PSYCHOMETRIKA—VOL. 48, NO. 3.
SEPTEMBER, 1983

MULTIDIMENSIONAL SCALING MODELS FOR REACTION TIMES AND SAME-DIFFERENT JUDGMENTS

YOSHIO TAKANE AND JUSTINE SERGENT

MCGILL UNIVERSITY

A method for joint analysis of reaction times and same-different judgments is discussed. A set of stimuli is assumed to have some parametric representation which uniquely defines dissimilarities between the stimuli. Those dissimilarities are then related to the observed reaction times and same-different judgments through a model of psychological processes. Three representation models of dissimilarities are considered, the Minkowski power distance model, the linear model, and Tversky's feature matching model. Maximum likelihood estimation procedures are developed and implemented in the form of a FORTRAN program. An example is given to illustrate the kind of analyses that can be performed by the proposed method.

Key words: Minkowski power distance model, feature matching model, maximum likelihood estimation, AIC.

Introduction

Statistical analysis often proceeds with data as they are given, and no serious concern is taken for possible data transformations in finding data representations. Psychological scaling, on the other hand, explicitly aims at both data transformations (response scaling) and data representations (stimulus scaling). These two scaling problems were traditionally viewed as one, and only fairly recently the importance of the distinction came to be fully realized. Since Shepard's [1962] and Kruskal's [1964a, b] landmark papers on nonmetric multidimensional scaling, it has quickly become a new tradition in psychological scaling to seek both the optimal data transformation and the optimal data representation based on a single optimality criterion [Young, de Leeuw & Takane, 1980].

Too often, however, the scale level of measurement (i.e., whether the given numerals constitute a ratio scale, an interval scale, an ordinal scale, etc.) is the only major data characteristic that is respected in response scaling. Classifying the data in terms of just the scale level of measurement seems to be too crude. For example, similarity ratings and similarity rankings both yield ordinal measures of similarity which are different in other important respects (e.g., distributional properties of the data). There are many different ways to collect similarity data, and they differ not only in the form of the data they provide, but also in the process by which the data are generated. Unfortunately it is rather rare to find scaling procedures which explicitly take into account the specific data generation process presumed to underlie each data collection method. (But see Takane, 1978, 1981; Takane & Carroll, 1981). In this paper we discuss a scaling method for two-choice reaction time data which incorporates a model of psychological processes linking a stimulus representation to the particular form of the observed data.

Because mental processes are embedded in real time, measuring the interval between

The work reported in this paper is supported by Grant A6394 to the first author from the Natural Sciences and Engineering Research Council of Canada. Portions of this study have been presented at the Psychometric Society meeting in Chapel Hill, N.C., in May, 1981. We thank Tony Marley, Jim Ramsay and anonymous reviewers for their helpful comments. MAXRT, a computer program which performs the computations described in this paper may be obtained by writing to the first author.

Requests for reprints should be sent to Yoshio Takane, Department of Psychology, McGill University, 1205 Avenue Docteur Penfield, Montreal, Quebec, H3A 1B1, Canada.

0033-3123/83/0900-4109$00.75/0

stimulus presentation and response production in a given task has proven an objective tool for the systematic observation of mental events [Posner, 1978]. Thus, variations in reaction times consequent to manipulations of task requirements or stimulus parameters provide data that can be used to study the time course of information processing in the human nervous system.

In the two-choice reaction time experiments to be considered in this paper, two stimuli are presented, either simultaneously or successively. The subject is instructed to respond, as quickly as possible, whether the two stimuli are "same" or "different." The time to respond and the type of judgment are recorded as data. The two-choice reaction time data are, unlike direct judgments of similarity, performance measures. That is, as in the mental testing situation, it is clear to the subject what constitutes a better performance (i.e., shorter reaction times and fewer errors). This is one advantage of the reaction time data, since it is relatively difficult to systematically fake the responses.

Curiously, however, scaling of the choice-reaction time data has not been too successful [Young, 1970] and there are at least two reasons conceivable for this. First, reaction time data are usually very noisy. For example, in the data sets we will discuss in the result sections, approximately 50 to 75% of the variability in reaction times are due to error. Consequently, one needs to get a lot of data in order to secure reliable estimation. Second, the two-choice reaction time data are intrinsically bivariate (one for same-different judgment and the other for reaction time), only meaningful in the context of the other. For example, the reaction time of, say, 800 ms may mean something entirely different depending on whether it is associated with a "same" or "different" judgment [Podgorny & Garner, 1979]. Similarly, a high confusion rate of one stimulus with another may be produced by two entirely different processes depending on whether it occurs with relatively short reaction times or long reaction times. It is far from trivial what constitutes a better performance for the two variables taken together. However, this simple fact is often ignored in order to apply conventional scaling procedures. The stimulus confusion data (arising from same-different judgments) and the reaction time data are often analyzed separately, each disregarding the existence of the other. By contrast the method we discuss in this paper takes full cognizance of the bivariate nature of the two-choice reaction time data. It finds a single common stimulus representation by joint analysis of reaction times and same-different judgments.

The method we discuss in this paper assumes that a set of stimuli have some parametric representation. Based on this representation a dissimilarity between each pair of stimuli is derived. The dissimilarity between the stimuli is then assumed to be error-perturbed in a specific way, and the error model specifies the nature of this perturbation process. The error-perturbed dissimilarity is then related to both the observed reaction time and the same-different judgment through models of psychological processes. These models, called the response models, specify the nature of the relationship between the observed data and the underlying processes. Thus, once all the relevant component models are specified in sufficient detail, the likelihood of the observed data can be stated in term of parameters in these models. Maximum likelihood (ML) estimation (or some other estimation procedures) can then be used to determine the estimates of the parameters.

Maximum likelihood estimation procedures are developed for three representation models of dissimilarities, the Minkowski power distance model, the linear model and Tversky's [1977] feature matching models. These are the models which came naturally to our mind in a search for the best representation model for our "face" data. Theoretical and empirical motivations behind fitting these models will be discussed in the result sections. The maximum likelihood estimation offers a number of advantages regarding statis-

tical model evaluations [Ramsay, 1977, 1978; Takane, 1978, 1981; Takane & Carroll, 1981]. The kinds of statistical model evaluations that are feasible, and that are particularly interesting in the context of the present paper will be demonstrated through a concrete analysis example.

The Method

The Model

As mentioned earlier we assume that a set of n stimuli has some parametric representation, from which a dissimilarity between two stimuli is uniquely defined. For stimuli i and j we may write this dissimilarity as

$$d_{ij} = d(\theta_i, \theta_j), \tag{1}$$

where θ_i and θ_j are vectors of parameters characterizing stimuli i and j, respectively, and d is a function expressing their combination rule. The parameter vectors (θ_i and θ_j) themselves may be functions of some other parameters. We consider in this paper three distinct models of dissimilarities, the Minkowski power distance model, the linear model, and Tversky's feature matching model [Tversky, 1977]. However, for the sake of generality we defer the explicit statement of these models until the result sections and proceed with the general form of the representation model.

We further assume that d_{ij} is error-perturbed by

$$\lambda_{ijkr} = d_{ij} e_{ijkr}, \tag{2}$$

where e_{ijkr} is the error random variable operating at replication r, and that

$$\ln e_{ijkr} \sim N(0, \sigma_k^2). \tag{3}$$

The subscript k may represent a subject, an experimental session, or their combination. (In this paper the data from different subjects are analyzed separately, so that k will always refer to an experimental session.) Note that the subscript k in σ_k^2 allows for possible differences in dispersion over experimental sessions, although it may alternatively be assumed constant for all k (i.e., $\sigma_k^2 = \sigma^2$). The σ_k^2 in (3), however, implies that the variance of λ_{ijkr} is equal for equal d_{ij}. (The variance of λ_{ijkr} generally increases as d_{ij} increases.) Alternatively, we may allow σ_k^2 to vary across stimuli i and j as well as k (i.e., $V[\ln e_{ijkr}] = \sigma_{ijk}^2$), and then assume more elaborate variance component models for σ_{ijk}^2 (e.g., $\sigma_{ijk}^2 = \sigma_k^2(\alpha_i^2 + \alpha_j^2)$, etc.) [Ramsay, 1982]. The estimation procedure in this case, however, would be much more complicated, and this possibility will not be pursued in this paper. The log normal distributional assumption made in (3) is essentially the same as that made by Ramsay [1977] in his maximum likelihood multidimensional scaling (MDS) procedure. (A further justification of the log normal assumption will be given later.)

Let T_{ijkr} denote the random variable for reaction time, and Y_{ijkr} the random variable for same-different judgment. (We use the corresponding lower case letter, t_{ijkr} and y_{ijkr}, to denote the observed values.) We define Y_{ijkr}as

$$Y_{ijkr} = \begin{cases} 1, & \text{if the judgment is "same" for the pair of stimuli } i \text{ and } j \text{ at replication } r \text{ in session } k. \\ 0, & \text{if the judgment is "different".} \end{cases} \tag{4}$$

The error-perturbed dissimilarity λ_{ijkr} defined in (2) is related to the observable quantities, T_{ijkr} and Y_{ijkr}, by specific response models. Let b_k represent a threshold value (assumed

constant throughout session k) such that

$$Y_{ijkr} = \begin{cases} 1, & \text{when } \lambda_{ijkr} < b_k \\ 0, & \text{when } \lambda_{ijkr} \geq b_k . \end{cases} \tag{5}$$

That is, when the error perturbed dissimilarity exceeds the threshold value, a "different" judgment is generated; otherwise, a "same" judgment is generated. We use this as the response model for the same-different judgment. Note that model (5) is essentially Thurstonian [Torgerson, 1958]. It also plays a central role in the signal detection theory [Green & Swets, 1966; Sorkin, 1962; see also Zinnes & Wolff, 1977; Takane, 1981]. Note that it is not really necessary to assume that the threshold is constant. If a variable threshold, b_k^*, follows the log normal distribution with $E[\ln b_k^*] = \ln b_k$, then by redefining $\sigma_k^2 = V[\ln b_k^* - \ln \lambda_{ijkr}]$, everything else remains intact. Since the basic line of the argument stays the same, we may proceed as if the threshold were constant.

From the distributional assumptions on λ_{ijkr} [(2) and (3)] we obtain

$$Q_{ijk} = \Pr(Y_{ijkr} = 1) = \int_{-\infty}^{v_{ijk}} \phi(z)\, dz = \Phi(v_{ijk}), \tag{6}$$

where

$$v_{ijk} = \frac{(\ln b_k - \ln d_{ij})}{\sigma_k}, \tag{7}$$

and $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the density function and the (cumulative) distribution function of the standard normal distribution. It follows that $\Pr(Y_{ijkr} = 0) = 1 - Q_{ijk}$.

The response model for reaction time should be specified separately for a "same" judgment and a "different" judgment. This is because there is some empirical evidence suggesting that the relation between reaction time and dissimilarity is opposite for the two types of judgments. That the (correct) reaction time is inversely related to stimulus dissimilarity for *different* pairs has long been a common sense knowledge among experimental psychologists. It is the basis for applying nonmetric MDS to reaction time data [Young, 1970; Shepard, Kilpatric & Cunningham, 1975] assuming an inverse monotonic function between the reaction time data and underlying distances (i.e., reaction time is considered a measure of similarity between two stimuli). Recently, however, just the opposite relation for "same" judgments was experimentally demonstrated. Podgorny and Garner [1979; Podgorny, 1978] in their ingenious experiments have shown that it takes less time to judge more similar stimuli "same". Thus, we may suppose that for the "same" judgments, reaction time works as a measure of dissimilarity.

For a "different" judgment, we assume

$$\ln T_{ijkr} \sim N[p_k (\ln d_{ij} - \ln b_k) + a_k,\ q_k^2]. \tag{8}$$

That is, the distribution of T_{ijkr} is log normal with $E[\ln T_{ijkr}] = p_k (\ln d_{ij} - \ln b_k) + a_k$ and $V[\ln T_{ijkr}] = q_k^2$. Here p_k (assumed to be negative) and a_k are, respectively, slope and intercept parameters relating the difference between the log dissimilarity ($\ln d_{ij}$) and the log threshold ($\ln b_k$) to the expected value of the log reaction time. The log normality of T_{ijkr} is assumed to follow from the log normality of λ_{ijkr}. Since $E[\ln \lambda_{ijkr}] = \ln d_{ij}$ and $V[\ln \lambda_{ijkr}] = \sigma_k^2$ from (2) and (3), it may well be assumed further that

$$q_k^2 = p_k^2 \sigma_k^2 \qquad (\text{or } q_k = -p_k \sigma_k). \tag{9}$$

Let T_{ijkr}^* denote the median of T_{ijkr}. With the log normal distribution we have

$$T_{ijkr}^* = \exp\{E[\ln T_{ijkr}]\},$$

$$(\text{or} \quad \ln T_{ijkr}^* = E[\ln T_{ijkr}]),$$

and hence

$$T^*_{ijkr} = a^*_k \exp\{p_k (\ln d_{ij} - \ln b_k)\}$$
$$= a^*_k \left(\frac{d_{ij}}{b_k}\right)^{p_k}, \tag{10}$$

where $a^*_k = \exp(a_k)$. For fixed b_k the above equation states that the median reaction time is a negative power function ($p_k < 0$) of d_{ij}, which includes, as its special case, the reciprocal function ($p_k = -1$) advocated by several authors [Curtis, Paulos, & Rule, 1973; Shepard, et al., 1975; Shepard, 1978] for correct "different" reaction times. Shepard's observations are particularly compelling, since they are based on the empirical relationship between "different" reaction times and underlying distances. He applied nonmetric MDS to several sets of reaction time data assuming that they were merely inversely related to underlying distances between stimuli, and found that the observed reaction times were nearly reciprocal to the derived distances.

Unfortunately no such evidence exists for "same" reaction times. In this case we may merely extrapolate the same relationship as above to the "same" reaction times. The relationship between reaction time and dissimilatiry, however, should be reversed in accordance with Podgorny and Garner's [1979] empirical findings. More specifically we may write the distributional assumption for a "same" judgment as

$$\ln T_{ijkr} \sim N[p_k (\ln b_k - \ln d_{ij}) + a_k, q^2_k], \tag{11}$$

which implies

$$T^*_{ijkr} = a^*_k \left(\frac{b_k}{d_{ij}}\right)^{p_k}. \tag{12}$$

Notice that positions of d_{ij} and b_k are reversed in (11) and (12) in reference to (8) and (10). For fixed b_k the median reaction time is now a positive power transformation of the stimulus dissimilarity.

The threshold parameter b_k plays an important role in (10) and (12). This parameter was originally introduced in (5) in order to capture the mechanism generating both "same" and "different" judgments. Equations (10) and (12), on the other hand, state that reaction times are functions of magnitude of stimulus dissimilarity relative to the threshold value. However, since b_k is assumed constant throughout an experimental session, T^*_{ijkr} is a monotonically decreasing function of d_{ij}, for "different" judgments, while it is a monotonically increasing function for "same" judgments. Consequently these two opposing functions have to cross each other somewhere, and b_k is precisely the point where they intersect. When $d_{ij} = b_k$, we obtain $T^*_{ijkr} = a^*_k$ in both (10) and (12). (See Figure 4, where estimated threshold values are indicated by symbol b). This makes intuitive sense, since by definition the threshold represents a point where the two types of judgments are indifferent in every respect. In both (8) and (11), b_k, p_k, a_k and q^2_k may be assumed equal across different experimental sessions.

The log normal distributional assumption [(8) and (11)] is particularly appealing in the present context, since it has distinct positive skewness, which is typical of most reaction time data. There are, of course, a number of other positively skewed distributions which might be used. These include the noncentral chi-square distribution [Hefner, 1958; Zinnes & Wolff, 1977], the gamma distribution [Restle, 1961; Marley, 1981], the general gamma distribution [McGill & Gibbon, 1965], the inverse Gaussian (or Wald) distribution [Ramsay, Note 1; see Johnson & Kotz, 1970, for general reference], the (composite) Weibull distribution [Ida, 1980], the log logistic distribution, as well as others based on the sequential sampling or random walk models [Stone, 1960; LaBerge, 1962; Link, 1975; Grice, Nullmeyer & Spiker, 1977; Krueger, 1978]. However, Chocholle's

[1940] observation (reported in Laming, 1973, p. 156) that for simple reaction time its standard deviation is roughly proportional to its mean makes the log normal distribution particularly favorable. It is a characteristic property of this distribution (as well as that of the gamma distribution.) Whether it carries over to choice reaction time situations still remains an empirical question, and the empirical validity of the distributional assumption will be critically examined in the discussion section in the light of specific analysis results. Note that the log normal distribution is not properly defined for zero dissimilarity. This will cause some problem in fitting a dissimilarity model to *same* pairs for which $d_{ii} = 0$. For example, we obtain $d_{ii} = 0$ in the Minkowski distance model, so that this model cannot be fitted to the *same* pairs under the log normal assumption. Some plausible modification of the Minkowski power distance model in order to accomodate the *same* pairs will be suggested in the discussion section.

It might be pointed out here that $\ln T_{ijkr}$ in (8) and (11) may be replaced by $\ln (T_{ijkr} - c)$, where c is some prescribed constant representing the minimum required time for just responding. This may allow us to analyze pure comparison and decision times independent of pure response times. It also has the consequence of avoiding T^*_{ijkr} approaching zero, as d_{ij} approaches either infinity or zero in (8) and (11). The value of c generally ranges from 100 to 300 ms. However, it is likely that its optimal value is situation-dependent, and its exact value seems difficult to specify *a priori*. Thus, for major portions of our analyses the value of c is assumed to be zero (i.e., (8) and (11) are fitted in their original form). Some rationale for this will be given later. However, MAXRT, the program which performs the computations described in this paper has a provision that allows us to analyze a same set of data repeatedly under different values of c. Thus, the optimal value of c may be chosen *a posteriori* on the basis of analysis results.

Let $g^{(s)}(t_{ijkr})$ and $g^{(d)}(t_{ijkr})$ represent the conditional densities of T_{ijkr} for a "same" judgment and a "different" judgment, respectively. The joint density $g(t_{ijkr}, y_{ijkr})$ of T_{ijkr} and Y_{ijkr} is then written as

$$g(t_{ijkr}, y_{ijkr}) = \{g^{(s)}(t_{ijkr})Q_{ijk}\}^{y_{ijkr}}\{g^{(d)}(t_{ijkr})(1 - Q_{ijk})\}^{1-y_{ijkr}}, \tag{13}$$

where Q_{ijk} is defined in (6). Note that

$$g(t_{ijkr}, y_{ijkr}) = \begin{cases} g^{(s)}(t_{ijkr})Q_{ijk}, & \text{when } y_{ijkr} = 1, \\ g^{(d)}(t_{ijkr})(1 - Q_{ijk}), & \text{when } y_{ijkr} = 0. \end{cases}$$

The $g^{(s)}(t_{ijkr})$ and $g^{(d)}(t_{ijkr})$ are as specified in (8) and (11).

The joint likelihood for the total set of observations can be stated as

$$L = \prod_k \prod_{i,j} \prod_r g(t_{ijkr}, y_{ijkr}), \tag{14}$$

where the products are taken over k (session), i and j (stimulus pair) and r (replication) for which observations are actually made. This requires statistical independence among pairs of observations (T_{ijkr}, Y_{ijkr}) across k, i and j, and r. While this may not be exactly true due to sequential effects [Laming, 1973], the dependency can be minimized by stretching inter-trial intervals. By taking the log of (14) we obtain

$$\ln L = \psi_1 + \psi_2, \tag{15}$$

where

$$\psi_1 = \sum_k \sum_{i,j} \sum_r \{y_{ijkr} \ln g^{(s)}(t_{ijkr}) + (1 - y_{ijkr}) \ln g^{(d)}(t_{ijkr})\} \tag{16}$$

and

$$\psi_2 = \sum_k \sum_{i,j} \{n^{(s)}_{ijk} \ln Q_{ijk} + n^{(d)}_{ijk} \ln (1 - Q_{ijk})\} + \text{constant} \tag{17}$$

with $n_{ijk}^{(s)} = \sum_r y_{ijkr}$ and $n_{ijk}^{(d)} = \sum_r (1 - y_{ijkr})$. From (8) and (11) we see that

$$\ln g^{(s)}(t_{ijkr}) = \text{constant} - \frac{1}{2}\left[\ln q_k^2 + \frac{\{\ln t_{ijkr} - p_k(\ln b_k - \ln d_{ij}) - a_k\}^2}{q_k^2}\right]$$

and

$$\ln g^{(d)}(t_{ijkr}) = \text{constant} - \frac{1}{2}\left[\ln q_k^2 + \frac{\{\ln t_{ijkr} - p_k(\ln d_{ij} - \ln b_k) - a_k\}^2}{q_k^2}\right]. \tag{19}$$

The approach leading to (14) has some resemblance to Ramsay's [1980] joint analysis of similarity and preference data, in which he dealt with the problem of relating a common stimulus configuration to two kinds of data. The major difference is that in his case the two kinds of data were similarity and preference ratings made on a same set of stimuli rather than reaction times and same-different judgments.

Estimation of Parameters

The log likelihood stated in (15) is maximized with respect to parameters in the representation model $(\boldsymbol{\theta}_i\,;\, i = 1, \ldots, n)$, $\tilde{b}_k$ ($\equiv \ln b_k$), p_k, a_k and q_k^2. (The σ_k^2 is not directly estimated. It follows from (9), once p_k and q_k^2 are obtained.) In order to facilitate the derivation we may further breakdown the log likelihood function.

Let

$$h_{ijk}^{(s)} = p_k(\tilde{b}_k - \ln d_{ij}) + a_k, \tag{20}$$

and

$$h_{ijk}^{(d)} = p_k(\ln d_{ij} - \tilde{b}_k) + a_k. \tag{21}$$

Let C_{ijk} and D_{ijk} represent the sets of indices of replications for which the judgments are "same" and "different", respectively, for stimuli i and j and in session k. Then the first part of the log likelihood in (15) can be written as

$$\psi_1 = -\frac{1}{2}\sum_k N_k \ln q_k^2 - \tilde{\psi}_1, \tag{22}$$

where N_k is the total number of observations in session k, and

$$\tilde{\psi}_1 = \frac{1}{2}\sum_k \sum_{i,j} \frac{\left[\sum_{r \in C_{ijk}} (\ln t_{ijkr} - h_{ijk}^{(s)})^2 + \sum_{r \in D_{ijk}} (\ln t_{ijkr} - h_{ijk}^{(d)})^2\right]}{q_k^2}. \tag{23}$$

Define

$$\psi_1^* = \frac{1}{2}\sum_k \sum_{i,j} \frac{[n_{ijk}^{(s)}(\bar{g}_{ijk}^{(s)} - h_{ijk}^{(s)})^2 + n_{ijk}^{(d)}(\bar{g}_{ijk}^{(d)} - h_{ijk}^{(d)})^2]}{q_k^2}, \tag{24}$$

(where $n_{ijk}^{(s)}$ and $n_{ijk}^{(d)}$ are, as defined previously, the numbers of elements in C_{ijk} and D_{ijk}, respectively) and

$$\psi_1^{**} = \frac{1}{2}\sum_k \sum_{i,j} \frac{\left[\sum_{r \in C_{ijk}} (\ln t_{ijkr} - \bar{g}_{ijk}^{(s)})^2 + \sum_{r \in D_{ijk}} (\ln t_{ijkr} - \bar{g}_{ijk}^{(d)})^2\right]}{q_k^2}, \tag{25}$$

where

$$\bar{g}_{ijk}^{(s)} = \frac{\sum_{r \in C_{ijk}} \ln t_{ijkr}}{n_{ijk}^{(s)}}, \tag{26}$$

and

$$\bar{g}_{ijk}^{(d)} = \frac{\sum_{r \in D_{ijk}} \ln t_{ijkr}}{n_{ijk}^{(d)}}. \tag{27}$$

The $\bar{g}_{ijk}^{(s)}$ and $\bar{g}_{ijk}^{(d)}$ above are simple algebraic means of log reaction times over replications. When either $n_{ijk}^{(s)}$ or $n_{ijk}^{(d)}$ is zero, $\bar{g}_{ijk}^{(s)}$ or $\bar{g}_{ijk}^{(d)}$ is undefined, and consequently the corresponding terms in (24) and (25) should be deleted from ψ_1^* and ψ_1^{**}. It can be easily verified that $\tilde{\psi}_1 = \psi_1^* + \psi_1^{**}$. Note that ψ_1^{**} does not involve any model parameters except q_k^2. Hence it can be ignored when parameters other than q_k^2 are estimated. This would save a lot of computer time, since ψ_1^* does not involve summations over replications. Furthermore, the numerator of ψ_1^{**} (which does involve summations over replications) has to be calculated once for all computations. Note also that $\bar{g}_{ijk}^{(s)}$ and $\bar{g}_{ijk}^{(d)}$ are minimum sufficient for $h_{ijk}^{(s)}$ and $h_{ijk}^{(d)}$, respectively. The log likelihood can now be written as

$$\ln L = \frac{1}{2} \sum_k N_k \ln q_k^2 - \psi_1^* - \psi_1^{**} + \psi_2. \tag{28}$$

The log likelihood can be maximized by one of a variety of numerical methods for optimization. We use Fisher's scoring algorithm for solving maximum likelihood equations in combination with the partial alternating least squares. The scoring algorithm has successfully been employed in a number of similar situations [Takane, 1978; Takane, 1981; Takane & Carroll, 1981; Takane, 1982]. This algorithm updates estimates of parameters ξ by solving $\varepsilon^{(t)}\mathbf{I}(\xi^{(t)})(\xi^{(t+1)} - \xi^{(t)}) = \mathbf{u}(\xi^{(t)})$ for $\xi^{(t+1)}$, where $\varepsilon^{(t)}$ is a stepsize parameter at iteration t,

$$\mathbf{u}(\xi^{(t)}) = \left(\frac{\partial \ln L}{\partial \xi}\right)_{|\xi = \xi^{(t)}} \tag{29}$$

and

$$\mathbf{I}(\xi^{(t)}) = E\left[\left(\frac{\partial \ln L}{\partial \xi}\right)\left(\frac{\partial \ln L}{\partial \xi}\right)'\right]_{|\xi = \xi^{(t)}}. \tag{30}$$

The $\xi^{(t)}$ is the current estimates of ξ. The Moore-Penrose inverse of the information matrix evaluated at the maximum of $\ln L$ gives variance-covariance estimates of estimated parameters [Ramsay, 1978].

The scoring algorithm may not be very efficient when it is used to update a large number of parameters simultaneously. When a large number of parameters are involved, we may partition the total set of parameters into several subsets, and update each subset successively within each iteration. As a consequence, more iterations are needed for convergence, but each iteration can be done much more efficiently, and there is usually a great economy in terms of total computation time. In the current MAXRT we first update parameters in the representation model, $\boldsymbol{\theta}$, with the other parameters ($\tilde{b}_k$, p_k, a_k and q_k^2) being fixed. We then update $\tilde{b}_k$, p_k, a_k and q_k^2 in this order. In updating one set of parameters those in the other subsets are always fixed at their current estimates.

One additional advantage of this conditional estimation scheme is that parameters specific to experimental sessions (those with subscript k) can be updated independently within each subset of parameters. (For example, p_k and $p_{k'}$ ($k \neq k'$) can be updated independently of each other.) This is because derivatives of any terms related to k in the log likelihood with respect to parameters related to k' ($k \neq k'$) are always zero. The infor-

mation matrix in this case is diagonal, implying that their estimates do not influence each other, given the other subsets of parameters. When the information matrix is diagonal, the solution of the updating equations is rather trivial. When the session-specific parameters are assumed equal, they are not session-specific anymore, but then there is only a single parameter in each subset, and they can be updated as efficiently as before.

The conditional estimate of a_k can be obtained in a closed form (again provided that the other parameters are fixed). The a_k is related to only ψ_1^* in (28), which is quadratic in a_k. Consequently the stationary equation involving a_k is linear. This estimate is given by

$$a_k = \frac{(g_{..k} - p_k z_{..k})}{N_k}, \tag{31}$$

where

$$g_{..k} = \sum_{i,k} (n_{ijk}^{(s)} \bar{g}_{ijk}^{(s)} + n_{ijk}^{(d)} \bar{g}_{ijk}^{(d)}),$$

$$z_{ijk}^{(s)} = \tilde{b}_k - \ln d_{ij}, \qquad z_{ijk}^{(d)} = -z_{ijk}^{(s)},$$

and

$$z_{..k} = \sum_{i,j} (n_{ijk}^{(s)} z_{ijk}^{(s)} + n_{ijk}^{(d)} z_{ijk}^{(d)}).$$

When $a_k = a$ for all k, it is given by

$$a = \frac{(g_{...} - pz_{...})}{\left(\frac{\sum_k N_k}{q_k^2}\right)}, \tag{32}$$

where $p_k = p$ is also assumed for all k, and

$$g_{...} = \sum_k \left(\frac{N_k g_{..k}}{q_k^2}\right),$$

$$z_{...} = \sum_k \left(\frac{N_k z_{..k}}{q_k^2}\right).$$

It is important to have good initial estimates for both faster convergence and avoiding nonglobal maxima. They are estimated as follows. First, initial estimates of stimulus dissimilarities are obtained by taking means of reciprocal reaction times ($1/T$) from correct "different" judgments. Parameters in the representation models are then estimated by assuming that they are ratio measures of true stimulus dissimilarities. For the Minkowski power distance model we apply the Young-Householder procedure [Torgerson, 1958] assuming that the model is tentatively euclidean. For the linear model and Tversky's feature matching model we apply the ordinary linear least squares regression analysis. (Tversky's feature matching model reduces to a linear model if $\beta_1 = 1$ is assumed in (69).) Threshold values are estimated by first taking averages of dissimilarities corresponding to "different" judgments and "same" judgments separately and then further averaging the two averages. Once d_{ij} and b_k are estimated, we may plug them in (24) to obtain p_k and a_k. Finally, q_k^2 is obtained by

$$q_k^2 = \frac{(g_{..k} - p_k z_{..k})^2}{N_k}.$$

The above initialization method seems to work reasonably well. In the Monte Carlo study to be reported in the discussion section, we obtained solutions using both the "true"

parameter values and the built-in initialization method. In all cases we obtained virtually indistinguishable results.

Derivatives

In this section we collect derivatives necessary to "operationalize" the algorithm. Let us first define several quantities in order to simplify the notation. Let

$$\phi_{ijk} = \phi(v_{ijk}), \tag{33}$$

where

$$v_{ijk} = \frac{-p_k(\tilde{b}_k - \ln d_{ij})}{q_k}, \tag{6'}$$

$$G_{ijk}^{(s)} = \frac{n_{ijk}^{(s)}(\bar{g}_{ijk}^{(s)} - h_{ijk}^{(s)})}{q_k^2}, \tag{34}$$

$$G_{ijk}^{(d)} = \frac{n_{ijk}^{(d)}(\bar{g}_{ijk}^{(d)} - h_{ijk}^{(d)})}{q_k^2}, \tag{35}$$

and

$$Q_{ijk}^{*} = \frac{n_{ijk}^{(s)}}{Q_{ijk}} - \frac{n_{ijk}^{(d)}}{1 - Q_{ijk}}. \tag{36}$$

Parameters in the representation models of dissimilarities ($\boldsymbol{\theta}$), the log threshold parameters ($\tilde{b}_k$), and the slope parameter (p_k) are related to only ψ_1^* and ψ_2 in (28). Consequently derivatives of the log likelihood with respect to these parameters all have the same structure:

$$\frac{\partial \ln L}{\partial \xi} = \frac{\partial(-\psi_1^*)}{\partial \xi} + \frac{\partial \psi_2}{\partial \xi}, \tag{37}$$

where ξ is either $\boldsymbol{\theta}$, $\tilde{b}_k$ or p_k.

For $\boldsymbol{\theta}$ we have

$$\frac{\partial(-\psi_1^*)}{\partial \boldsymbol{\theta}} = \sum_k \sum_{i,j} \left\{ G_{ijk}^{(s)} \left(\frac{\partial h_{ijk}^{(s)}}{\partial \boldsymbol{\theta}} \right) + G_{ijk}^{(d)} \left(\frac{\partial h_{ijk}^{(d)}}{\partial \boldsymbol{\theta}} \right) \right\}, \tag{38}$$

where

$$\frac{\partial h_{ijk}^{(s)}}{\partial \boldsymbol{\theta}} = -\frac{p_k}{d_{ij}} \left(\frac{\partial d_{ij}}{\partial \boldsymbol{\theta}} \right) \quad \text{and} \quad \frac{\partial h_{ijk}^{(d)}}{\partial \boldsymbol{\theta}} = -\frac{\partial h_{ijk}^{(s)}}{\partial \boldsymbol{\theta}}, \tag{39}$$

and

$$\frac{\partial \psi_2}{\partial \boldsymbol{\theta}} = \sum_k \sum_{i,j} Q_{ijk}^{*} \left(\frac{\partial Q_{ijk}}{\partial \boldsymbol{\theta}} \right), \tag{40}$$

where

$$\frac{\partial Q_{ijk}}{\partial \boldsymbol{\theta}} = -\frac{\phi_{ijk} p_k}{q_k d_{ij}} \left(\frac{\partial d_{ij}}{\partial \boldsymbol{\theta}} \right). \tag{41}$$

The additional derivative ($\partial d_{ij}/\partial \boldsymbol{\theta}$) necessary to fit a specific representation model of dissimilarities will be given where it is specified.

For the log threshold $\tilde{b}_k$, we obtain

$$\frac{\partial(-\psi_1^*)}{\partial \tilde{b}_k} = \sum_{i,j} (G_{ijk}^{(s)} + G_{ijk}^{(d)}) p_k, \tag{42}$$

and

$$\frac{\partial \psi_2}{\partial \tilde{b}_k} = \sum_{i,j} Q^*_{ijk}\left(\frac{\partial Q_{ijk}}{\partial \tilde{b}_k}\right) \tag{43}$$

where

$$\frac{\partial Q_{ijk}}{\partial \tilde{b}_k} = -\frac{\phi_{ijk}\, p_k}{q_k}. \tag{44}$$

When $\tilde{b}_k = \tilde{b}$ for all k, $\partial \ln L/\partial \tilde{b} = \sum_k (\partial \ln L/\partial \tilde{b}_k)$.

For the slope parameter p_k, we have

$$\frac{\partial(-\psi_1^*)}{\partial p_k} = \sum_{i,j} (G^{(s)}_{ijk} - G^{(d)}_{ijk})(\tilde{b}_k - \ln d_{ij}), \tag{45}$$

and

$$\frac{\partial \psi_2}{\partial p_k} = \sum_{i,j} Q^*_{ijk}\left(\frac{\partial Q_{ijk}}{\partial p_k}\right), \tag{46}$$

where

$$\frac{\partial Q_{ijk}}{\partial p_k} = -\frac{\phi_{ijk}(\tilde{b}_k - \ln d_{ij})}{q_k}. \tag{47}$$

When $p_k = p$ for all k, $\partial \ln L/\partial p$ is obtained by the sum of $\partial \ln L/\partial p_k$.

For q_k^2 we obtain

$$\frac{\partial \ln L}{\partial q_k^2} = \frac{1}{2}\frac{[N_k - (\psi_1^* + \psi_1^{**})]}{q_k^2} + \sum_{i,j} Q^*_{ijk}\left(\frac{\partial Q_{ijk}}{\partial q_k^2}\right), \tag{48}$$

where

$$\frac{\partial Q_{ijk}}{\partial q_k^2} = \frac{\phi_{ijk}\, p_k(\tilde{b}_k - \ln d_{ij})}{2q_k^2}. \tag{49}$$

When $q_k^2 = q^2$ for all k, $\partial \ln L/\partial q^2$ is again obtained by the sum of $\partial \ln L/\partial q_k^2$.

The information matrix also has the same structure for $\boldsymbol{\theta}$, b_k and p_k, i.e.,

$$\mathbf{I}(\xi) = \mathbf{I}_1(\xi) + \mathbf{I}_2(\xi), \tag{50}$$

where $\mathbf{I}_1$ and $\mathbf{I}_2$ correspond with ψ_1^* and ψ_2, respectively, and are given by

$$\mathbf{I}_1(\xi) = \sum \frac{\left[n^{(s)}_{ijk}\left(\frac{\partial h^{(s)}_{ijk}}{\partial \xi}\right)\left(\frac{\partial h^{(s)}_{ijk}}{\partial \xi}\right)' + n^{(d)}_{ijk}\left(\frac{\partial h^{(d)}_{ijk}}{\partial \xi}\right)\left(\frac{\partial h^{(d)}_{ijk}}{\partial \xi}\right)'\right]}{q_k^2}, \tag{51}$$

and

$$\mathbf{I}_2(\xi) = \sum\left[\frac{n_{ijk}}{Q_{ijk}(1 - Q_{ijk})}\left(\frac{\partial Q_{ijk}}{\partial \xi}\right)\left(\frac{\partial Q_{ijk}}{\partial \xi}\right)'\right], \tag{52}$$

where $n^{(s)}_{ijk} + n^{(d)}_{ijk}$. The summations above extend over i and j. They may also extend over k depending on ξ. For q_k^2, the $\mathbf{I}_2$ part in (52) remains the same. The $\mathbf{I}_1$ part should be modified into

$$\mathbf{I}_1(q_k^2) = \frac{N_k}{q_k^4}, \tag{53}$$

(or $\mathbf{I}(q^2) = N^*/q^4$ where $N^* = \sum_k N_k$.) Expressions for $\mathbf{I}_1$ and $\mathbf{I}_2$ were readily obtained by noting the equivalence of the Gauss-Newton method for certain weighted least squares problems and the scoring algorithm for maximum likelihood estimation when the assumed population distribution is one of exponential type.

Some indication that the procedure outlined above indeed works will be given in the discussion section.

Representation Models of Dissimilarities and their Goodness of Fit

In this section we present some empirical results obtained by the method described in the previous sections. In stating the model we intentionally left the representation model unspecified. We tried a number of alternatives to find the best representation model for our data. Our strategy here is strictly empirical in the sense that we actually fit various candidate models to the same set of data and choose the one which fits the data best according to some statistical criterion.

In evaluating the goodness of fit of the models we use the AIC statistic [Akaike, 1974] defined by

$$\text{AIC}(\pi) = -2 \ln L + 2 n_\pi, \tag{54}$$

where n_π is the effective number of parameters in model π. Only relative magnitudes of AIC are meaningful. (The model associated with a smaller value of AIC is considered the better model.) One of the major advantages of the AIC statistic is that it may be used to compare any models, contrary to the asymptotic chi-square goodness of fit test which typically requires that one of the models to be compared is a constrained counterpart of the other. Some useful applications of AIC to psychometric models can be found in Takane [1978, 1981], Takane & Carroll [1981] and Winsberg & Ramsay [1981]. Takane [1981] in particular discusses other advantages as well as some limitations of the AIC statistic.

When one of the two models to be compared is nested within the other (e.g., comparisons between different dimensionalities), we may apply the asymptotic chi-square significant test. Let L_1 and L_2 denote the maximum likelihoods of more restrictive and less restrictive models, respectively. We then calculate $C = 2(\ln L_2 - \ln L_1)$, and compare this value against an appropriate critical value of chi square with degrees of freedom equal to the difference in the effective numbers of parameters in the two models. Where applicable, we use both AIC and the asymptotic chi square to identify the best fitting model.

The Data

The stimuli we used were line drawings of 8 front-view faces derived from a larger set of faces [Sergent, 1982]. The faces were made by combining each of two levels of three features: hair style (*H*), eyes and eyebrows (*E*), and jaw and chin (*J*). Eight faces were thus constructed. The values taken on the three features are listed in Table 1, and the faces themselves are displayed in Figures 2 and 3. Each possible combination of *different* face pairs was photographed twice, one face above the other, so that for each pair a face was once above a central fixation point and once below. Pairs of *same* faces were photographed once each. The stimuli were rear-projected onto a rectangular translucent screen, 18 × 13 cm, the bottom of the above-face and the top of the below-face falling .6 cm from the black central fixation point. The faces, when projected, subtended a visual angle of 3°33′ in height and 2°52′ in width. The slides were placed in a random-access projector. Presentation was controlled by a computer, which selected the slides in a random order. Each *different* pair appeared eight times, and each *same* pair appeared 28 times, yielding a total of 448 trials per session, with an equal number of *same* and *different* pairs.

The subjects were tested individually. Each subject sat about 80 cm in front of the screen, in a dark room, his head adjusted in a chin-and-forehead rest so that his eyes were constantly at the level of the central fixation point. A 500-msec tone warned the subject to fixate the central point. The stimuli appeared 1 sec after onset of the warning tone, and the subject was to press the right key if the faces were *same* and to press the left key if the faces were *different*. The stimuli remained on the screen for 1 sec. The subjects were told to respond as quickly and accurately as possible. The response deadline was set at 1.5 sec, and no record was made of reaction times exceeding the deadline (There were only a few such trials.) Each subject was tested in seven sessions of 448 trials. The first session served as practice, and the six other were experimental and run on consecutive days. Two male subjects participated in the experiment. They both had normal vision.

After several weeks the above procedure was repeated with the same subjects, but with upside-down stimulus presentation. The results obtained from these data were very similar to those obtained from the upright condition, and will not be presented in this paper. However, they serve as a sort of replication data.

For the reason stated earlier (i.e., the incompatibility of the log normal distribution with zero dissimilarities in the Minkowski power distance model) only those portions of the data pertaining to *different* pairs (224 trials per session) were analyzed.

As mentioned earlier, we fit the Minkowski power distance model, the linear model and Tverksy's feature matching model to the above data. In fitting each model we state some theoretical and/or empirical motivations behind the model, which presumably reflects the way the subject performs the required task.

Linear Model 1: The Dominance Metric

There are two classes of models which frequently appear in the reaction time literature. One is based on the analytic process assumption, and the other based on the holistic process assumption [Nickerson, 1972]. In the analytic process model the subject is supposed to make an independent decision ("same" or "different") for each discriminable feature of the stimuli, an overall "different" decision being made when a difference is detected, and a "same" decision resulting from an absence of difference. In the holistic process model, on the other hand, information on different features of stimuli is first integrated into a global judgment of dissimilarity on which the same-different decision is

Table 1. The stimuli and their features.

	Features		
Stimulus	hair (H)	eye (E)	jaw (J)
1	Long	Dark	Angular
2	Short	Dark	Angular
3	Short	Light	Angular
4	Long	Light	Angular
5	Long	Dark	Round
6	Short	Dark	Round
7	Short	Light	Round
8	Long	Light	Round

supposedly made. The representation model we consider first belongs to the analytic class.

Suppose the subject attends to one discriminable feature of stimuli at a time. If the pair of stimuli differ in the first feature, he immediately gives a "different" response; otherwise he moves on to the next feature. He repeats the same process until he encounters a feature in which the two stimuli differ, or until he exhausts all relevant features without finding any differences between the stimuli. If he follows this self-terminating serial scanning process in a consistent manner, the dominance metric with externally identifiable dimensions is obtained. (See below.)

The discriminability of the facial features from which the faces were constructed was assessed in a previous experiment using reaction times to judge the two instances of each feature as "same" or "different" [Sergent, 1982, Experiment 2]. It was found that the jaws were discriminated significantly faster than the hairstyles which were in turn discriminated faster than the two types of eyes. While this may provide an objective measure of the respective discriminability of each pair of features considered individually, it does not necessarily imply that faces are compared serially as a function of the discriminability of their features. Ellis [1975] pointed out that the upper part of the face was paid more attention to than the lower part, and that subjects may analyze faces in a top-to-bottom sequence. However, Davies, Ellis, and Shepherd [1977] have shown that the degree of dissimilarity of the facial features was also an important factor in face comparison and Sergent [1982] suggested that the scanning strategy involved in face perception may combine a top-to-bottom analysis with a sensitivity to the more discriminable features. This indicates that objective and subjective salience may not correspond.

Remember that the stimuli were factorially constructed by combining two levels in each of three features. Thus, each pair of stimuli may be characterized by a set of three binary variables, x_1, x_2 and x_3, where

$$x_m = \begin{cases} 0, & \text{if the two stimuli are } \textit{same} \text{ in feature } m, \\ 1, & \text{otherwise.} \end{cases} \tag{55}$$

Note that there are eight possible patterns of zero and one in three binary variables. All possible pairs of stimuli (28 *different* pairs and 8 *same* pairs) can be classified into one of these eight groups. Suppose that features 1, 2, and 3 are salient in that order, and that the subject consistently checks these features in the same order. Then he must follow one of the four paths depicted in Figure 1 to reach his final judgment. If the two stimuli differ in the most salient feature, the scanning process ends in Stage 1 [Path (4)]. Stimulus pairs belonging to the last four groups should have the shortest reaction times. Furthermore, the reaction times should be identical for the four groups. If, on the other hand, the two stimuli do not differ in the most salient feature, the next feature is examined. If they differ in this feature, the process terminates in Stage 2 [Path (3)]. Stimulus pairs, identical in the first, but different in the second feature (group 3 and group 4), should have the next shortest reation times. Again, the reaction times should be identical for the two groups. The reaction times should be longest for the remaining two groups in which stimulus pairs are identical in the first two features. In these cases the subject has to go into Stage 3 in order to make a "same" judgment [Path (1)] or a "different" judgment [Path (2)].

The essential feature of this strict sequential search hypothesis is that the reaction time is a function of only the most salient dimension on which two stimuli differ. The most salient dimension is the one on which the two stimuli differ most. This leads to the dominance metric model, which is written, in the general form, as

$$d_{ij} = \max_m |\theta_{im} - \theta_{jm}|, \tag{56}$$

[Coombs, 1964]. Since for a "different" response the reaction time is inversely related to

the dissimilarity between two stimuli, a shorter reaction time is expected for stimuli that differ on a more salient dimension. However, once two stimuli differ on a more salient dimension, possible differences on less salient dimensions have no effects on the overall dissimilarity between the two stimuli.

Note that (56), as it is, presupposes no prescription of relevant dimensions. The dominance metric is extremely difficult to fit in this general form. However, the situation is greatly simplified, when, as in the present case, we may assume that the relevant dimensions are known, and coincide with the defining features of the stimuli. (Whether this assumption is justified is an empirical question.) The dominance metric in this case reduces to a simple linear model. Let

$$\alpha_m = |\theta_{im} - \theta_{jm}| > 0 \tag{57}$$

for any stimulus pair whose members differ on dimension m. Here $m = H$(hair), E(eye) or J(jaw). The dissimilarity between stimuli i and j can then be written as

$$d_{ij} = \sum_m x_{ijm}\, \alpha_m \tag{58}$$

where $x_{ijm} = 1$, if m is the most salient feature in which those stimuli differ, and $x_{ijm} = 0$, otherwise. We may estimate α_m directly. The only additional derivative necessary to fit (58) is

$$\frac{\partial d_{ij}}{\partial \alpha_m} = x_{ijm}. \tag{59}$$

The dominance metric, as expressed in the form of (58), was fitted under all six possible salience orders among three features. The results are reported in Table 2. The main entries of the table are the values of the AIC statistic. The n_π is taken to be 3 (α_H, α_E and α_J); parameters commonly used in all solutions are not counted, since for the purposes of comparing goodness of fit only relative magnitude of AIC matters.

Among the six solutions only two are admissible for each subject in the sense that

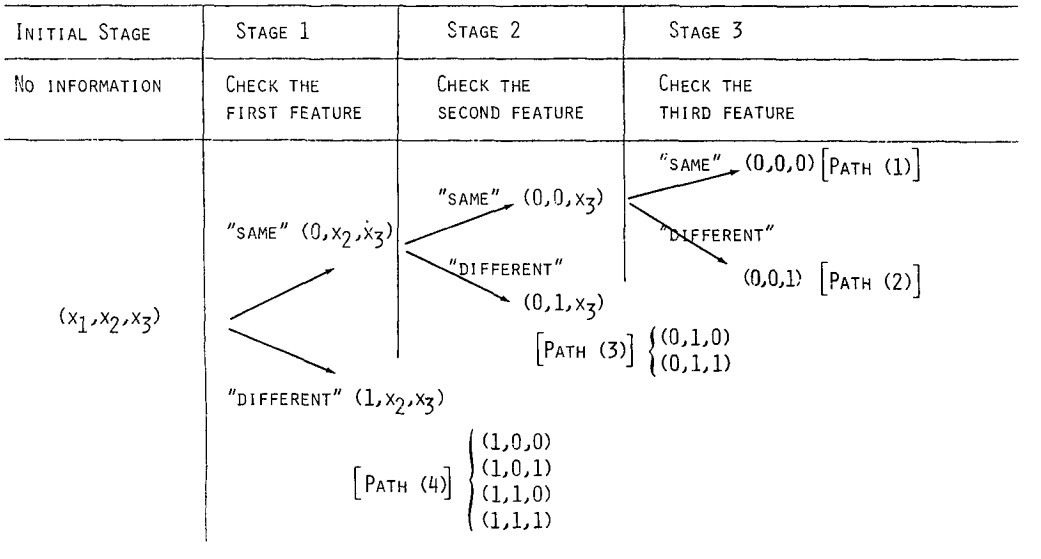

FIGURE 1.
The serial scanning process for same-different judgment.

Table 2. The values of AIC obtained by fitting the dominance metric model under six salience orders among features.

Dominance Order	Subject 1	Subject 2
1. H>J>E	-5026.5*	-3761.1
2. H>E>J	-4939.3	-3777.4*
3. J>H>E	-4936.2	-3255.7
4. J>E>H	-4789.1	-3066.7
5. E>H>J	-4733.2	-3341.0
6. E>J>H	-4700.1	-3062.0

The * indicates the minimum AIC solution.

Admissible solutions are underlined.

AIC = -2 ln(L) + 6.

estimated α's agree with prescribed salience orders (e.g., if $H > J > E$ is assumed, it must be that $\alpha_H > \alpha_J > \alpha_E$, etc.). The two solutions are underlined in the table. For both subjects the salience order of $H > J > E$ is among the admissible solutions. However, $J > H > E$ is admissible only for subject 1, whereas $H > E > J$ is admissible only for subject 2. For subject 1 both H and J are far more dominant than E, while the contributions of H and J do not differ greatly. For subject 2, however, H is far more dominant than both J and E, which themselves do not differ very much.

The minimum AIC criterion indicates that the salience order of $H > J > E$ fits the data best for subject 1, while $H > E > J$ fits best for subject 2. In both cases the hair is the most dominant feature. This is more or less consistent with the suggestion that spatially dense and/or global features tend to be processed faster [Sergent & Bindra, 1981]. The hair is both spatially dense and global. It is interesting to note that dissimilarity ratings of the same set of faces have revealed that either E or J is the most dominant feature, and H the least dominant. Perhaps the way different features are processed is different in time-pressed situations.

The dominance metric model fitted above, however, is perhaps too simple-minded. It assumes that the same scanning order is strictly maintained throughout experimental sessions. Mean log reaction times were calculated for the eight groups of stimulus pairs given in Figure 1, and are listed in Table 3. They clearly tend to decrease as more features differ between two stimuli. This contradicts the prediction from the strict sequential scanning hypothesis. Although there are other possibilities (e.g., probabilistic sequential scanning hypothesis), it may support the view that features are not processed in an independent manner [Lockhead, 1979]. That is, information on different features may somehow be combined before the final same-different decision is made [Lockhead, 1972; Monahan & Lockhead, 1977; Miller, 1978].

Linear Model 2: The Context Model

The problem now is how information on different features (or dimensions) is combined. We discuss Medin and Schaffer's [1978] cue context model in this section, and

Tversky's feature matching model and the euclidean model in the subsequent sections.

The context model defines similarity between two stimuli by a multiplicative rule. Let s_{ij} represent the similarity between stimuli i and j. Then the model is formally written as

$$s_{ij} = \prod_m f_{ijm}, \tag{60}$$

where

$$f_{ijm} = \begin{cases} 1, & \text{if stimuli } i \text{ and } j \text{ are } \textit{same} \text{ in feature } m, \\ \tau_m, & \text{otherwise.} \end{cases}$$

The τ_m takes values between zero and one (i.e., $0 < \tau_m < 1$). Taking the log of (60) we obtain

$$\begin{aligned} -\ln s_{ij} &= -\sum_m \ln f_{ijm} \\ &= \sum_m x_{ijm}\, \alpha_m, \end{aligned} \tag{61}$$

where

$$x_{ijm} = \begin{cases} 0, & \text{if } f_{ijm} = 1, \\ 1, & \text{if } f_{ijm} = \tau_m, \end{cases} \tag{62}$$

and

$$\alpha_m = -\ln \tau_m. \tag{63}$$

Since α_m is always positive, $-\ln s_{ij}$ is always nonnegative. Thus, $-\ln s_{ij}$ may be viewed as representing dissimilarity (d_{ij}) between stimuli i and j. That is,

$$d_{ij} = \sum_m x_{ijm}\, \alpha_m. \tag{61'}$$

This is linear in α_m.

It can be shown that model (61') can be interpreted as a special instance of the city-block distance model defined by

$$d_{ij} = \sum_m |\theta_{im} - \theta_{jm}|, \tag{64}$$

Table 3. Mean log reaction times (ms) for eight groups of stimulus pairs.

Features
0 same
1 different

	HJE	Subject 1	Subject 2
0.	000	6.910	6.860
1.	001	6.923	6.863
2.	010	6.877	6.846
3.	011	6.852	6.806
4.	100	6.847	6.677
5.	101	6.810	6.646
6.	110	6.796	6.623
7.	111	6.781	6.616

where θ_{im} is the coordinate of stimulus i on dimension m. It is assumed that each feature represents a dimension. Then $|\theta_{im} - \theta_{im}|$ is either zero or some positive constant for any pair of stimuli i and j. If it is zero, we set $x_{ijm} = 0$ in (61'), otherwise, we set $x_{ijm} = 1$ and $\alpha_m = |\theta_{im} - \theta_{jm}|$. Then (64) reduces to (61').

The context model was fitted to the present data in the form of (61'). The values of the AIC statistic obtained were -5133.2 and -3822.3 for subject 1 and subject 2, respectively. (The *df* for the representation model is 3 as in the case of dominance metric.) For both subjects they are substantially smaller than the minimum AIC's obtained from the dominance metric. This result conforms to our previous observation in Table 3.

The context model of Medin and Schaffer assumes that contributions of *same* features to overall dissimilarities are zero regardless of features and values the stimuli take on the features. For example, two stimuli are *same* in feature H, whether they both have long hair or they both have short hair, and the contribution of the hair dimension to the overall dissimilarity is assumed zero in both cases. We next consider a model which is free from this restriction.

The Feature Matching Model

Tversky's feature matching model postulates that the similarity between stimuli i and j is given by

$$s_{ij} = \beta_1 v(Q_i \cap Q_j) - \beta_2 v(Q_i - Q_j) - \beta_3 v(Q_j - Q_i), \tag{65}$$

where Q_i and Q_j are the sets of features possessed by stimuli i and j, respectively, v is a measure defined on a set of features, and β_1, β_2 and β_3 (all assumed nonnegative) are the weights representing the importance of the three components (three measures) in obtaining the overall similarity between the two stimuli. The $v(Q_i \cap Q_j)$ represents the (unweighted) contribution of features commonly possessed by stimuli i and j, $v(Q_i - Q_j)$ represents the distracting effect of features possessed by stimulus i but not by stimulus j, and $v(Q_j - Q_i)$ the distracting effect of features possessed by stimulus j but not by stimulus i. It is clear from (65) that the v-scale is determined up to an interval scale [Tversky, 1977], and consequently features common to all stimuli may be excluded from (65).

Some simplifying assumptions are necessary to fit the feature matching model. The following assumptions are similar to those made by Keren and Baggen [1981] who recently fitted the feature matching model to the confusion data.

(I) *Feature Independence.*

$$v(Q) = \sum_{w \in Q} v(w), \tag{66}$$

where w is a feature, and Q is a collection of features. The above equation implies that the combined effect of features is obtained by simple addition of contributions from elementary features, and that there are no interactive effects of more than one feature.

(II) *Symmetry.*

$$\beta_2 = \beta_3 = \tfrac{1}{2} \tag{67}$$

We simply make this assumption, since no systematic assymmetric effects are foreseen [Podgorny & Garner, 1979].

(III) *Log Linearity.*

$$\ln d_{ij} = -s_{ij} \tag{68}$$

The s_{ij} in (65) gives an interval scale of similarity. The $\exp(-s_{ij})$ will effectively transform the interval scale of similarity into a ratio scale of dissimilarity.

Finally, we assume that the three defining features of the stimuli (H, J, and E) are the

only relevant features. Each feature had two values. For the purposes of present analysis each level of the three features is taken as a "feature". (We distinguish the two uses of the word feature by putting quotation marks to the latter.) Thus, there are six "features" altogether. Each "feature" is either shared by a pair of stimuli, possessed by only one member, or not possessed by either member. Let

$$x_{ijm}^{(C)} = \begin{cases} 1, & \text{if stimuli } i \text{ and } j \text{ share "feature" } m, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$x_{ijm}^{(D)} = \begin{cases} 1, & \text{if either stimulus } i \text{ or } j \text{ (but not both) possesses "feature" } m \\ 0, & \text{otherwise.} \end{cases}$$

Then $-s_{ij}$ may be written as

$$-s_{ij} = \ln d_{ij} = \sum_m (\beta_1 x_{ijm}^{(C)} + x_{ijm}^{(D)})\alpha_m, \tag{69}$$

where α_m is the effect of "feature" m on the overall similarity between the two stimuli. The above model is slightly more complicated than the simple linear model due to the multiplicative parameter β_1. Derivatives necessary to fit (69) are given as follows:

$$\frac{\partial \ln d_{ij}}{\partial \alpha_m} = \beta_1 x_{ijm}^{(C)} + x_{ijm}^{(D)}, \tag{70}$$

and

$$\frac{\partial \ln d_{ij}}{\partial \beta_m} = \sum_m x_{ijm}^{(C)} \alpha_m. \tag{71}$$

The AIC values obtained by fitting (69) were -5125.1 (6 df) for subject 1, and -3818.9 (6 df) for subject 2. (There are 7 parameters in the model, but due to the interval scale nature of α_m, one of them can be arbitrarily fixed.) For both subjects they indicate poorer fits of the feature matching model than the context model discussed in the previous section. There are at least two possible reasons for this: It might be that the additional parameters introduced for possible differential effects of *same* features did not significantly improve the goodness of fit of the model. Or it may be that the log linearity assumption (68) is unrealistic (though no other assumptions are readily obvious). At the moment, however, there is no empirical evidence favoring one over the other.

Note that in all the models discussed so far we have assumed feature independence and external indentifiability of relevant features. It still remains a question whether the three physical properties used to construct the stimuli exhaust all relevant psychological dimensions that the subject utilizes in reaching same-different decisions. It is possible that some other aspects of the stimuli (e.g., length of forehead) may have been critical. It is also possible that an entirely new psychological dimension emerges from the objective physical components of the face. In fact, faces, as other visual shapes, have configural properties that result from some form of interrelation between the stimulus components and that exist in addition to the component properties [Garner, 1978].

The additive clustering (ADCLUS) analysis of Shepard and Arabie [1979] is interesting in this regard. The model decomposes the overall similarity between two stimuli into a weighted sum of contributions from features commonly possessed by the two stimuli. The most appealing aspect of the model is that it does not presuppose the knowledge of relevant features. Rather, the method tries to identify a set of features necessary and sufficient to account for the set of observed similarities. However, the ADCLUS model is

extremely difficult to fit due to the discrete nature of the model [Arabie & Carroll, 1980]. If, on the other hand, the features are assumed known *a priori*, this model reduces to a special case of Tversky's feature matching model ($\beta_1 = \beta_2 = 0$), which is easier to fit, but then the ADCLUS model loses its most appealing feature. Of course it is possible to identify potentially relevant features based on the least squares fitting of the ADCLUS model [Arabie & Carroll, 1980] to reaction time data. We may then fit only a few candidate models by the present method for further model comparisons.

The Euclidean Model

Another class of models which do not presuppose *a priori* identification of relevant stimulus attributes are the distance models used in the conventional multidimensional scaling (MDS). In fact one of the most appealing features of MDS is that we do not have to prescribe what aspects of stimuli are psychologically relevant, or how many aspects (or dimensions) are necessary to account for the data.

The distance model most widely used in multidimensional scaling is the Minkowski power distance model [Kruskal, 1964a, b] defined by

$$d_{ij} = \left\{ \sum_{m=1}^{M} |\theta_{im} - \theta_{jm}|^u \right\}^{1/u}, \tag{72}$$

where θ_{im} is the coordinate of stimulus i on dimension m, u (≥ 1) is the Minkowski power, and M is the dimensionality of the representation space. The derivative necessary to fit (72) is given by

$$\frac{\partial d_{ij}}{\partial \theta_{tm}} = (\delta_{it} - \delta_{jt}) |\theta_{im} - \theta_{jm}|^{u-1} \text{ signum } (\theta_{im} - \theta_{jm}) \frac{}{d_{ij}^{u-1}}, \tag{73}$$

where $\delta_{\cdot\cdot}$ is a Kronecker delta. The Minkowski power distance model generates a family of distance models which differ in how dimensional differences contribute to the overall distance, depending on the value of u. When $u = 2$, we obtain the familiar euclidean distance model.

The euclidean distance model was fitted in four two to five dimensions. The results are reported in Table 4. The values of AIC indicate that the best solution is four dimensional for both subjects. The asymptotic chi-squares also indicate that it is the best solution. The chi-squares representing the difference between the four and five dimensional solutions are 1.8 with 3 df for subject 1 and 4.2 with 3 df for subject 2. The fifth dimension thus seems nonsignificant. However, the fourth dimension seems significant. The chi-squares are 12.6 and 15.6 for subject 1 and subject 2, respectively, each with 4 df.

The AIC values of -5146.9 and -3850.2 are also both substantially smaller than those obtained from the context model, indicating that the four-dimensional euclidean model is the best model so far. Figures 2 and 3 display derived four-dimensional stimulus configurations. Stimulus coordinates on the fourth dimension are given in the lower right corner. The first three dimensions of the configurations roughly correspond with the three defining features of the stimuli, namely hair, jaw and eye, in this order. The third dimension in the four-dimensional solutions shows some curious pattern, which does not show up in the three-dimensional solution; on this dimension stimuli 2 and 3, and stimuli 5 and 8 are distinctly closer to each other than stimuli 1 and 4, or stimuli 6 and 7. This is partly compensated by the fourth dimension, on which those closer stimulus pairs are located far apart. The average standard error (as determined by the asymptotic variance-covariance estimates of the estimated parameters) is approximately 3/10 and 1/3 of the distance between stimuli 2 and 9 on the fourth dimension for subject 1 and subject 2, respectively. (The standard error estimate for each parameter is obtained by the square root of a

Table 4. Summary of the results obtained by fitting the euclidean model.

Dimensionality		Subject 1	Subject 2
	ln(L)	2598.5	1950.1
5	n_π	25	25
	AIC	-5146.9	-3850.2
	ln(L)	2597.6	1948.0
4	n_π	22	22
	AIC	-5151.2*	-3852.1*
	ln(L)	2591.3	1940.2
3	n_π	18	18
	AIC	-5146.6	-3844.4
	ln(L)	2548.1	1918.1
2	n_π	13	13
	AIC	-5070.2	-3810.1

*The minimum AIC solution

respective diagonal element in the Moore-Penrose inverse of the information matrix defined in (30).) From this, one can get a rough impression of how reliably parameters are estimated.

Furnas [Note 2] at Bell Laboratories suggested that the fourth dimension might

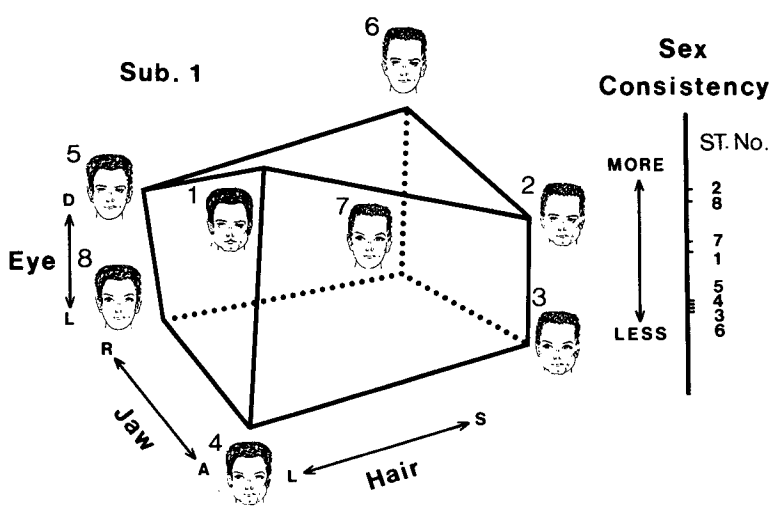

FIGURE 2.
Derived stimulus configuration for subject 1.

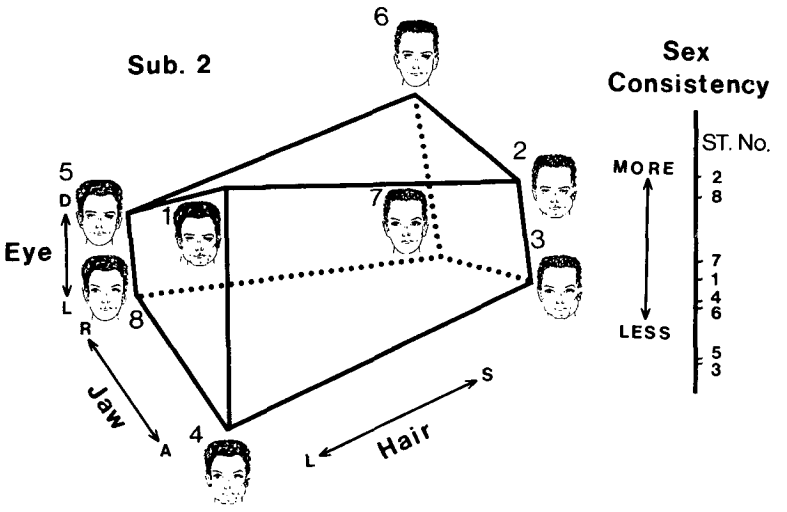

FIGURE 3.
Derived stimulus configuration for subject 2.

represent sex consistency. Short hair, angular jaw, and small dark eyes are considered typical male features, whereas long hair, round smooth jaw, and big light eyes are female features. The stimuli most consistent with these sex profiles (i.e., stimuli 2 and 8) are located at the top of the fourth dimension. For stimuli 1 and 7 the hair is inconsistent. For stimuli 3 and 5 both the hair and the jaw are inconsistent. These six stimuli are aligned in pairs roughly in the order given above from nearest to furthest from stimuli 2 and 8. (For subject 1 the last four stimuli, namely, 4, 6, 3, and 5, have approximately equal coordinates on the fourth dimension. This may be because for stimuli 3 and 5 it can also be said that only the eyes are inconsistent.)

The significant fourth dimension seems fairly persistent. It has appeared consistently for both subjects. Arrangements of the stimuli along this dimension are also very similar for both subjects. Furthermore, the significant fourth dimension was also obtained from the data taken under the upside down stimulus presentation. Again the stimulus configuration along this dimension was very similar to those obtained from the upright condition. It must be pointed out that this finding does not bear on the issue of the processes underlying face identification. Whereas it is well established that inverting the orientation of a face impairs its identification [e.g., Rock, 1973], the present experiment involved a discrimination between two simultaneously presented unfamiliar faces having no identity, and therefore required processes unlike those underlying the recognition of a face as that of a unique individual.

However, it is still possible that the fourth dimension has resulted from some misspecification of the model. We discuss two possibilities. One concerns with the function relating the distance to the reaction time, and the other pertains to the representation model.

Note that, by increasing the dimensionality from three to four, the distances remain intact for stimulus pairs differing in all three features. This implies that the three dimensional euclidean space (defined by the first three dimensions in the four dimensional solution) tends to underestimate small distances, which, in turn, implies that the fourth dimensions would not have emerged, if the function (10) relating d to T^* had yielded even smaller estimates of distances for small distances. That is, the estimated function may have been too steep. Kruskal [Note 3] suggested that this might be due to the fact that, as mentioned earlier, we set the baseline parameter c for reaction time to zero in all analyses

conducted so far. If, however, some positive value is specified for c the function would necessarily be less steep, since it is bounded from below by c.

The euclidean model was thus fitted for several prescribed values of c (c ranging from 100 to 500 ms in the step of 100 ms) under the assumption that T-c is log normal. The likelihood, however, was found to get consistently worse as the value of c was increased from zero. The estimated function (10) relating d to T^* is steepest for subject 2's last session. Even in this case the distance has to be more than four times larger than the largest existing distance to reach the point where the function predicts the mean reaction time of 300 ms, which is just about the absolute minimum for reaction times. Thus, at least for the difficulty level of the present discrimination task (and, consequently, for the range of reaction times we obtain) there is no indication that the function is too steep, though for much simpler tasks the parameter c may still be important.

Another possible cause for the emergence of the fourth dimension is inadequacy of the euclidean distance model as a combination rule. Suppose we have a right isosceles triangle, whose equal sides are oriented along the reference axes of the space in which the triangle is embedded. Suppose further that its hypothenuse is of a fixed length, say, h. Then the euclidean model predicts that the two isosceles sides have a length of $h/(2)^{1/2}$. The dominance metric (i.e., the case in which $u \rightarrow \infty$ in (72)), on the other hand, predicts that they are h, which is larger than $h/(2)^{1/2}$. Thus, the underestimation of small distances by the three-dimensional euclidean model may be corrected by the Minkowski power distance model with the power greater than two. The Minkowski power model is closer to the dominance metric in this case. The Minkowski power distance model was fitted in both three and four dimensions with u systematically varied between 2.5 and 5.0. In all cases larger likelihoods were obtained than in their euclidean counterparts. Note, however, the Minkowski power distance model ($u \neq 2$) generally uses more parameters than the euclidean model, since the former does not allow rotations of axes. This difference being taken into account in the AIC statistic, none of the Minkowski solutions ($u > 2$) were better than the euclidean four-dimensional solutions.

Shepard [1974] notes, as an observation first made by Arabie (see also Koopman & Cooper, Note 4), that if one finds a local maximum of a likelihood function (as a function of u) at $u = u_1$, one tends to find another local maximum (called a conjugate maximum) around u_2 where u_1 and u_2 satisfy $1/u_1 + 1/u_2 = 1$. For $u_1 > 2$, u_2 must be smaller than 2. This and the fact that the stimulus features (at least those physically manipulated) are fairly distinct (separable) tempted us to fit the Minkowski power distance model with $u < 2$ [Shepard, 1964; Hyman & Wells, 1967, 1968] including the city block distance ($u = 1$) [Attneave, 1950]. Shepard's [1974] remark was indeed true. However, the AIC's corresponding to the conjugate maximum of the likelihood function were found still larger than those for the four-dimensional euclidean solutions. Thus, even though the physically manipulated features are separable, the finding of a fourth dimension may be consistent with Garner's [1978] suggestion that integral (configural) properties can be derived from, and exist in addition to, the component features of a visual stimulus.

The Null Model

It was already mentioned that $\bar{g}_{ijk}^{(s)}$ in (26) and $\bar{g}_{ijk}^{(d)}$ in (27) are the minimum sufficient statistics for $h_{ijk}^{(s)}$ in (20) and $h_{ijk}^{(d)}$ in (21), respectively. Similarly, $\hat{Q}_{ijk} = n_{ijk}^{(s)}/n_{ijk}$ (where $n_{ijk} = n_{ijk}^{(s)} + n_{ijk}^{(d)}$) is minimum sufficient for Q_{ijk}. This implies that all the information pertinent to the estimation of $h_{ijk}^{(s)}$, $h_{ijk}^{(d)}$ and Q_{ijk} is contained in $\bar{g}_{ijk}^{(s)}$, $\bar{g}_{ijk}^{(d)}$ and $\hat{Q}_{ijk}$, respectively. In a sense the set, $\{\bar{g}_{ijk}^{(s)}, \bar{g}_{ijk}^{(d)}, \hat{Q}_{ijk}\}$, provides a model of reaction times and same-different judgments, which does not assume any representation or response model. This benchmark model is called the null model here. It is the most general model in the sense that it yields

the largest possible likelihood among all conceivable models that treat replications as such. Thus, it is of some interest to compare the four dimensional euclidean solution against this null model.

The null model has yielded the log likelihoods of 2804.9 and 2153.6 for subject 1 and subject 2, respectively, as opposed to 2597.6 and 1948.0 in the four dimensional euclidean solution. However, the null model uses 363 and 378 parameters for subject 1 and subject 2, respectively, whereas the four-dimensional euclidean model uses only 46 parameters. (Note that 22 df given in Table 4 does not include parameters in the error and the response models.) The differences in the log likelihoods (207.3 and 205.6) are not as large as the differences in the effective numbers of parameters (317 and 332) in the two models. Both the asymptotic chi-square and the AIC statistics in this case would favor the more restrictive model, which in the present case is the four-dimensional euclidean model.

Discussion

After conducting a series of analyses using the method introduced earlier in the paper, we are now in the position to give overall evaluations of the method and to suggest possible future studies.

Recovery of the Original Information

The validity of our results given in the previous sections depends on many factors. Above all it depends on the ability of the proposed method to recover faithful stimulus representations. In this section we give some indication that the method really works with the amount of data we have collected in our experiment. We will assume, throughout this section, that our model is correct. (Whether it is correct for our data will be discussed later.)

For those who do not trust the asymptotic standard error estimates, the jack-knife type of reliability assessment (which is conceptually similar to the split-half type of assessment) may help observe how reliably model parameters have been estimated. For each subject the data analyzed in this paper consisted of six experimental sessions. Thus, we may analyze the data from each session separately, and see how the derived configurations compare with those obtained from the entire data sets. In each case the euclidean four-dimensional model (which was the best fitting model for the entire data sets) was fitted. It was found that in all cases similar results were obtained to those obtained from the entire data sets. Mean correlations between the stimulus configurations derived from each session separately and the overall configurations were .95 (s = .03) for subject 1 and .93 (s = .4) for subject 2. While these are by no means extraordinary, considering that the mean correlation between two random configurations could be as large as .56 (s = .07) with eight stimuli in four dimensions, they are still quite high, particularly in the presence of generally large error variance in reaction time data (50 to 75% of the total variance in log reaction times in the present case) and under the influence of massive data reduction (as much as 1/6 of the entire data sets).

The above correlations, however, tend to overestimate the correlation between "true" and estimated configurations for two reasons. First, same portions of the data were repeatedly used in deriving both the overall and session-wise stimulus configurations. Secondly, the stimuli used in our empirical study are all well separated from one another, perhaps more than usual. (This is due to the factorial nature of the stimuli.) A small Monte Carlo study was thus conducted in order to investigate the goodness of recovery of the original information using randomly generated stimulus configurations. In order to maximize the comparability of the results all parameters were set as equivalent as possible to those set (or obtained) in our empirical study. It was assumed that there were eight

stimuli represented in the four-dimensional euclidean space. A different stimulus configuration was generated for each data set by uniform random numbers. The variance in reaction times was set at approximately 65% of the variance in expected log reactions times, as perhaps typical of the reaction time data comparable to ours. Ten independent samples of data were generated for each of three sample sizes: one-session equivalent (i.e., 8 replications for each of 28 *different* pairs), three-session equivalent, and six-session equivalent. Solutions were obtained in both four and five dimensions.

Mean correlations between the "true" configurations and the estimated four dimensional configurations were found to be .86 ($s = .05$) for the one-session case, .93 ($s = .02$) for the three-session case, and .97 ($s = .01$) for the six-session case. This implies that we need as much data as can be collected in six sessions (48 replications for each stimulus pair) to secure a stable .95 or above correlation with a "true" stimulus configuration. On the other hand, three sessions are sufficient, if we are satisfied with the minimum correlation of about .90. Three sessions were also found to be sufficient to rely on the AIC statistic and the asymptotic chi-square goodness of fit statistic representing the difference between four and five dimensional solutions. (The ratio of the number of observations to the number of parameters is approximately 15 in this case.) Standard errors were, on average, 2.1 and 1.5 times larger in the one-session and the three-session cases, respectively, than in the six-session case.

The above study is obviously of a rather limited nature. But at least it indicates that our empirical results are quite reliable, assuming that our model is correct.

The Representation Model

From the results presented in the previous sections it may be safely concluded that the four-dimensional euclidean model gives a reasonable approximation to our data. The results, however, may depend on the particular set of stimuli used in the present study. Quite possibly, other representation models may fit better for other kinds of stimuli. This possibility will be explored in a subsequent paper [Sergent & Takane, Note 5].

One of the major difficulties of the Minkowski power distance model as defined in (72) is that it cannot be directly fitted to reaction times for *same* pairs. Although the *same* pairs were excluded from the present analysis, their importance in choice-reaction time experiments should not be neglected [Podgorny, 1980]. Two-choice reaction time data and same-different judgments are usually recorded for both *same* and *different* pairs. Sizable portions of the data would then be discarded, if only *different* pairs are analyzed.

The importance of including the *same* pairs may be seen in Figure 4, in which estimated functions [(10) and (12)] relating the dissimilarity to the reaction time are depicted for a selected subset of experimental sessions. Functions for the "different" judgments are indicated by dotted curves, while those for the "same" judgments are indicated by solid curves. Note that the "same" function for subject 1's session 6 predicts longer reaction times than for session 1. This has been caused by the lowered threshold in session 6. (Estimated threshold values are indicated by b_1 and b_6 in the figure.) By session 6 subject 1 became completely familiar with the task and made only a few errors (i.e., "same" judgments for different pairs). The threshold should have been lowered accordingly. Evidently this was due to the fact that only different pairs were analyzed. It could be avoided only if the *same* pairs were included in the analysis.

The joint analysis of *same* and *different* pairs is possible with the linear model and Tversky's feature matching model without any modification. The Minkowski power distance model, on the other hand, needs some modification. One possibility is to define

$$\tilde{d}_{ij} = (d_{ij}^u + \theta_{i0} + \theta_{j0})^{1/u}, \tag{74}$$

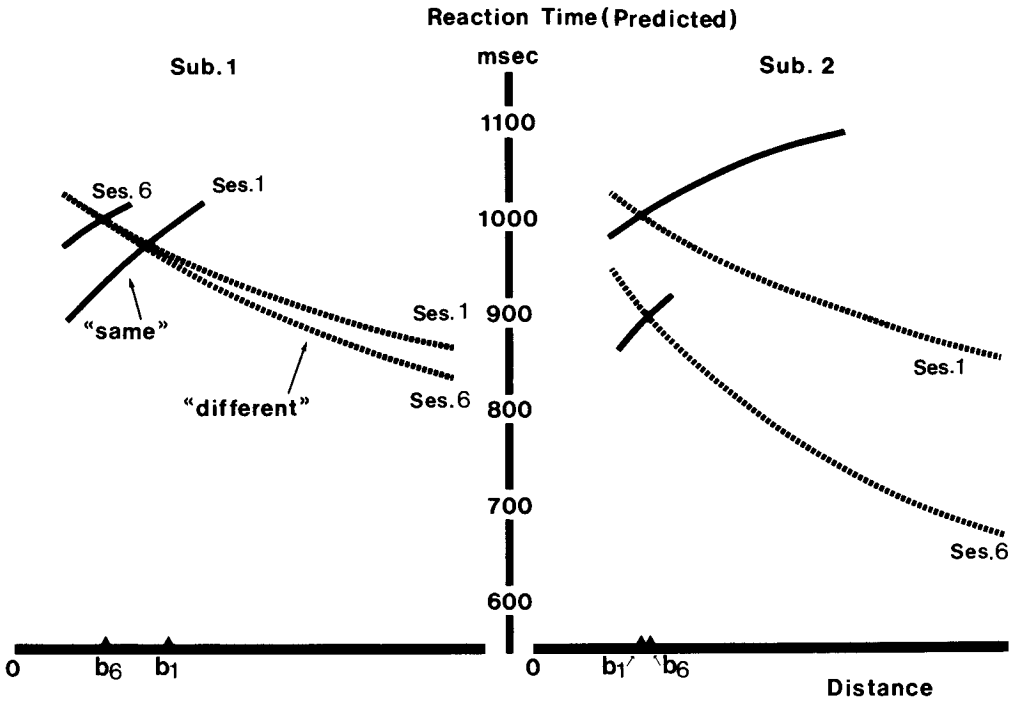

FIGURE 4.
Estimated functions relating dissimilarities to reaction times.

where d_{ij} is the Minkowski power distance as defined in (72), and θ_{i0} and θ_{j0} are additive parameters representing stimulus complexity of stimuli i and j, respectively. These additive parameters are motivated by the empirical observations [Nickerson, 1967] that it takes more time to identify two complex stimuli than simple stimuli. It is assumed that $\theta_{i0} > 0$, so that for $i = j$, $\tilde{d}_{ii} = (2\theta_{i0})^{1/u} > 0$, which conveniently introduces some random component into $\tilde{\lambda}_{iikr} = \tilde{d}_{ii} e_{iikr}$ (under the distributional assumptions (3) and (4), $V(\tilde{\lambda}_{iikr}) = 0$, if $\tilde{d}_{ii} = 0$). The above model is similar in form to Krumhansl's [1978] distance-density model, which is written, under the present notation, as

$$\tilde{d}_{ij} = d_{ij} + \theta_{i0} + \theta_{j0}. \tag{75}$$

(Note that (75) is a somewhat simplified version of Krumhansl's original model in that it has no provision for asymmetric dissimilarity.) The major difference between (74) and (75) is that the effect of θ_{i0} diminishes in the former as d_{ij} increases. In Krumhansl's model θ_{i0} is interpreted as representing spatial density surrounding stimulus i. The θ_{i0} in (74), on the other hand, may be interpreted as stimulus specificity, analogous to a specific variance component in common factor analysis. This interpretation can be afforded by the fact that the coordinate matrix for model (74) is obtained by appending a diagonal matrix of $\theta_{i0}^{1/u}$ to the original coordinate matrix for (72).

Of course there are other possibilities which are equally plausible. For example, the noncentral chi-square distribution mentioned earlier allows for $d_{ii} = 0$. It has other desirable properties as well (e.g., positive skewness, increasing variance for increasing mean, etc.). However, considerable numerical approximations are necessary to evaluate its integral, and it is only appropriate for the (squared) euclidean model. Furthermore, $d_{ii} = 0$ implies that all stimuli are equally similar to themselves, which seems far from the truth [Tversky, 1977; Podgorny & Garner, 1979]. Thus, we believe that the redefinition of d_{ii} is

not only mathematically convenient (the log normal distribution can still be used), but also empirically pertinent [Krumhansl, 1978].

The Error and the Response Models

We made a rather specific distributional assumption to relate the representation models of dissimilarities to the observed data. Whether it is adequate or not is of crucial importance in evaluating the goodness of fit of the model. The validity of the maximum likelihood estimation method employed in the present method critically depends on the veracity of the distributional assumption.

Figure 5 presents four normal quantile plots obtained from the four-dimensional euclidean solutions. Two plots were randomly sampled from each of the total of six sessions for the two subjects. The vertical axis in each plot represents normalized log residuals (i.e., $\ln t_{ijkr} - h_{ijk}^{(s)}$ or $\ln t_{ijkr} - h_{ijk}^{(d)}$ depending on the type of judgment; $h_{ijk}^{(s)}$ and $h_{ijk}^{(d)}$ are defined in (20) and (21).), and the horizontal axis normal quantile scores. If the distributional assumption is correct, the plots should exhibit 45° lines going from lower left to upper right, as indicated by solid lines. Visual inspections of the plots indicate that the fit is reasonably good in all cases, although in almost all cases we see some irregularities (departures from the solid lines) near extreme ends of the distribution. It does not seem too obvious whether any other distribution (than the log normal distribution) can correct these irregularities. Furthermore, there seem to be small, but fairly systematic individual differences. Whereas subject 1 tends to give too extreme reaction times in both tails of the distribution (more often in the upper tail), subject 2 tends to do so in the lower tail of the distribution (i.e., subject 2 tends to give too short reaction times). This difference may reflect an important difference in the two subjects' response style.

The adequacy of the log normal distributional assumption for the same-different

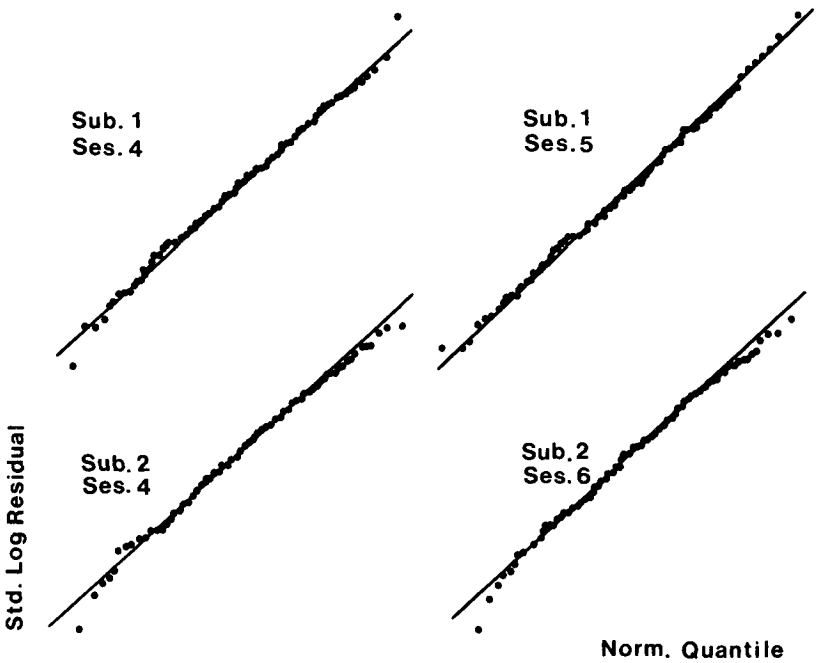

FIGURE 5.
Normal quantile plots of log residuals.

judgments may be partially checked by the likelihood ratio of predicted Q_{ijk} to observed Q_{ijk} (i.e., $\hat{Q}_{ijk}$). Minus twice the log likelihood ratio may be considered approximate chi-square with 122 df, which is equal to 28 stimulus pairs times 6 sessions (the number of $\hat{Q}_{ijk}$'s) minus 46 estimated parameters. This value was found to be 121.5 for subject 1 and 167.1 for subject 2. While it may be a bit large for subject 2, it is still much smaller than 244, twice the corresponding degrees of freedom. (This criterion is equivalent to the AIC statistic.) Furthermore, the 46 parameters were used to predict not only the probabilities of the same-different judgments but also the reaction times. The 122 df is probably an underestimation in this case. We may thus conclude that the log normal assumption is adequate for the same-different judgments as well.

In connection with the too short reaction times it may be noted that the present model almost invariably predicts relatively long reaction times for incorrect judgments. A majority of incorrect reaction times are indeed long. However, there is a fraction of incorrect reaction times which are disturbingly short. These short incorrect reaction times may be due to fast guesses [Ollman, 1966; Yellot, 1971]. If this is the case there should be approximately equal number of fast guesses which happened to be correct. At present no provision is made in our model to cope with possible fast guesses. However, guessing parameters may be incorporated into our model much the same way as in Birnbaum's three-parameter logistic model [Lord & Novick, 1968] in mental test theory. In this case we distinguish two kinds of responses, stimulus-controlled response and guessing response. The model developed in this paper is presumed to apply to the stimulus-controlled responses. The subject, however, is assumed to elicit a certain proportion of guessing responses. If the probability and other distributional properties of the guessing responses are parametrized, the distribution of the observed reaction times and same-different judgments can be specified by a mixture of two distributions, one for the stimulus-controlled responses and the other for the guessing responses. It is also possible to construct a model in which the guessing rate may vary as a function of stimulus dissimilarity and the threshold. Allowing the guessing parameters in the model in this way may also account for the nonmonotonic "latency-probability" curves found in certain choice reaction situations [Link, 1971; Link & Tindall, 1971; Petrusic & Jamieson, 1978].

It will be interesting to see how well our response model can put up with various task requirements in reaction time experiments. The data analyzed in this paper were taken under a specific condition, a clear viewing, a long exposure duration, no impending response deadlines. This is so-called resource (or process) limited condition as opposed to data (or state) limited condition [Garner, 1978]. While reaction times are typically taken under the resource limited condition, accuracy measures (error rates) are taken under the data limited condition. The joint analysis of both types of data has opened up possibilities of analyzing a whole range of data collected under varying conditions between these extremes. We have seen that our method works reasonably well under the resource limited condition. However, whether it will work equally well in other situations, or how parameters in the response model will respond to the experimental condition, is yet to be seen.

Concluding Remarks

In this paper we developed and evaluated a method for joint analysis of reaction times and same-different judgments. As a model of psychological process this method attempts to explain both types of data by a single underlying process. As a data analysis tool, it is capable of fitting the Minkowski power distance model, the linear model and a somewhat restricted version of Tversky's feature matching model. However, it is relatively straightforward to extend the method to cover other representation models, such as Car-

roll and Chang's [1970] weighted distance model, Johnson's [1967] hierarchical tree model, Fishburn's [1980] lexicographic additive difference model, etc. An interesting possibility is the probabilistic sequential scanning model, which is a proper generalization of the strict (=nonprobabilistic) search model discussed in this paper. In the former a particular scanning order of features is assumed followed with a certain probability (rather than consistently). However, at present we do not know what representation model this probabilistic process model reduces to.

We made the specific assumptions both about the distributional properties of the error and about the form of the function relating the represented dissimilarity between stimuli and the observed reaction time and same-different judgment. We have seen that these assumptions are satisfied to a reasonable degree by the data examined in this paper. Their validity, however, need be carefully inspected in each specific instance to which the present method is applied. This point should be emphasized, because at the moment we have no evidence regarding how much violation the method can tolerate. Estimates of the representation model may be fairly robust as is usually the case, though the goodness of fit statistics (AIC and asymptotic chi-square) may be more susceptible to the violation. In any case some systematic study is necessary on this point.

One may still argue that the present method lacks that remarkable generality that traditional nonmetric MDS has enjoyed. It is our contention, however, that the richness of analyses (i.e., variety, detailedness and decisiveness) we enjoy with the proposed method more than compensates the loss of generality.

REFERENCE NOTES

1. Ramsay, J. O. Some models for similarity. A paper presented at the European meeting of the Psychometric Society, Groningen, The Netherlands, 1980.
2. Furnas, G. W. Personal communication.
3. Kruskal, J. B. Personal communication.
4. Koopman, R. F., & Cooper, M. Some problems with Minkowski distance models in multidimensional scaling. Paper presented at the Psychometric Society meeting, Stanford, 1974.
5. Sergent, J. & Takane, Y. Structures in two-choice reaction time data. A manuscript in preparation.

REFERENCES

Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, *19*, 716–723.

Arabie, P. & Carroll, J. D. MAPCLUS: a mathematical programming approach to fitting the ADCLUS model. *Psychometrika*, 1980, *45*, 211–235.

Attneave, F. Dimensions of similarity. *American Journal of Psychology*, 1950, *63*, 516–556.

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

Chocholle, R. Variation des temps de reaction auditifs en fonction de l'intensite a diverses frequences. *L'Annee Psychologique*, 1940, *41*, 65–124.

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

Curtis, D. W., Paulos, M. A., & Rule, S. J. Relation between disjunctive reaction time and stimulus difference. *Journal of Experimental Psychology*, 1973, *99*, 167–173.

Davies, G. M., Ellis, H. D., & Shepherd, J. W. Cue saliency in faces as assessed by the "Photofit" technique. *Perception*, 1977, *6*, 263–269.

Ellis, H. D. Recognising faces. *British Journal of Psychology*, 1975, *66*, 409–426.

Fishburn, P. C. Lexicographic additive differences. *Journal of Mathematical Psychology*, 1980, *21*, 191–218.

Garner, W. R. Aspects of a stimulus: Features, dimensions, and configurations. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization*. Hillsdale, N.J.: Erlbaum, 1978.

Green, D. M. & Swets, J. A. *Signal detection theory and psychophysics*. New York: Krieger, 1966.

Grice, G. R., Nullmeyer, R. & Spiker, V. A. Application of variable criterion theory to choice reaction time. *Perception & Psychophysics*, 1977, *22*, 431–449.

Hefner, R. A. Extensions of the law of comparative judgement to discriminable and multidimensional stimuli. Unpublished doctoral dissertation, University of Michigan, 1958.

Hyman, R. & Well, A. Judgment of similarity and spatial models. *Perception & Psychophysics*, 1967, *2*, 233–248.
Hyman, R. & Well, A. Perceptual separability and spatial models. *Perception & Psychophysics*, 1968, *3*, 161–165.
Ida, M. The application of the Weibull distribution to the analysis of the reaction time data. *Japanese Psychological Research*, 1980, *22*, 207–212.
Johnson, N. L. & Kotz, S. *Distributions in statistics: Continuous univariate distributions—1*. Boston: Houghton Mifflins, 1970.
Johnson, S. C. Hierarchical clustering scheme. *Psychometrika*, 1967, *32*, 241–254.
Keren, G. & Baggen, S. Recognition models of alphanumeric characters. *Perception & Psychophysics*, 1981, *29*, 234–246.
Krueger, L. E. A theory of perceptual matching. *Psychological Review*, 1978, *85*, 278–304.
Krumhansl, C. L. Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, 1978, *84*, 445–463.
Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 1964a, *29*, 1–29.
Kruskal, J. B. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 1964b, *29*, 115–129.
LaBerge, D. A recruitment theory of simple behavior. *Psychometrika*, 1962, *27*, 375–396.
Laming, D. *Mathematical Psychology*. London: Academic Press, 1973.
Link, S. W. Applying RT deadlines to discrimination reaction time. *Psychonomic Science*, 1971, *25*, 355–358.
Link, S. W. The relative judgment theory of two choice response time. *Journal of Mathematical Psychology*, 1975, *12*, 114–136.
Link, S. W. & Tindall, A. D. Speed and accuracy in comparative judgments of line length. *Perception & Psychophysics*, 1971, *9*, 284–288.
Lockhead, G. R. Processing dimensional stimuli: a note. *Psychological Review*, 1972, *79*, 410–419.
Lockhead, G. R. Holistic versus analytic process models: A reply. *Journal of Experimental Psychology: Human Perception and Performance*, 1979, *5*, 746–755.
Lord, F. M. & Novick, M. R. *Statistical theories of mental test scores*. Menlo Park, Calif.: Addision-Wesley, 1968.
Marley, A. A. J. Multivariate stochastic processes compatible with "aspect" models of similarity and choice. *Psychometrika*, 1981, *46*, 421–428.
McGill, W. J. & Gibbon, J. The general gamma distribution and reaction times. *Journal of Mathematical Psychology*, 1965, *2*, 1–18.
Medin, D. L. & Schaffer, M. M. Context theory of classification learning. *Psychological Review*, 1978, *85*, 207–238.
Miller, J. Multidimensional same-different judgments: evidence against independent comparisons of dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 1978, *4*, 411–422.
Monahan, J. S. & Lockhead, G. R. Identification of integral stimuli. *Journal of Experimental Psychology: General*, 1977, *196*, 94–110.
Nickerson, R. S. Same-different reaction times with multi-attribute stimulus differences. *Perceptual and Motor Skills*, 1967, *24*, 543–554.
Nickerson, R. S. Binary-classification reaction time: a review of some studies of human information-processing capabilities. *Psychonomic Monograph Supplements*, 1972, *4*, 275–318.
Ollman, R. Fast guesses in choice reaction time. *Psychonomic Science*, 1966, *6*, 155–156.
Petrusic, W. M. & Jamieson, D. G. Relation between probability of preferential choice and time to choose changes with practice. *Journal of Experimental Psychology: Human Perception and Performance*, 1978, *4*, 471–482.
Podgorny, P. Deciding that objects are the same. Unpublished doctoral dissertation, Yale University, 1980.
Podgorny, P. and Garner, W. R. Reaction time as a measure of inter- and intraobject visual similarity: letters of the alphabet. *Perception & Psychophysics*, 1979, *26*, 37–52.
Posner, M. I. *Chronometric Explorations of Mind*. Hillsdale, N.J. Erlbaum, 1978.
Ramsay, J. O. Maximum likelihood estimation in multidimensional scaling. *Psychometrika*, 1977, *42*, 241–266.
Ramsay, J. O. Confidence regions for multidimensional scaling analysis. *Psychometrika*, 1978, *43*, 145–160.
Ramsay, J. O. Joint analysis of direct ratings, pairwise preferences and dissimilarities. *Psychometrika*, 1980, *45*, 149–165.
Ramsay, J. O. Some statistical approaches to multidimensional scaling. *Journal of the Royal Statistical Society, Series A*, 1982.
Restle, F. *Psychology of judgment and choice*. New York: Wiley, 1961.
Rock, I. *Orientation and form*. New York: Academic Press, 1973.
Sergent, J. About face: Left-hemisphere involvement in processing physiognomies. *Journal of Experimental Psychology: Human Perception and Performance*, 1982, *8*, 1–14.

Sergent, J. & Bindra, D. Differential hemispheric processing of faces: methodological considerations and reinterpretation. *Psychological Bulletin*, 1981, *89*, 541–554.

Shepard, R. N. The analysis of proximities: multidimensional scaling with an unknown distance function, I & II. *Psychometrika*, 1962, *27*, 125–140 & 219–246.

Shepard, R. N. Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, 1964, *1*, 54–87.

Shepard, R. N. Representation of structure in similarity data: problems and prospects. *Psychometrika*, 1974, *39*, 373–421.

Shepard, R. N. The circumplex and related topological manifolds in the study of perception. In Shye, S. (Ed.), *Theory construction and data analysis in the social sciences*. San Francisco: Jossey-Bass, 1978.

Shepard, R. N. & Arabie, P. Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, 1979, *86*, 87–123.

Shepard, R. N., Kilpatric, D. W. & Cunningham, J. P. The internal representation of numbers. *Cognitive Psychology*, 1975, *7*, 82–138.

Sorkin, R. D. Extension of the theory of signal detectability to matching procedures in psychoacoustics. *The Journal of Acoustic Society of America*, 1962, *34*, 1745–1751.

Stone, M. Models for choice-reaction time. *Psychometrika*, 1960, *25*, 251–260.

Takane, Y. A maximum likelihood method for nonmetric multidimensional scaling: I. The case in which all empirical pairwise orderings are independent—theory and evaluations. *Japanese Psychological Research*, 1978, *20*, 7–17 and 105–114.

Takane, Y. Multidimensional successive categories scaling: a maximum likelihood method. *Psychometrika*, 1981, *46*, 9–28.

Takane, Y. Maximum likelihood additivity analysis. *Psychometrika*, 1982, *47*, 225–241.

Takane, Y. & Carroll, J. D. Nonmetric maximum likelihood multidimensional scaling from directional rankings of similarities. *Psychometrika*, 1981, *46*, 389–405.

Torgerson, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

Tversky, A. Features of similarity. *Psychological Review*, 1977, *84*, 327–352.

Winsberg, S., & Ramsay, J. O. Analysis of pairwise preference data using integrated B-splines. *Psychometrika*, 1981, *46*, 171–186.

Yellot, J. I. Correction for fast guessing and the speed-accuracy trade-off in choice reaction time. *Journal of Mathematical Psychology*, 1971, *8*, 159–199.

Young, F. W. Nonmetric scaling of line length using latencies, similarity, and same-different judgments. *Perception & Psychophysics*, 1970, *8*, 363–369.

Young, F. W., de Leeuw, J., & Takane, Y. Quantifying qualitative data. In Lantermann, E. D., and Feger, H. (Eds.) *Similarity and choice*. Vienna: Hans Huber, 1980.

Zinnes, J. L., & Wolff, R. P. Single and multidimensional same-different judgments. *Journal of Mathematical Psychology*, 1977, *16*, 30–50.

Manuscript received 5/10/82
Final version received 3/7/83