .$$

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

$$(11) \qquad \phi^2(\underline{X},\underline{W},\underline{D}^*) = \sum_i \sum_j \sum_k^{j-1} (d^{*2}_{ijk} - d^2_{ijk})^2,$$

where d^{*2}_{ijk} is an element of $\underline{D}^*_i$, where d^2_{ijk} is defined by Eq. (1) and where (11) is subject to suitable normalization conditions. Since (11) is in the STRESS framework, but differs due to being defined on squared distances d^2_{ijk} and squared estimates d^{*2}_{ijk}, (note that d^{*2}_{ijk} is the least squares estimate of d^2_{ijk}, not the square of the least squares estimate of d_{ijk}) we refer to the formula as SSTRESS. Hayashi (1974) and Obenchain (1971) have developed multidimensional scaling procedures within the SSTRESS framework, and Young (1972b) has discussed the index.

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

$$t^o[o^2_{ijk}] = [d^{*2}_{ijk}]$$

defined on o^2_{ijk} and d^{*2}_{ijk} 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^2_{ijk} and d^{*2}_{ijk}. We will investigate this inconvenience more later on, but suffice it to say here that this difficulty can be surmounted, allowing us to state that the measurement restrictions

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

and

(12') $\quad t[o^2_{ijk}] = [d^{*2}_{ijk}]$

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. Let's say that we have the following two cases:

(A) $d_{ijk}=2 \quad d^*_{ijk}=1$

(B) $d_{ijk}=5 \quad d^*_{ijk}=6.$

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) $d_{ijk}=5 \quad d^*_{ijk}=4$

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^2_{ijk}. However, with SSTRESS we have

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

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

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 is not straightforward (though algebraically tractable) and not entirely illuminating, since 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.

3.0 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, 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 (1941), and follow its development through many researchers, culminating in the NILES/NIPOLS work of Wold and associates (Wold and 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 to estimate 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 (1969) in the nonmetric principal components situation, Young (1972) for initial values in the polynomial conjoint scaling situation, de Leeuw (1975b) for the canonical analysis of categorical data, de Leeuw, Young & Takane for nonmetric ANOVA (1975) and Young,

de Leeuw & Takane for nonmetric multiple and canonical regression (1975). 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 $\underline{D}^*$ under the assumption that the configuration $\underline{X}$ and the weights $\underline{W}$ are constants. That is, we solve the conditional least squares problem which minimizes SSTRESS (Eq. 11) under the condition that $\underline{X}$ and $\underline{W}$ are not variables. Notationally, we indicate this as $\underset{\underline{D}^*}{\text{MIN}}[\phi^2(\underline{D}^*|\underline{X},\underline{W})]$. The second major phase involves two separate minimization subphases, the first solving the problem $\underset{\underline{W}}{\text{MIN}}[\phi^2(\underline{W}|\underline{X},\underline{D}^*)]$ and the second the problem $\underset{\underline{X}}{\text{MIN}}[\phi^2(\underline{X}|\underline{W},\underline{D}^*)]$. 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 $\underline{X}$ and $\underline{W}$ directly from $\underline{O}$ using a modification of Schonemann's algebraic solution.

1. Optimal Scaling Phase

   1.1 Calculate the squared weighted Euclidian distances $\underline{D}$ using $\underline{X}$ and $\underline{W}$.

   1.2 Normalize appropriately.

   1.3 Obtain the optimally scaled (least squares estimated) disparities $\underline{D}^*$ from the normalized $\underline{D}$, the observations $\underline{O}$, and the relevant measurement restrictions. Use the de Leeuw, Young & Takane method.

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 $\underline{W}$ from the old $\underline{X}$ and the new $\underline{D}^*$ (from step 1.3) by regression techniques.

3.2 Impose nonnegativity constraints on $\underline{W}$, if necessary, by an ALS technique developed here.

3.3 Calculate the new least squares estimate of the configuration $\underline{X}$ from the new weights just calculated in steps 3.1 and 3.2 and the $\underline{D}^*$ computed in step 1.3, by using Gill & 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.

3.1 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 Eq. (8) for the error-free ratio measurement level case.

Let us suppose that there are N scalar product matrices $\underline{P}_i$ (one for each of the N subjects i) of order n (there are n stimuli) which satisfy

$$\underline{P}_i = \underline{X}\underline{W}_i\underline{X}' \tag{13}$$

where the symbols are defined as in Eq. (8) (recall that $\underline{W}_i$ is a diagonal matrix of weights for subject i, whereas $\underline{W}$ is a rectangular matrix of weights for all subjects). The problem is to recover $\underline{X}$ and $\underline{W}_i$ from the $\underline{P}_i$, under the assumption that $\underline{X}$ is of full column rank, and that the diagonal elements of $\underline{W}_i$ are strictly positive for at least one subject. For any non-singular diagonal matrix T of order t (there are t dimensions) we have

$$\underline{P}_i = \underline{X}\underline{T}(\underline{T}^{-1}\underline{W}_i\underline{T}^{-1})\underline{T}\underline{X}' \tag{14}$$

and consequently we must make some restriction on the size of the $\underline{W}_i$ for identification purposes. Thus, we assume that

$$\underline{D}_\cdot = \frac{1}{N}\sum_{i=1}^{N}\underline{W}_i$$

is equal to the identity, implying that $\underline{P}_{\cdot} = \underline{X}\underline{X}'$ (where $\underline{P}_{\cdot}$ is the average $\underline{P}_i$). 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 $\underline{P}_{\cdot\cdot}$. Call this arbitrary solution $\underline{Y}$. It follows that

$$\underline{X} = \underline{Y}\underline{K},$$

with $\underline{K}$ a rotation matrix. We also know that

$$(\underline{X}'\underline{X})^{-1}\underline{X}'\underline{P}_i\underline{X}(\underline{X}'\underline{X})^{-1} = \underline{W}_i$$

should be diagonal for each i. It follows that we should select our rotation $\underline{K}$ in such a way that

$$\underline{K}'(\underline{Y}'\underline{Y})^{-1}\underline{Y}'\underline{P}_i\underline{Y}(\underline{Y}'\underline{Y})^{-1}\underline{K} = \underline{W}_i \tag{15}$$

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

Let

$$\underline{C}_i = (\underline{Y}'\underline{Y})^{-1}\underline{Y}'\underline{P}_i\underline{Y}(\underline{Y}'\underline{Y})^{-1}. \tag{16}$$

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

$$\underline{C}_e = \sum_{i=1}^{N} \underline{e}_i\underline{C}_i \tag{17}$$

to find $\underline{K}$ and compute $\underline{W}_i$ from (15). Thus we have obtained the configuration $\underline{X}$, and the weights $\underline{W}$. 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 $\underline{e}$ 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 Eq. (13). In the fallible case in which the relationship is only approximately true we need to make two choices. First, we need to define $\underline{P}_.$, and second we need to define $\underline{e}$. The first problem is quite easily solved by simply double centering the elements of each data matrix $\underline{O}_i$ with elements o^2_{ijk} (and dividing by -2) to obtain a matrix $\underline{P}_i$ of scalar products for each subject, and then averaging over subjects to obtain $\underline{P}_.$, which can be decomposed into

$$\underline{P}_. = \underline{Y}\underline{Y}' \quad , \tag{18}$$

to obtain $\underline{Y}$, the arbitrarily oriented configuration which best reproduces the averaged scalar products. Note that a) if the data are asymmetric we average over the triangular portions of each matrix 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) that 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 $\underline{K}$ which simultaneously diagonalizes the matrices $\underline{C}_i$ as much as possible. The method suggested by de Leeuw & Pruzansky (1975) is used.

Since the procedure just outlined assumes that the data are metric, it is possible to obtain negative weights, espe-

cially 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 (Eq. 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.

3.2 Optimal Scaling Phase

In the optimal scaling phase we wish to optimally scale the squared observations $\underline{O}$ to obtain the disparities $\underline{D}^*$ which a) meet the selected measurement restrictions, and b) are least squares estimates of the squared distances $\underline{D}$, 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: 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 $\underline{D}^*$ are free to vary, with the stimulus configuration $\underline{X}$ and

the subject weights $\underline{W}$ being held constant. Thus we solve the conditional least squares problem $\underset{\underline{D}^*}{\text{MIN}}[\phi^2(\underline{D}^*|\underline{X},\underline{W})]$.

3.2.1 Compute distances

The first step in the optimal scaling phase is to compute the $\underline{D}_i$ from the current $\underline{X}$ and $\underline{W}$ by Eq. (1).

3.2.2 Normalize

The second step in the optimal scaling phase is to normalize the model space. As has been discussed by Carroll & Chang (1970) two of the three aspects of the optimization problem represented by Eqs. (1) and (11) (the data, the weights, and the configuration) must be normalized, with the remaining aspect being left unnormalized. While the choice is arbitrary, and the actual details of the normalization are also arbitrary, we choose to continue the conventions adopted by Carroll & Chang. Sepcifically, 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 & Chang normalize the data, we normalize the distances, which is equivalent. Our reason for not normalizing the data is to permit analysis of qualitative as well as quantitative data (it is difficult to normalize qualitative data). It would seem that the most compelling alternative would be to normalize the optimally scaled data (which are quantitative), however, the relationship between the optimally scaled data $\underline{D}^*$ and the distances $\underline{D}$ is such that the choice is completely

arbitrary (Kruskal & Carroll, 1969; Young, 1972). Thus, the squared distances are normalized so that their sum of squares (i.e., the sum of each distance raised to the fourth power) is $N[n(n-1)/2]$.

There is one additional normalization consideration, and that is the conditionality of the data. If the data are unconditional (all observations are comparable) then we normalize as stated. However, if the data are conditional (either matrix or row conditional) then we must normalize each subject's distances separately since there is no way to compare between subjects. Thus, in these cases we normalize so that the sum of squares of the squared distances equals $n(n-1)/2$ for each matrix. Note that the conditionality of the data and the resulting difference in normalization has certain implications for interpretating the weights. These implications are discussed in section 5.1.

Either of the above normalizations permit the optimization of the normalized loss function

$$(20) \qquad \phi'^2(\underline{X},\underline{W},\underline{D}^*) = \phi^2(\underline{X},\underline{W},\underline{D}^*) \Big/ \sum_{n}^{N} \sum_{j}^{n} \sum_{k}^{j-1} d_{ijk}^4$$

while actually operating on the unnormalized loss function (Eq. 11), as has been discussed by de Leeuw, Young & Takane (1975). This characteristic is very convenient, since we do not have to deal directly with the normalized function (which is the ratio of two biquadratic forms) whose partial derivatives are considerably more complicated than those for the unnormalized function.

As was briefly mentioned in section 2.3, Hayashi (1974) has proposed a multidimensional scaling procedure which is within the SSTRESS framework. His proposal is not,

however, to minimize an index precisely equivalent to our index (Eq. 22), but rather one which divides the sum of squared differences by a factor proportional to the variance of the squared distances. That is, Hayashi's method optimizes

$$(21) \qquad \xi^2(\underline{X},\underline{W},\underline{D}^*) = \phi^2(\underline{X},\underline{W},\underline{D}^*) \Big/ \sum_i^N \sum_j^n \sum_k^{j-1} (d^2_{ijk} - \bar{d}^2)^2$$

where $\bar{d}^2$ indicates the mean squared distance. (Hayashi's method is actually restricted to non-individual differences models, so he assumes that $w_{ia}=1$ for all i and a, and that N=1). If we compare Hayashi's function with our function (Eq. 20) we note that they differ in a way which parallels the differences between Kruskal's STRESS formulas 1 and 2 (Kruskal & Carroll, 1969) which are normalized, respectively, by the sum of squared distances and the variance of the distances. Since our normalization is the sum of squares of the squared distances, and Hayashi's is the variance of the squared distances, it would be appropriate to refer to our formula as SSTRESS formula 1, and Hayashi's as SSTRESS formula 2.

3.2.3 Optimal Scaling

The third and final step in the optimal scaling phase is to actually perform the optimal scaling. For all of the various restrictions the transformation can be defined as a linear transformation of the squared distances. That is,

$$(22) \qquad d^{*2}_{ijk} = t(d^2_{ijk}),$$

where t now indicates a linear transformation paralleling the measurement restrictions used to define t earlier. Furthermore, t defines d^{*2}_{ijk} so that SSTRESS (Eq. 11) is minimized for fixed values of $\underline{W}$ and $\underline{X}$, in a least squares sense. (For notational convenience we use the same symbol t. However, there is a definite difference in the two usages of t. Whereas t was used as a functional relationship between observations

and disparities, it is used here as the transformation which relates distances to disparities.)

We do not discuss the specific features of these transformations here since a detailed account is already presented in an earlier paper (de Leeuw, Young & Takane, 1975b). Instead, we present here a simplified characterization of t using matrix notation. Since we are regressing d^2_{ijk} onto o^2_{ijk} in the least squares sense under the various measurement restrictions mentioned above, t may be represented by a projection operator of the form

$$t: \quad \underline{E} = \underline{Z}(\underline{Z}'\underline{Z})^{-1}\underline{Z}' \tag{23}$$

where $\underline{Z}$ is, in general, a matrix of vectors defining the space onto which the vector of d^2_{ijk} is regressed.

For the ratio transformation t^r $\underline{Z}$ is simply the vector $\underline{O}$ of squared observations. For the interval transformation t^i $\underline{Z}$ 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 $\underline{Z}$ 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 $\underline{Z}$ indicates that distances which correspond to categorically equivalent observations are to be tied. The obvious least squares estimates in this case simply involves category means. Finally, for the continuous-nominal case the

matrix $\underline{Z}$ 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 & Takane's (1975a) pseudo-ordinal procedure to determine the optimal boundaries.

Note that for some transformations $\underline{Z}$ is known before the analysis is made, and in other cases it is not. Specifically, for all discrete transformations except the discrete-ordinal transformation $\underline{Z}$ is known *a priori*, and for the remainder $\underline{Z}$ is only known after the analysis is made. Furthermore, in these cases $\underline{Z}$ varies from iteration to iteration depending on the nature of the distances.

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 Eq. (23). This means that if we define a column vector $\underline{d}$ containing the $Nn(n-1)/2$ elements d^2_{ijk} and another column vector $\underline{d}^*$ containing the corresponding elements d^{*2}_{ijk}, then we can make the important observation that

$$\underline{d}^* = \underline{E}\underline{d}. \tag{24}$$

Furthermore, this equation, which is implicitly in terms of unconditional data, can be easily extended to conditional data. For matrix-conditional data we define $\underline{Z}_i$ for each individual separately and then construct a block-diagonal supermatrix $\underline{Z}$ with the $\underline{Z}_i$'s on the diagonal. For row-conditional data we define $\underline{Z}_{ij}$ for every row of every individual's data matrix and then construct the block-diagonal supermatrix $\underline{Z}$ with these $\underline{Z}_{ij}$'s on the diagonal.

In both cases $\underline{E}$ remains defined as before, thus the projection operator notion and Eq. (24) 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 Eq. (24) is that we can now easily express SSTRESS entirely in matrix notation, and entirely in terms of the distances. If we define $\tilde{\underline{E}} = \underline{I} - \underline{E}$, then SSTRESS (Eq. 11) can be rewritten as

$$(25) \qquad \phi^2(\underline{X},\underline{W},\underline{D}^*) = \underline{d}'\tilde{\underline{E}}\underline{d}$$

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

$$(26) \qquad \phi'^2(\underline{X},\underline{W},\underline{D}^*) = \underline{d}'\tilde{\underline{E}}\underline{d} \;/\; \underline{d}'\underline{d}$$
$$= \underline{d}'\tilde{\underline{E}}(\underline{d}'\underline{d})^{-1}\tilde{\underline{E}}\underline{d} \; .$$

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

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 & Abelson, 1956). The problem is as follows. When we assume that the observations are defined at the interval level this means that

$$(27) \qquad d^*_{ijk} = a(o_{ijk}) + b,$$

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

$$d^{*2}_{ijk} = a^2(o_{ijk})^2 + 2ab(o_{ijk}) + b^2 \quad , \tag{28}$$

which is clearly different from the simple regression problem of d^2_{ijk} to o^2_{ijk}, the assumption implied in a linear relationship between d^2_{ijk} and o^2_{ijk} (unless b=0 as in the ratio case).

While it is possible to directly solve Eq. (28), it is much simpler to redefine the problem as

$$d^{*2}_{ijk} = \alpha + \beta(o_{ijk}) + \gamma(o_{ijk})^2 \tag{29}$$

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

$$\beta^2 = 4\alpha\gamma \quad . \tag{30}$$

We now introduce three definitions. First we define the parameter vector $\underline{\chi}' = [\alpha,\beta,\gamma]$. 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 $\underline{O}$ (note that this is not the same $\underline{O}$ as used in other sections of the paper). Finally, we define a column vector $\underline{d}$ having the N[n(n-1)/2] elements d^2_{ijk} arranged in the same manner as the o_{ijk} in $\underline{O}$.

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

$$\phi^2(\underline{\chi},\lambda|\underline{O},\underline{d}) = (\underline{d}-\underline{O}\underline{\chi})'(\underline{d}-\underline{O}\underline{\chi}) + \lambda(\beta^2-4\alpha\gamma) \tag{31}$$

which we seek to minimize by solving for $\underline{\chi}$ and λ (the Lagrangian multiplier).

The least squares estimate for the constrainted parameters is

$$\hat{\underline{\chi}} = (\underline{O}'\underline{O})^{-1}\underline{O}'\underline{d} \quad . \tag{32}$$

To solve for the Lagrangian multiplier we define

$$(\underline{O}'\underline{O})^{-1}\underline{g} = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} \quad , \tag{33}$$

where $\underline{g}' = [-2\gamma, \beta, -2\alpha]$ is the derivatives of Eq. (30).

Then we must solve

$$(\hat{\beta}+\lambda q_2)^2 = 4(\hat{\alpha}+\lambda q_1)(\hat{\gamma}+\lambda q_3) \quad . \tag{34}$$

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

3.3 Termination phase.

The termination phase is extremely simple. We must only determine the value of SSTRESS on the current iteration (Eq. 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 & Takane, 1975).

3.4 Model estimation phase.

Whereas in the optimal scaling phase we solved the conditional least squares problem $\underset{\underline{D}^*}{\text{MIN}}[\phi^2(\underline{D}^*|\underline{X},\underline{W})]$, in the model estimation phase we solve two conditional least squares problems successively. The first subphase solves the conditional least squares problem $\underset{\underline{W}}{\text{MIN}}[\phi^2(\underline{W}|\underline{X},\underline{D}^*)]$, whereas the second subphase

solves the problem $\underset{\underline{X}}{\text{MIN}}[\phi^2(\underline{X}|\underline{W},\underline{D}^*)]$. In this section we discuss both of these problems.

3.4.1 Compute weights

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

$$\phi^2(\underline{Y},\underline{W},\underline{D}^*) = \text{tr}(\underline{D}^*-\underline{WY}')'(\underline{D}^*-\underline{WY}') \quad , \tag{35}$$

from which we see that the least squares estimates of $\underline{W}$ are

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

3.4.2 Nonnegativity weight constraint

There is one difficulty using the regression approach just outlined for obtaining $\underline{W}$: The non-negativity constraints placed on the weights (Eq. 1) may be violated. Thus we now turn to a discussion of a way to incorporate 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 prosented in Eq. (36) is independent for each individual. That is, the values estimated for the weights for one individual do not effect the estimated values of the weights for any other individual. This can be seen from the fact that SSTRESS (Eq. 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 from those for the others we can impose non-negativity 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 from 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 $\underline{W}$ by Eq. (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 (the constrained least squares estimate under the condition that all the other weights are constant), and then, for another weight, re-estimate its value under the assumption that all other weights are constant. The conditional least squares estimate for a single weight is

$$(37) \qquad w_{ia} = (\underline{d}_i^* - \sum_{b \neq a} w_{ib}\underline{y}_b)'\underline{y}_a / (\underline{y}_a'\underline{y}_a) \quad ,$$

where $\underline{y}_a$ is the a'th column of matrix $\underline{Y}$ (Eq. 35) which contains 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.

3.4.3 Compute coordinates

The second subphase of the model estimation phase is to determine the stimulus coordinates $\underline{X}$. This subphase is somewhat more complicated than the weight estimation subphase since the partial derivatives of SSTRESS with respect to the elements of $\underline{X}$ are not linear in the x_{ja}'s. Rather, SSTRESS is quartic in the x_{ja}'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 from a theoretical point of view, would be to analytically solve the system of m=n*t simultaneous cubic equations for the m unknowns, as has been suggested by Obenchain (1971). 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_{ja} 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 $\underline{W}$ and $\underline{D}^*$) 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, although 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, and it seems reasonable to assume that the quickest method lies somewhere in between the two extremes. That is, it may be best to estimate a block of x_{ja}'s simultaneously, making sure that the number of coordinates being simultaneously estimated is not so large that it slows down the entire process so much that 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_{ja} $(a=1,\ldots,t)$ simultaneously for a specific j and successively for each stimulus j $(j=1,\ldots,n)$. This is the same 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 $\underset{\underline{x}_j}{\text{MIN}}[\phi(\underline{x}_j|\underline{x}_k,\underline{W},\underline{D}^*),\ (j\neq k)]$. Thus our approach is to place the Newton-Raphson method within the ALS framework to solve a single quartic equation, again demonstrating 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 since 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 & Murray (1972) which ensures the positive definiteness of the Hessian at the current point, thus ensuring that we are proceeding in a down hill 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 & Murray) deliberately chosen values are added to the diagonal to force it to be positive definite, thus avoiding convergence to a maximum or to some other stationary point which is not a minimum.

We now provide the first and second derivatives of SSTRESS with respect to x_{ja} for fixed j and $a=1,\ldots,t$. To simplify the

derivation we note that

$$(38) \qquad d^{*2}_{ijk} - d^{2}_{ijk} = d^{*2}_{ijk} - \sum_{a}^{t} w_{ia}(x_{ja}-x_{ka})^2$$

$$= d^{*2}_{ijk} - \sum_{a}^{t} w_{ia}x^{2}_{ja} + 2\sum_{a}^{t} w_{ia}x_{ja}x_{ka} - \sum_{a}^{t} w_{ia}x^{2}_{ka}$$

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

$$(39) \qquad h_{ijk} = d^{*2}_{ijk} - \sum_{a}^{t} w_{ia}x^{2}_{ka} \quad .$$

We organize these terms into a vector $\underline{h}_j$ which contains all h_{ijk} for fixed j and for $k \neq j$ (this vector has N(n-1) elements). We also define a supermatrix $\underline{G} = [\underline{G}_1, \underline{G}_2]$, with N(n-1) rows and 2t columns. The two submatrices are defined as follows:

$$(40) \qquad \underline{G}_1 = -2 \begin{bmatrix} w_{11}x_{11} & \cdots & w_{1t}x_{1t} \\ \vdots & & \vdots \\ w_{11}x_{k1} & \cdots & w_{1t}x_{kt} \\ \vdots & & \vdots \\ w_{11}x_{n1} & \cdots & w_{1t}x_{nt} \\ \vdots & & \vdots \\ w_{N1}x_{n1} & \cdots & w_{Nt}x_{nt} \end{bmatrix} , \quad k \neq j ,$$

and

$$(41) \qquad \underline{G}_2 = \begin{bmatrix} w_{11}\underline{u} & \cdots & w_{1t}\underline{u} \\ \vdots & & \vdots \\ w_{N1}\underline{u} & \cdots & w_{N1}\underline{u} \end{bmatrix} ,$$

where $\underline{u}$ is an n-1 component column vector of unities. We also define a 2t component supervector $\underline{\alpha}_j$ consisting of a vector $\underline{x}_j$ of coordinates of stimulus j on t dimensions, and a vector whose elements are the squares of the elements in $\underline{x}_j$. Since it is possible to express the squared elements as the product of a diagonal matrix and a vector, we further define $\underline{X}_j$ 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 $\underline{X}$). Then

$$(42) \qquad \underline{\alpha}_j = \begin{bmatrix} \underline{x}_j \\ \underline{X}_j\underline{x}_j \end{bmatrix} .$$

We can now define SSTRESS as

$$(43) \qquad \phi^2(\underline{X},\underline{W},\underline{D}^*) = \frac{1}{2}\sum_j^n(\underline{h}_j-\underline{G\alpha}_j)'(\underline{h}_j-\underline{G\alpha}_j).$$

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

The gradient vector (first derivatives of SSTRESS with respect to $\underline{x}_j$) can be expressed as

$$(44) \qquad \underline{g} = -[\underline{I},2\underline{X}_j]\underline{G}'\underline{h}_j + [\underline{I},2\underline{X}_j]\underline{G}'\underline{G\alpha}_j$$

$$= [\underline{G}_1'\underline{G}_1 + 2\underline{X}_j\underline{G}_2'\underline{G}_1 + \underline{G}_1'\underline{G}_2\underline{X}_j + 2\underline{X}_j\underline{G}_2'\underline{G}_2\underline{X}_j]\underline{x}_j$$

$$- [\underline{G}_1\underline{h}_j + 2\underline{X}_j\underline{G}_2\underline{h}_j] \quad .$$

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

$$(45) \qquad h_{ab} = [\underline{e}_a', 2x_{ja}\underline{e}_a']\underline{G}'\underline{G}\begin{bmatrix} \underline{e}_b \\ x_{jb}\underline{e}_b \end{bmatrix} \quad ,$$

where $\underline{e}_a$ is a vector with unity in the a'th position and zeros elsewhere. The a'th diagonal element of the Hessian is

$$(46) \qquad h_{aa} = -[\underline{0},\ 2\underline{e}_a']\underline{G}'\underline{h}_j + [\underline{e}_a',\ 2x_{ja}\underline{e}_a']\underline{G}'\underline{G}\begin{bmatrix}\underline{e}_a \\ 2x_{ja}\underline{e}_a'\end{bmatrix} + [\underline{0}',\ 2\underline{e}']\underline{G}'\underline{G}\begin{bmatrix}\underline{x}_j \\ \underline{X}_j\underline{x}_j\end{bmatrix}$$

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

$$(47) \qquad \underline{x}_j^{(\ell)} = \underline{x}_j^{(\ell-1)} - \theta\tilde{\underline{H}}^{(\ell-1)^{-1}}\underline{g}^{(\ell-1)}$$

where θ is a stepsize determined to ensure that $\phi^{2(\ell)} < \phi^{2(\ell-1)}$, where $\tilde{\underline{H}}=\underline{H}$ when $\underline{H}$ is positive definite, and where $\tilde{\underline{H}}=\underline{H}+\underline{F}$ for $\underline{F}$ a diagonal matrix with positive diagonal values when $\underline{H}$ is not positive definite. The matrix $\underline{F}$ is determined according to Gill & Murray's developments. While it is the case that SSTRESS must be evaluated several times in determining the estimate of $\underline{x}$, each point's coordinate vector, it is a very simple and quick evaluation since the optimal scaling $\underline{D}^*$ 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' (1972) 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 then defines the last step of a single

iteration, and 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 (Eq. 43) is a quartic equation; therefore its gradient (Eq. 44) is a cubic which may have <u>two</u> minima. However, the procedure we have proposed converges on one of the minima without ensuring that it is the optimal one. While numerical analysis results indicate that we will most often converge on the optimal minimum(especially if the two minima have rather different function values), we will at least occasionally converge on the non-optimal minimum. Alternative procedures could be proposed which would circumvent the necessity of checking the function at both minima, with one possible procedure being along the line of the procedure proposed in section 3.2.3 for optimal scaling with the interval level of measurement. We have not yet investigated such a procedure, however, and are of the opinion that the 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 estimation process is for 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 difficulty has little practical effect.

4.0 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.

4.1 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 & Waddington, 1973) 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 choose 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 shown 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, a situation which, in our experience and in the experience of Carroll (personal communication) is optimal for obtaining a robust

Insert following p. 45

Table 1-C

Values of $\gamma^{(\ell)}$

error level	dimension	
ℓ	I	II
1	0	0
2	.180	.286
3	.600	.953

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

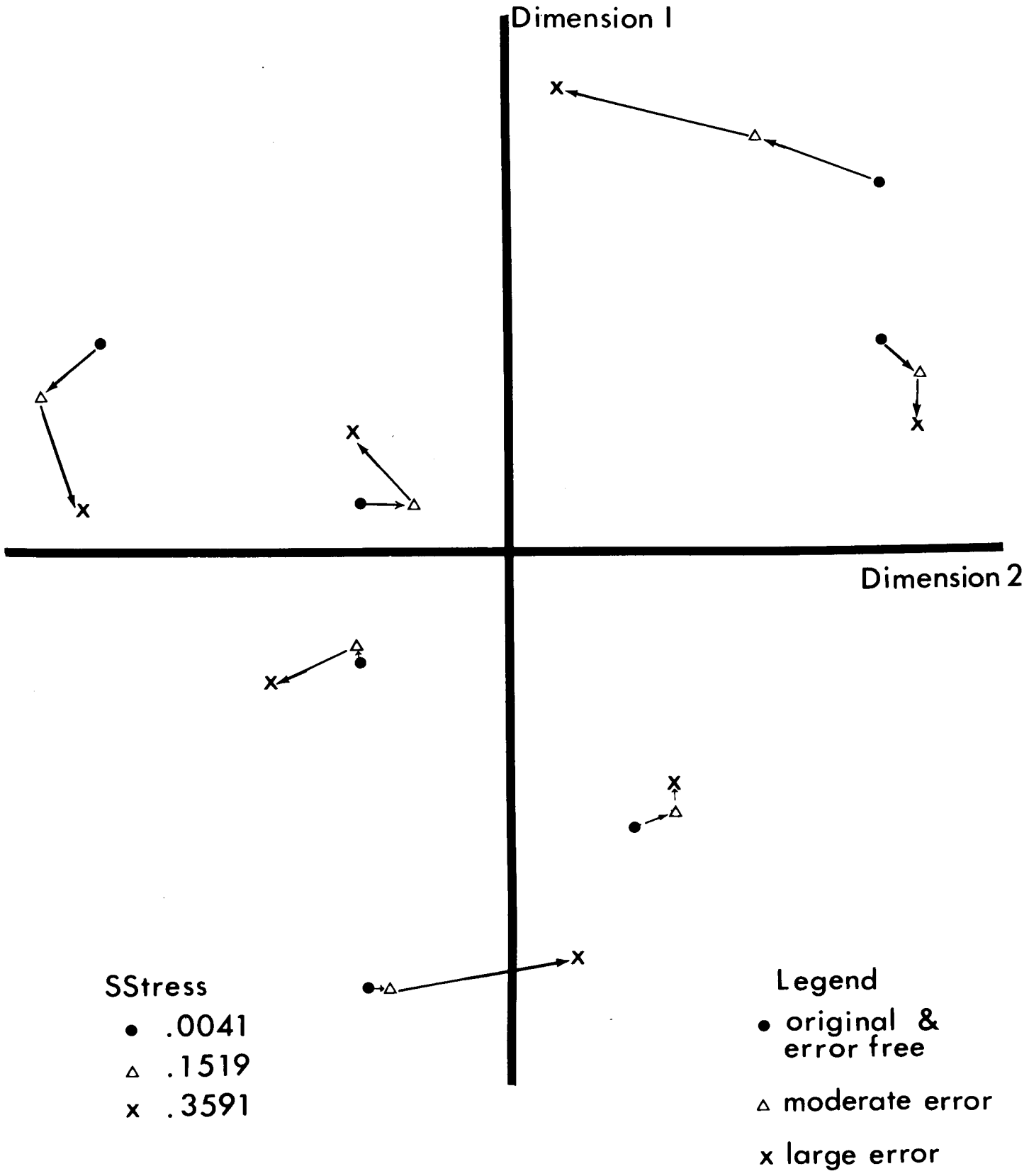


Figure 1. Monte Carlo study: Effects on the stimulus configuration when the data are assumed to be ordinal.

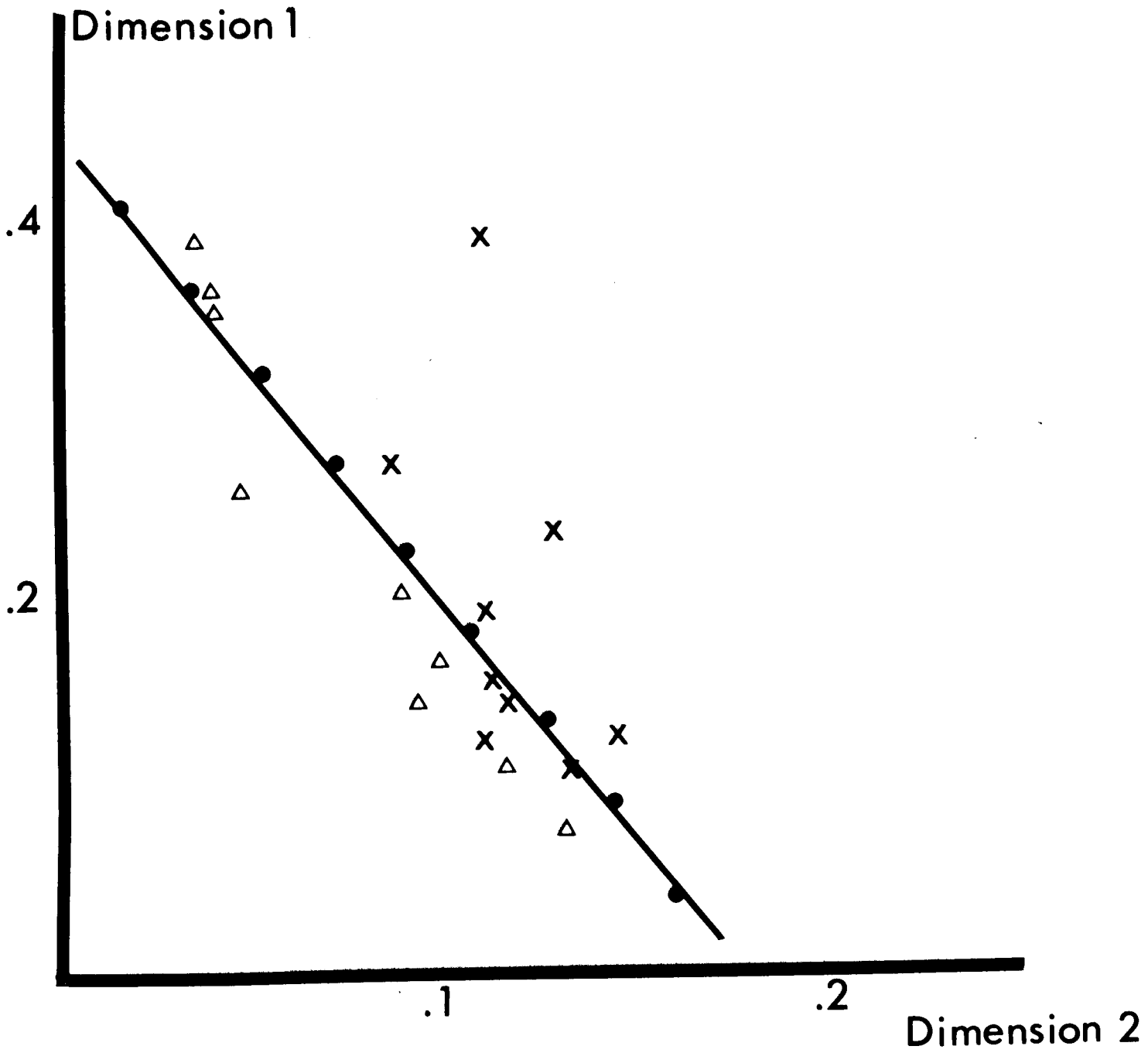


Figure 2. Monte Carlo study: Effects on the weight space when the data are assumed ordinal (symbols as in Fig. 1).

and meaningful analysis with INDSCAL. Note also that subjects generally attach relatively more importance to dimension I than 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 choose to follow the procedure of Young (1970) in which independent random normal error is added to the stimulus coordinates, with such error being generated anew for each pair of stimuli. Thus d^2_{ijk}, under the ℓ^{th} degree of error perturbation, is generated by

$$(48) \qquad d^{(\ell)}_{ijk} = \left[\sum_{a=1}^{t} w_{ka}(x_{ia}-x_{ja}+(z_{ijka})\gamma^{(\ell)}_a)^2\right]^{1/2}$$

where $z_{ijka}=z_{ija}-z_{ika}$, where $z_{ija}\sim N(0,1)$ $(i,j=1,\ldots,7),(a=1,2)$ and where $\gamma^{(\ell)}_a$ is a parameter specifying the variability of the errors. Note that $d^{(\ell)2}_{ijk}$ does not follow the noncentral chi-squares distribution (as it does in Young (1970)) since the variability is different across dimensions ($\gamma^{(\ell)}_a$ 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 $\gamma^{(\ell)}_a$ are shown in Table 1-c. Since z_{ija} and z_{ika} are independent, the variance of $(z_{ijka})\gamma^{(\ell)}_a$ is $2(\gamma^{(\ell)}_a)$. Note that the stimulus configuration x_{ja}'s are standardized so that they have unit variances for both dimensions. We refer to the case when $\ell=1$ as the error free case, $\ell=2$ as the moderate error case, and $\ell=3$ as the large error case. (For $\ell=3$ the error variance is .720

for dimension I and 1.816 for dimension II which is much larger than used in most Monte Carlo studies).

Next we introduce systematic monotonic error by either squaring the randomly perturbed distances in equation (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 from the error-free dissimilarities. The structure, which is indistinguishable from the true structure, is presented in the 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.

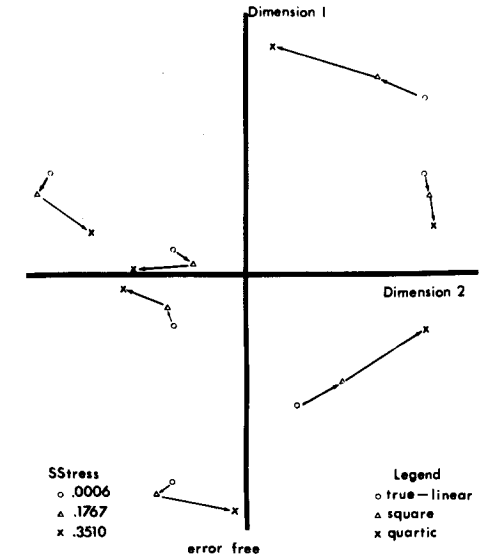
Figures 1 and 2 also display the structures derived by ALSCAL when there is moderate random error (the triangles) and when there is large random error (the squares). Note that there is, once again, no discernible effect for systematic distortion when the data are assumed to be ordinal, with all three 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, although the entire configuration is rotated from the true orientation towards an orientation which is more nearly like the principal components of 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, we refrain from definitive comments at this stage of investigation, particularly considering that the same z_{ijka} is added across different error levels.

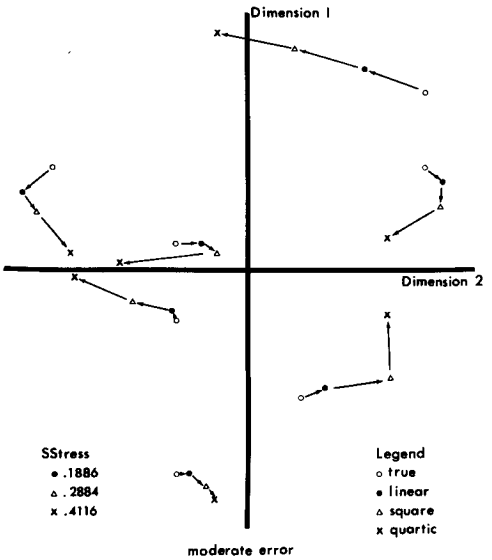
The weights, on the other hand, simply show a nonsystematic deterioration as the amount of error increases. Although the relatively heavier weighting on dimension I is preserved, the order of individual subjects along the dimensions of the weight space is destroyed, let alone the ratio of an individual's 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 discuss 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, showing 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. It does, as can be seen from 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 remaining intact (though modified). When we recall that there was no effect of systematic monotonic distortion when the





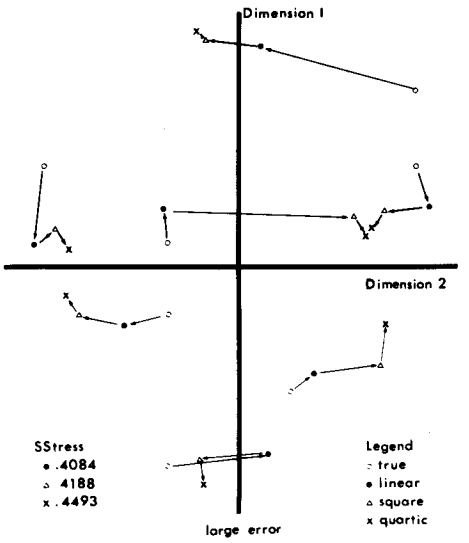
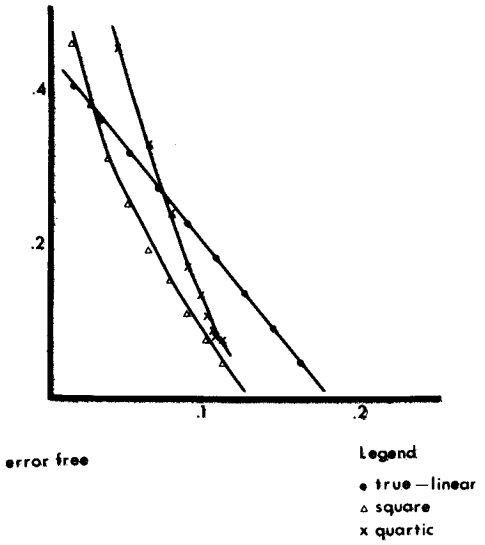
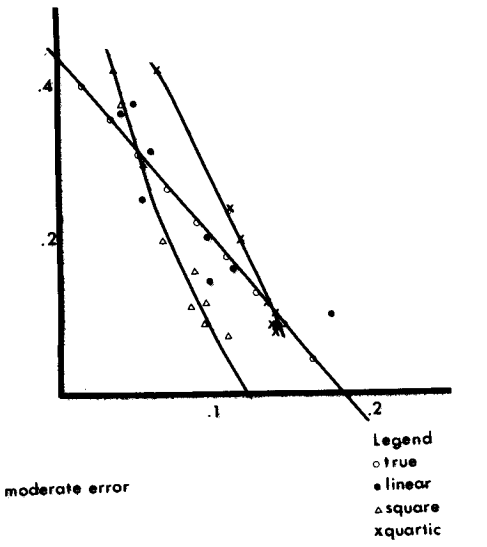


Figure 3. Monte Carlo study: Effects on the stimulus configuration when the data are assumed to be ratio.





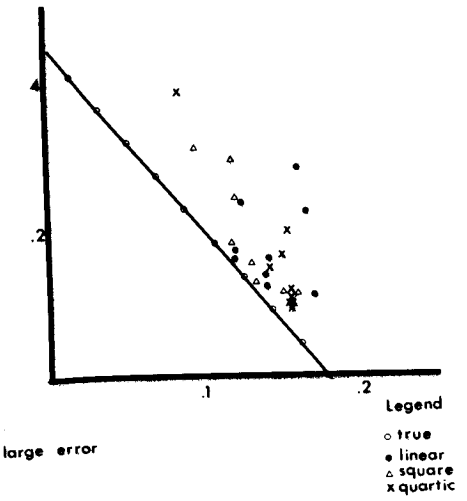


Figure 4: Monte Carlo study: Effects on the weight space when the data are assumed to be ratio.

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; and 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 together, 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 dimension I. Specifically, the variance of a points' 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 than I. Correspondingly, we see in the weight space that the variance of weights on dimension II decrease faster than for dimension I as error increases, suggesting 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 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.

4.2 Real data & 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 & 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 & 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 structure was essentially identical to that obtained by Jones & Young (who used the ratio assumption). However, the ALSCAL weight structure was more homogeneous than the one found by Jones & 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 & Young interpretation, even though it was not identical. Note that the weight homogeneity is at least partly a function of measurement level, but that we anticipate more homogeneous weights than with the INDSCAL method, as will be discussed in section 5.1. Finally, we note that these analyses assumed the data were matrix-conditional, which is, implicitly, the assumption made by Jones & 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 (1975; Young, 1975).

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 (although there are 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 regards 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 ages, the ages being 6-year-olds, 8-year-olds, 10-year-olds and adults. In our analysis we included only the youngest and oldest children 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 adults and children weights 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, with the groups 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 the light of the previous interpretation this indicates that six-year 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 import.

4.3 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,

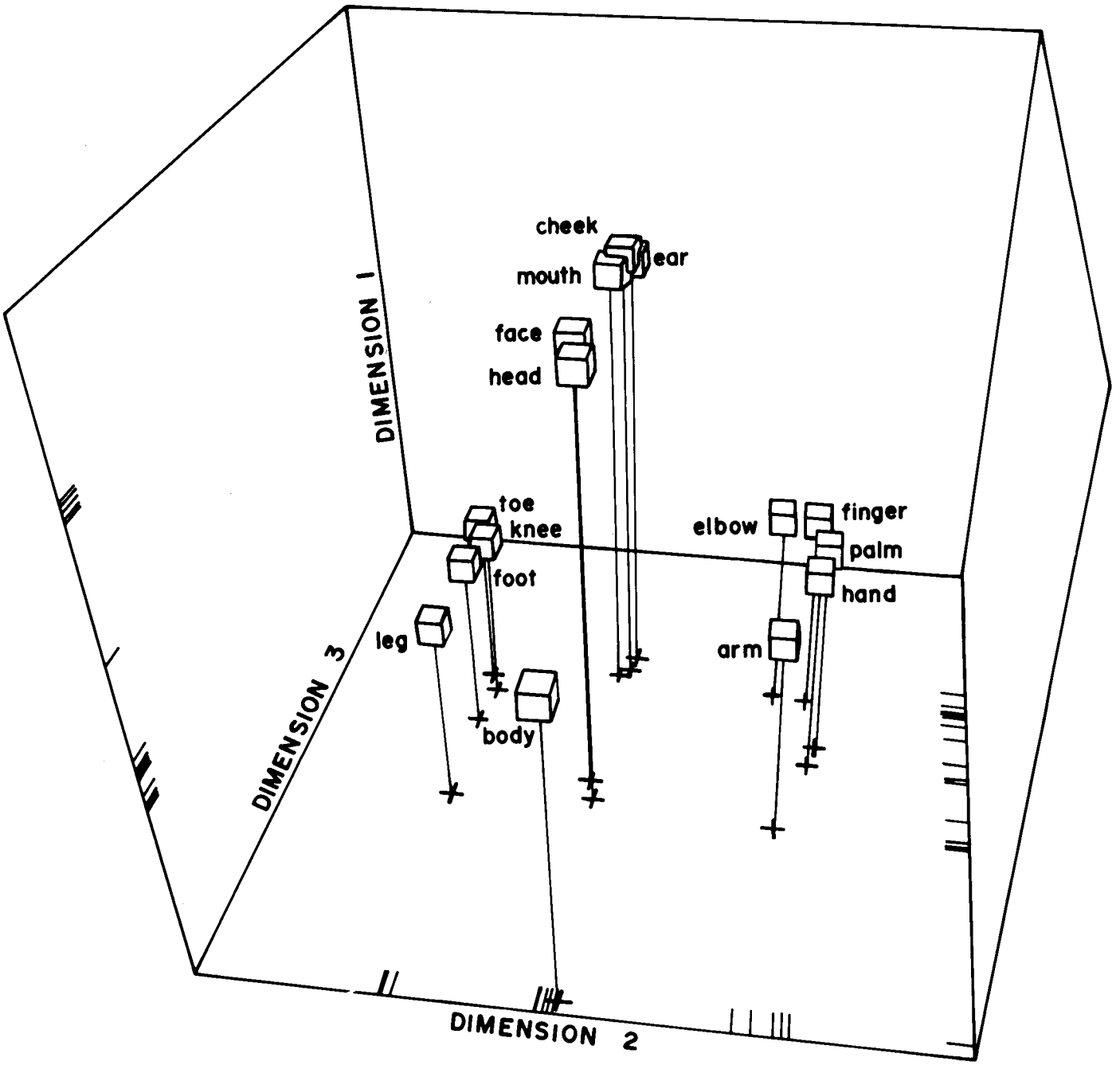


Figure 5: Jacobowitz Body-parts data: Three-dimensional stimulus space.

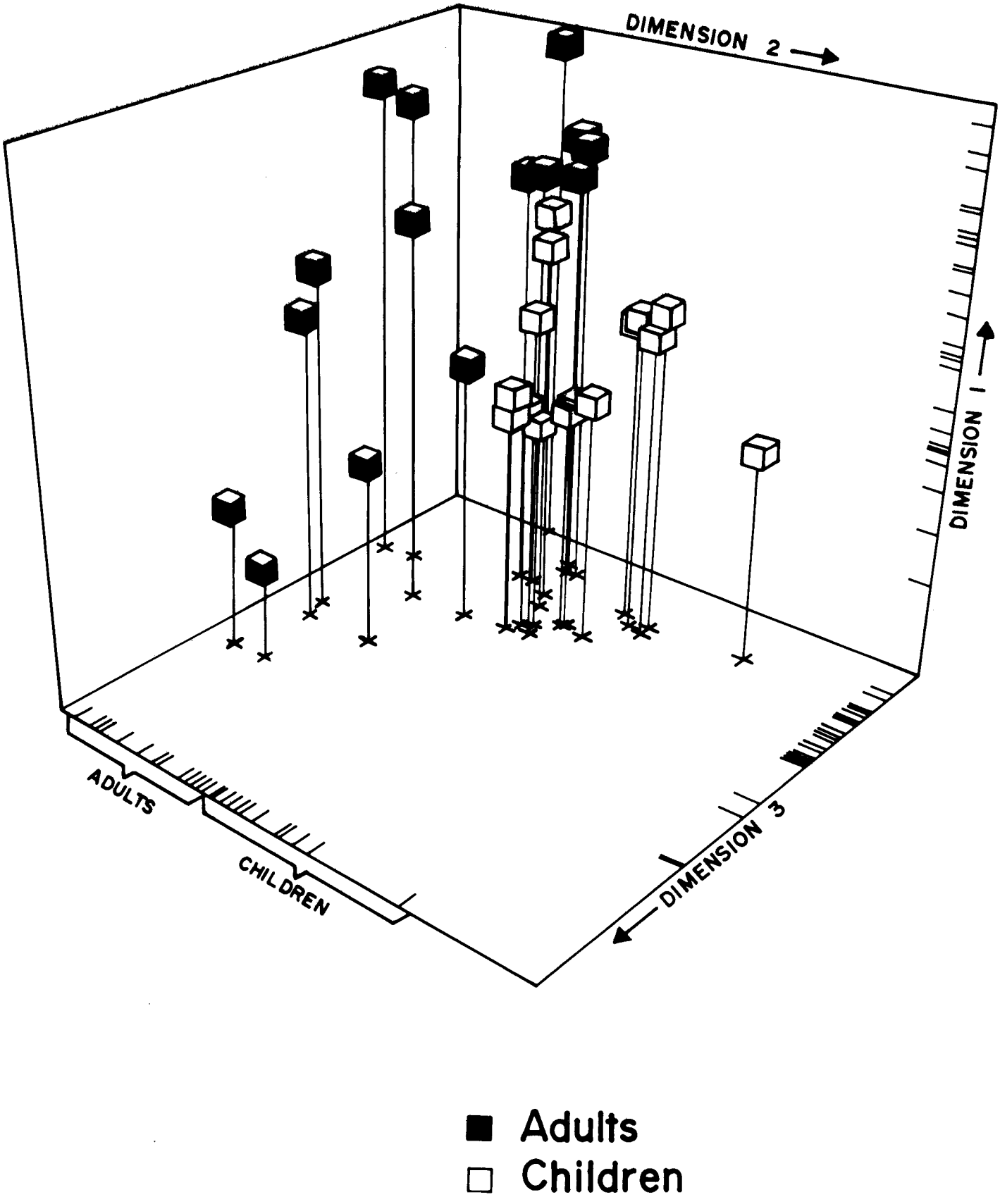


Figure 6: Jacobowitz Body-parts data: Three-dimensional subject weight space.

p. 511), Ekman's color data (1954, p. 468), Miller & Nicely's sound data (1953), Peterson & Barney's vowel data (1952), Green & Rao's breakfast menus data (1972), Hayashi's rice data (1974), among others. In all cases the obtained stimulus configuration was virtually indistinguishable from the published results (even though the published results 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 weaker measurement assumptions 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), including the fact 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) values used by Hayashi. Thus, we see that ALSCAL converges to the same solution independently of the initial category values, though, of course, the number of iterations required to reach convergence is larger for the random values.

Finally, we reanalyzed Ekman's color data using the nominal measurement assumption, and collapsing the number of observation categories to nine categories, by combining Ekman's categories.

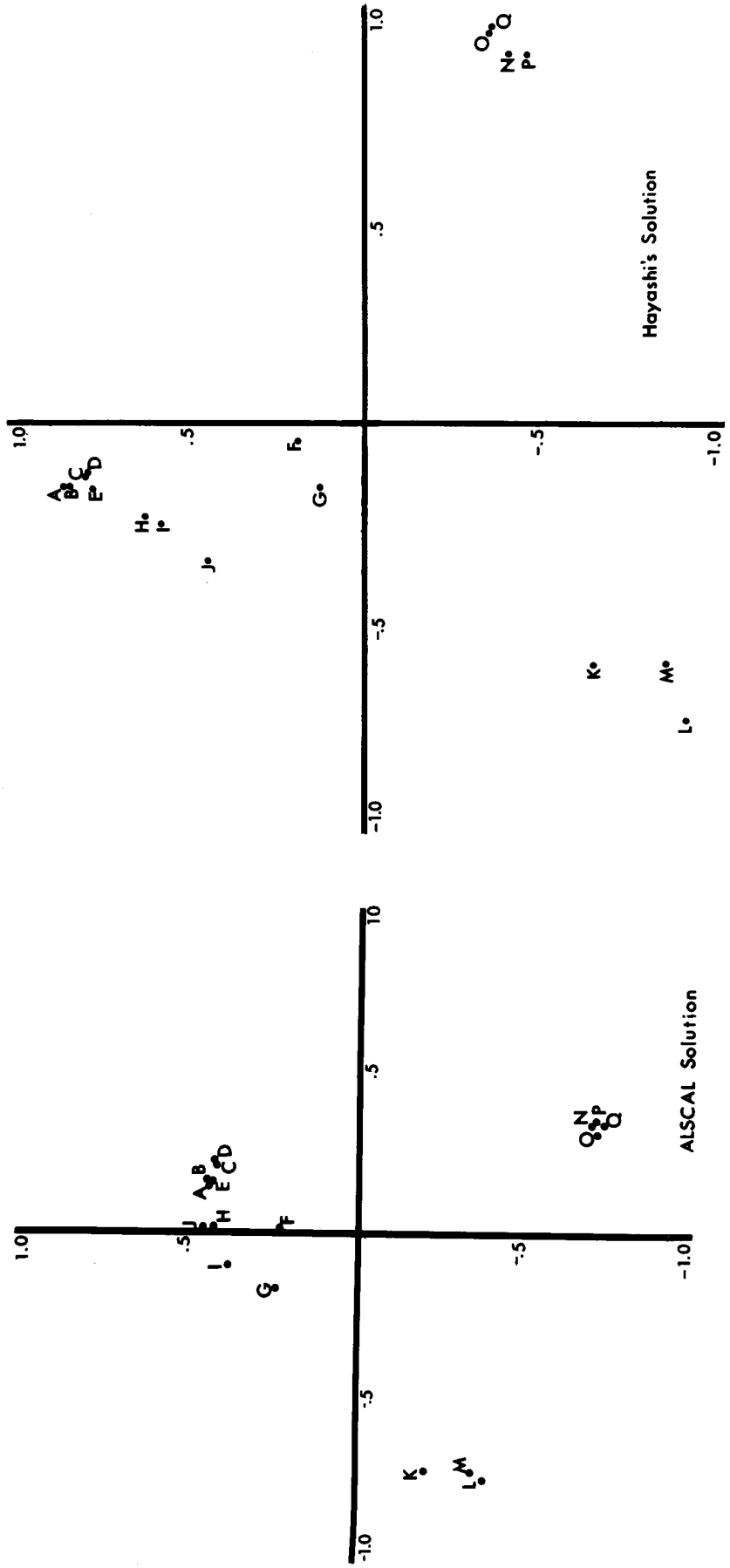


Figure 7: Hayashi's rice-strain data: Stimulus space (unweighted model) from ALSCAL and from Hayashi's paper.

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 assigned values to the observation categories is the reverse of the order of the desired distances. In the case of 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.

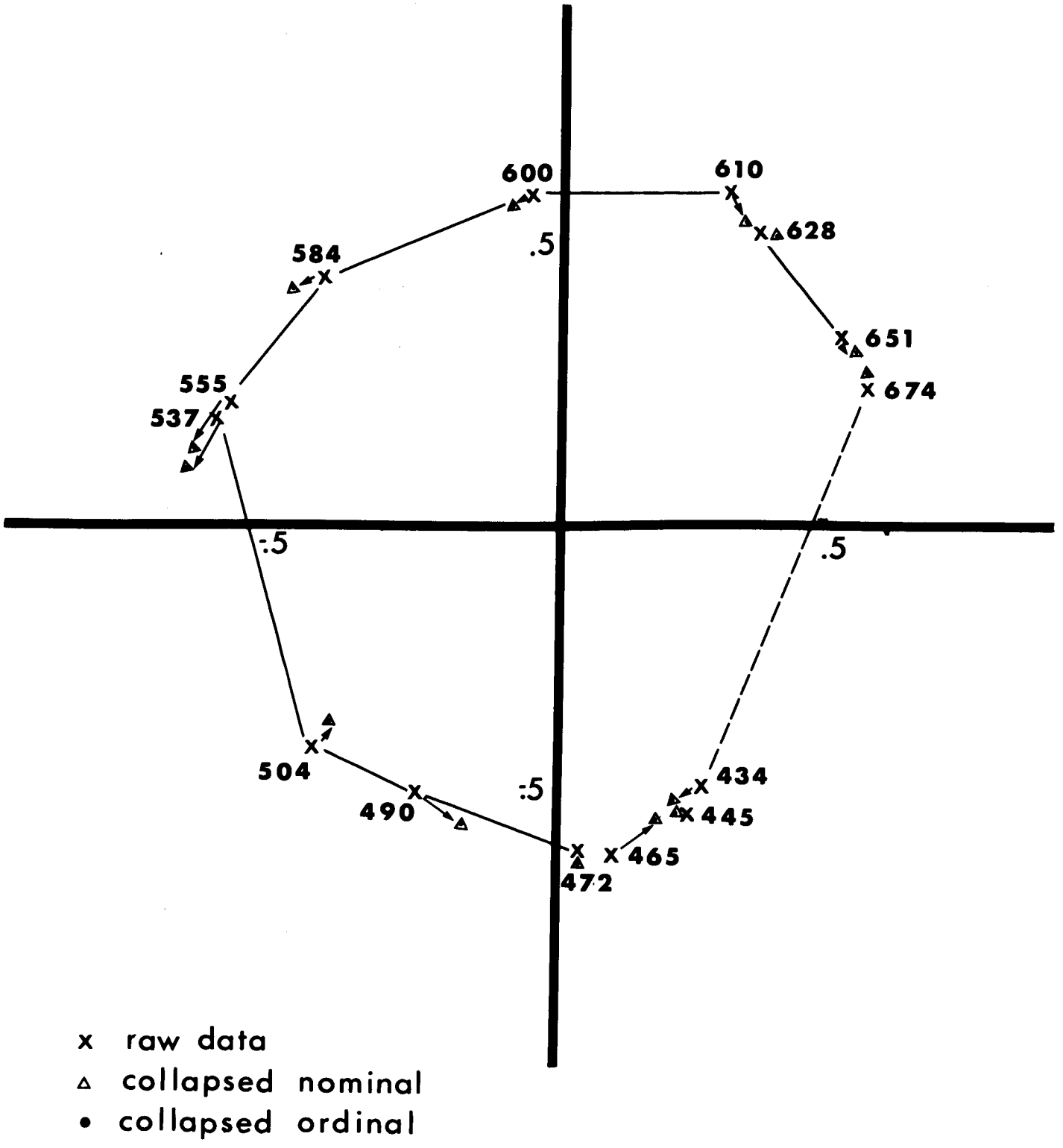


Figure 8: Ekman's color data: Stimulus space (unweighted model) with wavelength of each stimulus indicated in nanometers.

5.0 Discussion

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

5.1 Interpretation of X and W

We do not dwell at length on the interpretation of $\underline{X}$ and $\underline{W}$ since Carroll & Chang (1970) have already done so. The interpretation of $\underline{X}$ is in every way identical to the earlier work ($\underline{X}$ represents the stimuli as points in an unrotatable Euclidian space with dimensions of unit length), but there are three subtle differences in the interpretation of $\underline{W}$, although its general nature is unchanged ($\underline{W}$ represents the subjects as vectors whose direction indicates the relative importance of each dimension to each subject).

The first difference in the interpretation of $\underline{W}$ is that with unconditional data it is permissible to make direct inter-subject weight comparisons, whereas for conditional data (of either type) and for Carroll & Chang's proposal (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, and subject B has weights of .40 and .60, then for any type of data we may say that subject places 1.33=.80/.60 more weight on dimension one than he places on dimension two, and that subject B 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 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) indicate 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, and that is 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, but the difference is that with the Carroll & Chang approach this "variance accounted for" is being optimized, whereas in our procedure it is not. As was noted by Carroll & Chang, the "variance accounted for" notion is only precisely true when the configuration is exactly orthonormal ($\underline{X}'\underline{X}=\underline{I}$). 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 entirely zero weights. For the Carroll & Chang situation zero weights for a subject means that the model of his judgments consists of a scalar product matrix which is entirely zero, whereas for our situation the subject would have a distance matrix which was entirely zero. Now for the Carroll & Chang approach the model's zero scalar products matrix is being 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, for our approach the zero distance matrix is being fit to the optimally scaled observations which do not have a zero mean, thus the means of the two matrices are not the same. Therefore, for our approach a vector of zero weights is going to contribute relatively more to the apparent lack of fit than for the Carroll & Chang approach. In a practical sense this means that zero (or nearly zero) weights are less likely to occur with our approach, and that the weight structure obtained with our approach may be similar to the weight Carroll & 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 & Chang procedure. This may account, in part, for some of the results observed in the previous section, both in the Monte Carlo study (where the weights became more homogeneous as the error increased) and in the analysis of the Jones & Young (1970) data (where the weights were more homogeneous than in the Carroll & Chang analysis, although the effect interacted with the assumed measurement level and was least prevalent with the

measurement level used by Carroll & Chang).

5.2 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. These models correspond to whether we allow for individual differences only in the response process (i.e., response bias), only in the judgmental process (including perceptual and cognitive processes), or in both the response and judgmental processes. It should be pretty

clear by now that individual differences in judgmental processes is reflected by the weights of the weighted Euclidian model, thus we must choose this model if we are interested in allowing for this type of individual differences (the Horan (1969) and Carroll & 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, and 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 <u>set</u> 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, 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, one that allows for replicated data, but assumes that the replications arise from subjects with identical judgmental and response processes.

5.3 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, 1972), PARAFAC, a model which permits individuals to have weighted oblique dimensions (Harshman, 1970), 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 (Eq. 8) with the definition of the weights matrix changed in different ways for the different models.

As has been discussed by Carroll & Chang (1972), 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

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

or, in matrix notation,

$$(50) \qquad d^2_{ijk} = (\underline{x}_i-\underline{x}_j)\underline{R}_k(\underline{x}_i-\underline{x}_j)'$$

where $\underline{x}_i$ is a row vector of coordinates for point i, and where $\underline{R}_k$ 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 $\underline{R}_k$ to be a diagonal matrix.

The other models mentioned at the beginning of this section are obtained as follows: For Carroll & Chang's model we decompose $\underline{R}_k$ so that

(51) $\underline{R}_k = \underline{U}_k \underline{W}_k \underline{U}'_k$

where $\underline{U}_k$ is orthogonal and $\underline{W}_k$ is diagonal, and where $\underline{U}_k$ can be interpreted as a subject's orthogonal rotation of the original coordinates $\underline{X}$ to a new orientation, and where his weights $\underline{W}_k$ 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 $\underline{X}$ is not unique.)

For Harshman's model we decompose $\underline{R}_k$ so that

(52) $\underline{R}_k = \underline{W}_k \underline{C} \underline{W}'_k$

where $\underline{W}_k$ is diagonal and $\underline{C}$ is square symmetric with unit diagonals, $\underline{C}$ is interpreted as a matrix of cosines of angles between oblique dimensions, and $\underline{W}_k$ 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 $\underline{X}$ (whose orientation is uniquely determined). Note that all subjects weight the same oblique dimensions.

For Tucker's model we decompose $\underline{R}_k$ so that

(53) $\underline{R}_k = \underline{W}_k \underline{C}_k \underline{W}'_k$

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 $\underline{C}_k$, 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 $\underline{X}$ is not unique here.) This decomposition of $\underline{R}_k$ 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 3.4.1). The modification is to redefine the matrix $\underline{Y}$ so that all pairs of dimensions are present as well as all pairs of points. Thus, if we define $\underline{Y}$ to be an order $n(n-1)/2$ by $t(t+1)/2$ matrix with general element $(x_{ia}-x_{jb})^2$, and then apply Eq. (36) we would obtain least squares estimates of the $\underline{R}_k$, which can then be decomposed in the desired way.

5.4 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:

$$(54)\quad d_{ijk}^{\ell} = \sum_{a=1}^{t} w_{ia} |x_{ja}-x_{ka}|^{\ell}, \quad w_{ia} \geq 0, \quad 1 \leq \ell \leq \infty,$$

so that ℓSTRESS would become

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

then with some rather minor modifications in section 3.4.3 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 space (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.

5.5 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 ($\underline{X}$, $\underline{W}$, $\underline{D}^*$, 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, not discrete notion. While it is obviously the case that only certain discrete points on the measurement level continuum are axiomatizable, 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 to investigate the nature of the optimal scaling $\underline{D}^*$. 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 ($\underline{D}^*$) deviate from linearity. Formally, we

might obtain the Pearson correlation between $\underline{D}^*$ and the set of data $\underline{O}$, as a descriptive indication of deviation from linearity (note that this is obtained for the ordinal level analysis for which the Spearman rank order correlation between $\underline{D}^*$ and $\underline{O}$ 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, and 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 preceeding 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 $\underline{D}^*$ and the distances $\underline{D}$. 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 a longer

way along the measurement continuum. However, this is not because we have reached a higher degree of quantification, but because 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 $\underline{O}$ and $\underline{D}^*$ for the nominal analysis would be unity, telling us that the data are actually ordinal when analyzed by the chosen model. The Pearson correlation between $\underline{D}^*$ and $\underline{D}$ would be the same for the two analyses (as is the SSTRESS) telling us that no more quantification was possible under the nominal assumption than under the ordinal assumption, implying that the data are ordinal. Finally, the Pearson correlation between $\underline{D}^*$ and $\underline{D}$ 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.

5.6 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 metric 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 both 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 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 & Young (1972) data as a function of dimensionality and measurement level. These times are CPU time only, with no I/O time included.

Table 2

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

We have set the convergence criterion to a value of δ=.001 where δ is the improvement in SSTRESS from one iteration to the next. (Note we use ϕ, that is, the square root of Eq. 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 & 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 (the latter were both optimizing Kruskal's second STRESS formula whereas ALSCAL was not, which accounts for the larger stress value obtained from ALSCAL), and the CPU times required to analyze the Jones & Young (1972) data in three dimensions by ALSCAL and INDSCAL. 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.

Table 3

Program	CPU time	Iterations	STRESS 2	STRESS 1	Improvement	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.

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 procedure in that they do not go precisely to the minimum in each sub-problem, but over-step the minimum. The over-stepping is referred to as overrelaxation. For these procedures it has been found that they are the 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 have not yet tried this, however.

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). Furthermore, no matter how frequently we solve one of the sub-problems 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 effected, 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 sub-problem is solved the same number of times in an iteration. Thus, it is usually more efficient to solve each sub-problem 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.

6.0 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, 1975; Young, de Leeuw & Takane, 1975) has shown that these two keystones yield viable results with linear models, and 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. The very general approach to algorithm construction has been tried by one of the current authors (Young, 1973) with mixed success, and it is our conclusion that it is more efficacious to develop several "highly tuned" algorithms as we have done here.

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 sub-problems, is only as simple as the simplest sub-problem. In our previous work with linear models each of the subproblems was very simple. However, with the current work one of the sub-problems, 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 sub-problem, 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. With this confidence we now turn to the bilinear model and focus our research on the nonmetric principal components situation.

References

Bloxom, B. Individual differences in multidimensional scaling. Research Bulletin 68-45. Princeton, N.J.: Educational Testing Service, 1968.

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

Bôcher, M. Introduction to Higher Algebra. New York: MacMillan, 1907.

Carroll, J.D. and Chang, J.J. IDIOSCAL (Individual Differences in Orientation Scaling). Paper presented to the Spring, 1972 meeting of the Psychometric Society, Princeton, N.J.

Carroll, J.D. and Chang, J.J. Some methodological advances in INDSCAL. Paper presented at the Spring meeting of the Psychometric Society, Stanford University, August 28-29, 1974.

Coombs, C.H. A Theory of Data. New York: Wiley, 1964.

de Leeuw, J. The positive orthant method for nonmetric multidimensional scaling. Research note RN 001-70. Datatheorie, University of Leiden, The Netherlands, 1970.

de Leeuw, J. Canonical Analysis of Categorical Data. University of Leiden, The Netherlands, 1973.

de Leeuw, J. An initial estimate for INDSCAL. Unpublished note, 1974.

de Leeuw, J. On the balanced least squares transformation. Psychometrika, 1975a (in press).

de Leeuw, J. Canonical discriminant analysis of relational data. Research Bulletin RB004-75. Datatheorie, University of Leiden, The Netherlands, 1975b.

de Leeuw, J. and Pruzansky, S. A new computational method to fit the weighted Euclidian model (SUMSCAL). Mimeographed notes, Bell Telephone Laboratories, 1975.

de Leeuw, J., Young, F.W. and Takane, Y. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. Psychometrika (in press), 1976.

Eckart, C. and 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 Edition. Edinburgh: Oliver and Boyd, 1946.

Funk, S., Horowitz, A., Lipshitz, R. and Young, F.W. The perceived structure of American ethnic groups: The use of multidimensional scaling in stereotype research. Sociometry, (in press) 1976.

Gill, P.E. and Murray, W. Two methods for the solution of linearly constrained and unconstrained optimization problems. NPL Report NAC 25, November, 1972.

Green, P.E. and 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.

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.

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

Harshman, R.A. Foundations of the PARAFAC procedure: Models and conditions for an explanatory multi-modal factor analysis. Working papers in phonetics (No. 16), University of California at Los Angeles, 1970.

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.

Horst, P. The Prediction of Personal Adjustment. Bulletin 48 of the Social Science Research Council. New York, 1941.

Jacobowitz, D. The acquisition of semantic structures. Unpublished doctoral dissertation, University of North Carolina, 1975.

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. and Wadington, J. Sensitivity of INDSCAL to simulated individual differences in dimension usage patterns and judgmental error. Paper delivered to the Spring meeting of the Psychometric Society, Chicago, Illinois, 1973.

Jones, L.E. and 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.

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

Kruskal, J.B. and Carroll, J.D. Geometric models and badness-of-fit functions. In Multivariate Analysis, Vol. 2. New York: Academic Press, Inc., 1969.

Kruskal, J.B., Young, F.W., and Seery, J.B. How to use KYST, a very flexible program to do multidimensional scaling and unfolding. Unpublished manuscript, Bell Telephone Laboratories, 1973.

Lawson, C.L. and 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. and 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. and Roskam, E.E. A mathematical and empirical analysis of two multidimensional scaling algorithms. Psychometrika, 1973, 38 (monograph supplement).

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. and Abelson, R.P. The additive constant problem in multidimensional scaling. Psychometrika, 1956, 21, 1-15.

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

Obenchain, R. Squared distance scaling as an alternative to principal components analysis. Mimeographed notes, Bell Telephone Laboratories, 1971.

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

Roskam, E.E. Data theory and algorithms for nonmetric scaling (parts 1 and 2). Unpublished manuscript, Catholic University Nijmegen, Netherlands: Department of Mathematical Psychology, 1969.

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

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.

Shepard, R.N. The analysis of proximities: Multidimensional scaling with an unknown distance function: I and II. Psychometrika, 1962, 27, 125-140, 219-246.

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 Some mathematical notes on three-mode factor analysis. Psychometrika, 1966, 31, 279-311.

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

Wilf, H.S. The numerical solution of polynomial equations. In Ralston, A. and W.S. Wilf (Eds.) Mathematical Methods of Digital Computers, Vol. 1. New York: Wiley, 1960, 233-241.

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

Yates, A. Nonmetric individual-differences multidimensional scaling with balanced least squares monotone regression. Paper presented to the Spring meeting of the Psychometric Society, Princeton, N. J., 1972.

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. Nonmetric multidimensional scaling: Recovery of metric information. Psychometrika, 1970, 35, 455-474.

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

Young, F.W. Polynomial conjoint analysis: Some second order partial derivatives. L.L. Thurstone Psychometric Laboratory Report, No. 108, July 1972b, Chapel Hill, North Carolina.

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. Scaling replicated conditional rank-order data. Sociological Methodology, 1975a, 129-170.

Young, F.W. Methods for describing ordinal data with cardinal models. Journal of Mathematical Psychology, 1975b, 12 416-436.

Young, F.W., de Leeuw, J. and Takane, Y. Multiple and canonical regression with a mix of qualitative and quantitative variables: An alternating least squares method with optimal scaling features. Psychometric Laboratory Report No. 146, Chapel Hill, North Carolina, 1975.

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

Young, F.W. and Torgerson, W.S. TORSCA, a fortran-IV program for nonmetric multidimensional scaling. Behavioral Science, 1968, 13, 343-344.