Toward a New Paradigm for the Study of Multiattribute Choice Behavior

Spatial and Discrete Modeling of Pairwise Preferences

J. Douglas Carroll — *Graduate School of Management, Rutgers University*
Geert De Soete — *University of Ghent, Ghent, Belgium*

Some recent probabilistic choice models for paired comparisons data are reviewed. First, the differences between various forms of stochastic transitivity are explained. It is then argued that in order to be realistic, a pairwise choice model should be a moderate utility model; that is, it should imply moderate stochastic transitivity, but not necessarily strong stochastic transitivity. Subsequently, three recent families of pairwise choice models are introduced that satisfy this requirement. These models are formulated mainly in an informal manner, stressing their rationale rather than their formal derivation. The models are illustrated from data about celebrities collected by Rumelhart and Greeno (1971). In the conclusion it is argued that the present models may very well form the basis for a new paradigm for studying multiattribute choice behavior.

Preferential choice data compose one of the most important classes of data on human judgment to be understood and modeled by psychologists and other behavioral scientists (e.g., sociologists, political scientists, and marketing researchers). Whether one is (a) a social psychologist or a sociologist concerned with understanding the basis for social preferences, (b) a political scientist concerned with political choice by voters (or potential voters) or—even more importantly—by political figures such as congressmen and senators, or (c) a marketer concerned with understanding and predicting consumer behavior involving choice among different brands or products, preferential choice is a pervasive and exceedingly important human behavioral function. In fact, because every human behavior can be viewed as defining a choice among some set of available alternatives, one might argue that choice is *the* most fundamental characteristic of human behavior (or of the behavior of nonhuman organisms).

Restricting ourselves to human behavior, we may define a choice experiment as one in which a human subject must choose one from a finite set of alternatives on each of a number of different occasions. Perhaps the simplest such choice experiment is the forced choice, paired comparisons experiment in which the subject chooses one of two presented alternatives on each occasion (or trial). A complete paired comparisons experiment is one in which each subject makes such choices for all pairs of some finite set of n stimuli or objects. A replicated paired comparisons experiment is one in which the subject makes choices among each pair more than one time, with an assumed independence of the choices involving the same pair from trial to trial. With many classes of stimuli (e.g., well-known political candidates) it is unrealistic to assume independent replications because memory of the choice made on the previous trial(s), coupled with a tendency toward (apparent) consistency in behavior, would tend to induce a strong positive correlation among these choices from trial to trial. Therefore, in most cases, replicated data are only possible by replicating over different individual subjects, who are often treated as if they were independent replications of a single subject.

Given a single (unreplicated) set of paired comparisons data, we can construct a rank order of preference for the subject making those judgments, if the data are completely consistent. By consistent, we mean that if a is preferred to b, and b to c, then a is always (by transitivity) preferred to c. However, inconsistencies or failures of transitivity (sometimes called *circular triads*) often occur in actual choice data. Even in this case, one can usually construct an approximate rank order that minimizes the number of choices inconsistent with that order (or some other criterion), which will typically give a fairly good account of the data. With replicated data, rather than an all-or-none choice for one element of each pair, we have an observed frequency that can be easily transformed into a relative frequency estimating the probability of choosing a over b, written as

Editor's note. Articles based on APA award addresses that appear in the *American Psychologist* are scholarly articles by distinguished contributors to the field. As such, they are given special consideration in the *American Psychologist's* editorial selection process.

This article was originally presented as a Distinguished Scientific Contributions award address at the 98th Annual Convention of the American Psychological Association in Boston in August 1990.

Author's note. Geert De Soete was supported as "Bevoegdverklaard Navoser" of the Belgian "Nationaal Fonds voor Wetenschappelijk Onderzoek."

Correspondence concerning this article should be addressed to J. Douglas Carroll, Graduate School of Management, Rutgers University, University Heights, 92 New Street, Newark, NJ 07102-1895.

 0003-066X/91/$2.00
Vol. 46, No. 4, 342-351

$$p(a, b) = \text{Prob}(a \succ b),$$

where $\succ$ means *is preferred to*. Of course, in a forced choice situation, $p(a, b) + p(b, a) = 1$, because we are excluding the option of no choice or indifference. In this case, the notion of transitivity in a deterministic sense (if $a \succ b$ and $b \succ c$, then $a \succ c$) must be replaced by a weaker and more probabilistic notion called *stochastic* transitivity. First, we need to define the concept of stochastic dominance. Because we do not necessarily assume complete consistency over replicated choices, $p(a, b)$ will generally deviate from 0 or 1, being somewhere in between. We say that stimulus a stochastically dominates b if $p(a, b) > ½$ (i.e., if a is chosen over b more often than not). The weakest form of stochastic transitivity states that if a stochastically dominates b, and b stochastically dominates c, then a stochastically dominates c. In probability notation, this can be stated as $p(a, b) > ½$ and $p(b, c) > ½$ implies $p(a, c) > ½$. Actually, weak stochastic transitivity (WST) states something a bit more general, in which $\succ$ is replaced by $\succcurlyeq$, meaning *is preferred to* or *is equally preferred as*, reflecting the fact that if a is equally preferred to b, and b to c, we would expect the same to hold for a and c (i.e., indifference is assumed transitive, as well as stochastic dominance). Later in this article, we define some stronger forms of stochastic transitivity, moderate stochastic transitivity (MST) and strong stochastic transitivity (SST), which are central to our overall thesis.

Assume for the moment that for a given individual there is a single underlying preference order of the stimuli. It is plausible in this case to assume that preference order reflects an underlying scale or dimension of preference, which, to use a more economic term, might be associated with the utility of the stimulus for that subject. The more utility a stimulus has, the greater the subject's degree or order of preference for that stimulus. When we translate this into probability of choice of a over b, it is reasonable to assume that in general, the greater the difference in the utility, the more that probability will deviate from a neutral value of ½. One plausible assumption in fact is to assume that the probability of preferring a to b is a strictly increasing function of the difference in utility of stimulus a and b, or

$$p(a, b) = F[u(a) - u(b)], \qquad (1)$$

where $u(a)$ denotes the utility of a and where F is an increasing function so that $F[0] = ½$ (i.e., no difference in utility values leads to indifference between the alternatives). Because $p(a, b) + p(b, a)$ must, as discussed, equal 1, F must satisfy $F[u(a)] = 1 - F[-u(a)]$. It turns out that a large and very important class of probabilistic or stochastic choice models leads to probabilities of pairwise choice of exactly this form. Any pairwise model that satisfies Equation 1 is a strong utility model (SUM). Perhaps the best-known and most widely used SUM is Thurstone's (1927) model of comparative judgment Case V, in which the function F is simply the standard normal ogive (i.e., the standard normal cumulative distribution function), whereas the utility defining preference scale values, $u(a), u(b), \ldots$, constitutes a latent variable that is estimated from replicated paired comparisons data (e.g., Torgerson, 1958). Another well-known example of a SUM is what is sometimes called the Bradley–Terry–Luce model (Bradley & Terry, 1952; Luce, 1959), in which F is the logistic function. Although Bradley and Terry had originally proposed this model simply because it had useful statistical properties and seemed empirically to describe human choice behavior quite well, Luce derived this form from more fundamental principles in his seminal work on individual choice behavior.

Although stochastic choice models of this form seem very intuitively appealing, they all have strong stochastic transitivity, which has counterintuitive (and empirically invalid) consequences for certain classes of stimuli that belong to strongly multidimensional domains. Strong stochastic transitivity implies that if stimulus a stochastically dominates b and b stochastically dominates c, then $p(a, c)$ must be at least as large as the largest of $p(a, b)$ and $p(b, c)$. Although this property seems appropriate for highly unidimensional stimuli, such as different amounts of money or grades of beef, differing only in a well-defined attribute of quality (in which utility is a monotonic or order-preserving function of this single underlying dimension), it is not so clearly appropriate in domains in which the stimuli are inherently multidimensional. For example, a strong utility model implies that if one is indifferent to the choice between a vacation to Hawaii and a vacation to Acapulco, and the probability of choosing Hawaii-plus-$1 over Hawaii-alone is 1.0, then the probability of choosing Hawaii-plus-$1 over Acapulco must also be equal to 1.0. Although realistically the addition of the dollar to the same vacation (say, Hawaii) presumably would always be chosen (by a "rational" decision maker) over that vacation alone, the addition of a dollar to the Hawaii vacation package would be very unlikely to increase the probability of choosing that very slightly enhanced Hawaii package (over that of a vacation to Acapulco) to any measurable degree at all.

Thus, it would seem that for many realistic multidimensional stimulus domains SST seems *too* strong. On the other hand, WST seems too weak. Weak stochastic transitivity would allow $p(a, c)$ to be only, say, 0.51, although $p(a, b) = p(b, c) = 0.99$, for example, which certainly seems counterintuitive in most realistic choice situations. An intermediate condition, moderate stochastic transitivity, would seem more appropriate. It states that if $p(a, b) \geq ½$ and $p(b, c) \geq ½$, then $p(a, c)$ must simply be at least as great as the smaller of $p(a, b)$ and $p(b, c)$. Thus, in our Hawaii, Acapulco, and Hawaii-plus-$1 example, the probability of choosing Hawaii-plus-$1 over Acapulco need only be at least as large as that of choosing Hawaii (alone) over Acapulco.

A model that exhibits MST, but not necessarily SST, is called a moderate utility model (MUM; Halff, 1976). As already suggested, the essential ingredient for a MUM is that the stimuli be multidimensional. In our vacation example, there were at least two clear dimensions—type

of vacation (Hawaii vs. Acapulco) and money (the extra dollar).

Although choices for multidimensional stimuli may satisfy SST (e.g., when the subject, as it were, combines the dimensions in some fashion to define a single utility dimension, which then determines the choice probabilities by Equation 1), the existence of more than a single dimension seems almost a *conditio* sine qua non for MST, but not SST, to hold. Thus, we assume henceforth that there are at least two, but possibly more, psychological dimensions in terms of which the stimuli can be described. Perhaps the simplest model for combining two or more dimensions to get a single utility value is to simply take a linear combination of them to arrive at a utility $u = w_1x_1 + w_2x_2 + \cdots + w_Rx_R$, where $x_1, x_2, \ldots, x_R$ are values on R stimulus dimensions, $w_1, w_2, \ldots, w_R$ are R weights, and u is the utility value resulting from taking the weighted combination of the R dimensions defined by these weights. If these weights are fixed once and for all, then we have the situation in which a single utility scale determines choice behavior on all trials or occasions. Suppose, however, that these weights are not fixed, but vary according to some multivariate distribution function from trial to trial. This is the central assumption of the first and simplest model we discuss in this article—the wandering vector (WV) model.

Wandering Vector and Wandering Ideal Point Models

To introduce the WV model, we need to define more precisely the idea of a vector in a multidimensional space. For present purposes, a vector is a directed line segment in an R-dimensional space, emanating from the origin of the coordinate system. The terminus of a vector defines its direction from the origin and its length. A vector can be uniquely defined by the coordinates of its terminus on the R dimensions, or coordinate axes, that characterize the R-dimensional space. These coordinates can be thought of as simply the perpendicular projections of the terminus of the vector on the R coordinate axes. Given the vector, these coordinates are uniquely defined; conversely, given the coordinates, the vector is uniquely defined. A vector $\mathbf{w}$, then, can be thought of as simply a collection of coordinates $(w_1, w_2, \ldots, w_R)$. A point in a space is, geometrically, simply a special kind of vector that is conceived in terms of the locus of the terminus only, rather than as a directed line segment. So a point, $\mathbf{x}$, in an R-dimensional space with a fixed origin and coordinate system can also be represented as a vector $\mathbf{x} = (x_1, x_2, \ldots, x_R)$.

An operation that is very important in the mathematics of vector spaces is the *scalar product*. Given the coordinate representation of two vectors, say $\mathbf{w} = (w_1, w_2, \ldots, w_R)'$ and $\mathbf{x} = (x_1, x_2, \ldots, x_R)'$, the scalar product of $\mathbf{w}$ and $\mathbf{x}$, often written $\mathbf{w} \cdot \mathbf{x}$, can be defined to be the number (or scalar)

$$\mathbf{w} \cdot \mathbf{x} = \sum_{r=1}^{R} w_r x_r = w_1x_1 + w_2x_2 + \ldots + w_Rx_R.$$

Thus, the scalar product of two vectors is a weighted combination of the coordinates of one vector ($\mathbf{x}$), with the coordinates of the other ($\mathbf{w}$) defining the weights. The geometric interpretation of the scalar product operation is that its value is simply the length of the projection of the point $\mathbf{x}$ onto the vector $\mathbf{w}$, multiplied by the length of vector $\mathbf{w}$. Or, to turn this around, the projection of $\mathbf{x}$ onto $\mathbf{w}$ is the scalar product of $\mathbf{x}$ and $\mathbf{w}$ divided by the length of $\mathbf{w}$. But the Pythagorean theorem tells us (at least for Euclidean spaces, which we will assume for now) that the length of a vector $\mathbf{w}$, denoted by $||\mathbf{w}||$, is simply the square root of the scalar product with itself; that is

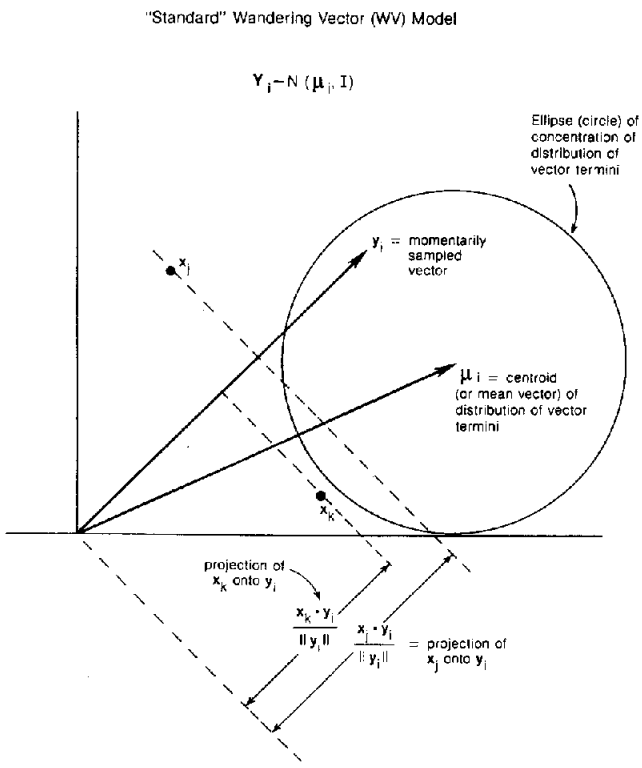
$$||\mathbf{w}|| = \sqrt{\mathbf{w} \cdot \mathbf{w}} = \sqrt{\sum_{r=1}^{R} w_r^2}.$$

So, the projection of $\mathbf{x}$ onto $\mathbf{w}$ is proportional to the scalar product of $\mathbf{x}$ and $\mathbf{w}$, with $1/||\mathbf{w}||$ being the constant of proportionality. Thus, if we assume that the utility of a stimulus, $\mathbf{u}$, is defined by a weighted linear combination of the R stimulus dimensions, it is geometrically equivalent to saying that utility is proportional to the projections of the stimulus points onto a vector ($\mathbf{w}$) in that space. If $\mathbf{w}$ is a unit length vector, so that $||\mathbf{w}|| = 1$, then the utility equals this projection.

In the WV model (Carroll, 1980; De Soete & Carroll, 1983) we assume that the vector that determines the weights of the linear combination is not fixed, but varies from trial to trial. It is, however, fixed on a particular trial. Our assumption is that, on each paired-comparisons trial, the subject projects each of the two stimuli being compared onto the momentarily fixed vector, and chooses the stimulus with the largest projection. Because we have seen the connection between projecting a point onto a vector and defining a weighted combination of dimensions, we can equally well say that the subject evaluates the same weighted combination of the values of the two stimuli on the R dimensions, and chooses the element of the pair for which that weighted combination value yields the higher value. This assumed process is illustrated geometrically in Figure 1, for the standard WV model. In this standard WV model, the distribution of vectors is assumed to be multivariate normal, with mean or centroid vector μ, and with zero covariances and equal variances for all dimensions. This simply means that the particular vector $\mathbf{y} = (y_1, y_2, \ldots, y_R)'$ sampled on a given trial is sampled from a distribution in which each component y_r ($r = 1, 2, \ldots, R$) is itself a (univariate) normally distributed random variable with a mean μ_r (where μ_r is the rth component of the centroid vector μ) and a common variance that can (without loss of generality) be assumed equal to 1. The fact that the covariances of every pair of coordinates, say y_r and y_s ($r \neq s$) are zero, means (for the multivariate normal distribution) that y_r and y_s are independent random variables.

We summarize this by saying that Y_i, where Y_i denotes the multivariate random variable whose value on a particular trial is $\mathbf{y}_i$ (the momentarily sampled vector), has a distribution that is $N(\mu_i, \mathbf{I})$, which denotes a mul-

Figure 1
Illustration of the Wandering Vector Model for a Single Subject i



tivariate normal distribution with mean or centroid vector μ_i and covariance matrix equal to the identity matrix **I**. (**I** is a square matrix with all off-diagonal entries = 0, and all diagonal entries = 1.) As we can see in Figure 1, this process will result in different momentary orderings of preference. In the example shown, we can see that stimulus j, represented by point $\mathbf{x}_j$, is preferred to stimulus k by subject i on that particular trial, as the projection of $\mathbf{x}_j$ onto the momentarily sampled vector $\mathbf{y}_i$ is greater than that of $\mathbf{x}_k$ onto $\mathbf{y}_i$ (so the corresponding momentary weighted linear combination of the $R = 2$ dimensions is greater for j than for k, in the judgment of subject i). It is easy to see that the order of projections of j and k on other vectors (sampled on other trials) could easily reverse that momentary preference—in fact, the projections on the centroid vector, μ_i, which can be viewed as the most likely vector to be sampled for subject i, reverses that order.

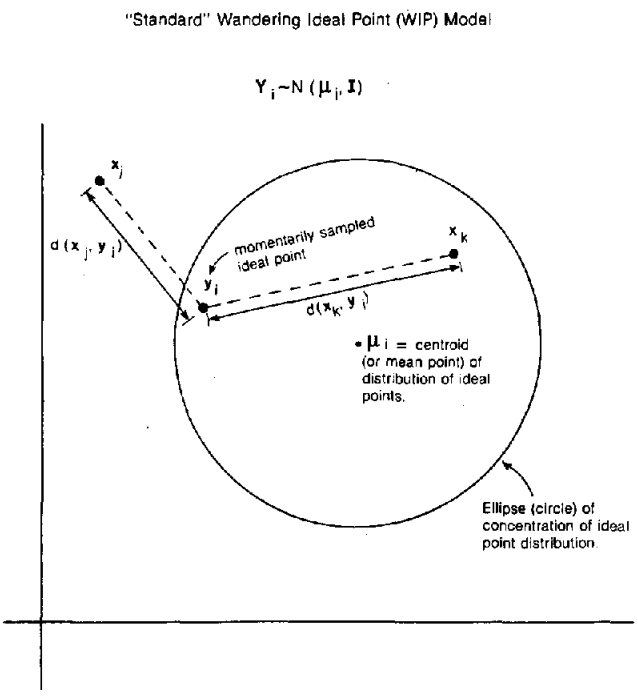
The circle shown in Figure 1 is what is called the *ellipse of concentration* for the bivariate distribution of the two dimensions of the wandering vector, $\mathbf{Y}_i$, for subject i. We can think of this as defining the region of the space in which some fixed percentage (say 95%) of the termini of the vectors will fall. For bivariate normal distributions, these regions of concentration will generally be ellipses of arbitrary orientation and degree of ellipticity. For the special case in which the covariance matrix is a scalar matrix (all variances are equal and all covariances are zero), the ellipse of concentration becomes a circle. In the standard WV model, the covariance matrix is a special scalar matrix, called the *identity matrix*, in which the value of the variances is 1. If R were greater than 2, we would be dealing with the generalized multivariate case. Here the ellipse of concentration becomes an ellipsoid, whereas in the special case of an identity covariance matrix (or some other scalar matrix), the circle becomes an R-dimensional sphere. The WV model can also account for individual differences in preference behavior by different subjects (or different subgroups of subjects) by assuming each subject (or subgroup) to have a different centroid vector μ_i. In the standard WV model, all of the subjects would be assumed to have the same (identity) covariance matrix—but in the general formulation of the WV model (De Soete & Carroll, 1986, in press), different subjects (or subgroups) could have different covariance matrices, $\mathbf{\Sigma}_i$, as well. For a single subject (or subgroup), one can always assume, without loss of generality, that the (single) covariance matrix is an identity (so the standard WV model is general enough for a single subject), but in the general case of more than one subject, this assumption imposes real constraints on the model! Returning to Figure 1, then, we may think of the circle (or sphere) of concentration in the standard WV model as defining the region in which (say) 95% of termini of the sampled vectors will fall. Sampling a vector with a terminus outside that region is possible, but unlikely.

The next model we consider is the wandering ideal point (WIP) model (De Soete, Carroll, & DeSarbo, 1986). This generalizes the Coombs (1950, 1964) and the Bennett and Hays (1960) multidimensional unfolding model, in which preference is assumed to be based on closeness to an ideal stimulus point, rather than to projections of stimuli onto a vector. The standard WIP model (in which, again, the covariance matrix $\mathbf{\Sigma}_i$ is constrained to be an identity matrix) is illustrated in Figure 2. Here, the general principle is exactly the same as in the WV model, except that the individual, i, is assumed to evaluate subjectively the distance (assumed here to be Euclidean) between $\mathbf{x}_j$ and $\mathbf{y}_i$, and compare that with the distance of $\mathbf{x}_k$ to $\mathbf{y}_i$. Stimulus j is chosen over k whenever $d(\mathbf{x}_j, \mathbf{y}_i) < d(\mathbf{x}_k, \mathbf{y}_i)$, whereas stimulus k is preferred to j if the opposite order of these distances obtains. In the paired comparisons choice trial illustrated in Figure 2, $\mathbf{x}_j$ is closer to $\mathbf{y}_i$ than is $\mathbf{x}_k$, so j is preferred to k on that trial.

In the WIP model, unlike the WV model, the assumption that $\mathbf{\Sigma}_i = \mathbf{I}$, even for the case of only one subject (or subgroup), cannot be made without loss of generality, and so puts definite constraints on the model. In the standard model, which is illustrated in Figure 2, this assumption is made, so the ellipse (or ellipsoid) of concentration (of the distribution of wandering ideal points) again becomes a circle (or sphere).

Although the WV model appears to be quite different, it can be shown that it is a special limiting case of the WIP model, which obtains when the ideal points be-

Figure 2
Illustration of the Wandering Ideal Point Model for a Single Subject i



come infinitely distant from the stimuli, in some direction. As derived in De Soete and Carroll (1983) and De Soete et al. (1986), a general expression for the probability of preferring j to k, written as p_{jk}, for the standard versions of the WV and WIP models (for one subject) can be stated as follows:

$$p_{jk} = \Phi\left(\frac{u_j^* - u_k^*}{d(\mathbf{x}_j, \mathbf{x}_k)}\right),$$

with

$$u_j^* = u_j = \boldsymbol{\mu} \cdot \mathbf{x}_j$$

for the WV model (where $u_j = \boldsymbol{\mu} \cdot \mathbf{x}_j$ is simply a mathematical shorthand for the scalar product of the centroid vector $\boldsymbol{\mu}$ and the stimulus vector $\mathbf{x}_j$, as discussed earlier), whereas

$$\begin{aligned} u_j^* &= \boldsymbol{\mu} \cdot \mathbf{x}_j - \frac{1}{2}\,||\mathbf{x}_j||^2 \\ &= u_j - \frac{1}{2}\,||\mathbf{x}_j||^2 \end{aligned} \qquad (2)$$

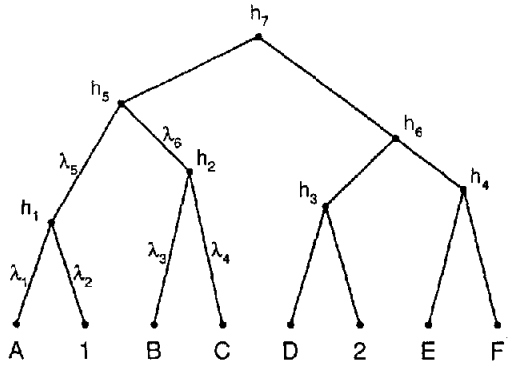
for the WIP model. The quantity $d(\mathbf{x}_j, \mathbf{x}_k)$ denotes the Euclidean distance between the points representing stimuli j and k. This general form of the choice model, in which the probability is an increasing function F (such as Φ), of a difference in utility between j and k, divided by a metric function defining dissimilarities between j and k, was shown by Halff (1976) to provide a necessary and sufficient condition for such a model to be a MUM. It turns out that the WIP model is of the same general form, but simply with the term $\frac{1}{2}||\mathbf{x}_j||^2$ subtracted from u_j. We can also see from this equation that the WV model is a special case of the WIP model. This occurs when the squared lengths of the vectors representing the stimulus points, $\mathbf{x}_j$, all become "vanishingly small" relative to the quantities u_j. This happens when the centroid $\boldsymbol{\mu}$ grows infinitely large relative to the typical (say, the maximum) length of vectors representing the stimuli (assuming the centroid of those stimulus points is constrained, without loss of generality, to be at the origin of the coordinate system). Stating this in a more precise mathematical way, as the ratio of the length of $\boldsymbol{\mu}$ to the maximum (over all stimuli j) of the length of any stimulus vector approaches infinity, u_j^* approaches u_j, so that the equation for p_{jk} in the WIP model approaches that for p_{jk} in the WV model. Thus, the WV model is a limiting case of the WIP model with the centroid of the ideal point distribution approaching infinity (in some direction). Although we have stated this only for the special case of the WIP model with only one subject, and with identity covariance matrix, the same is true when the covariance matrix is more general, and when there is more than one subject (cf. De Soete & Carroll, 1986, in press; De Soete, Carroll, & DeSarbo, 1989). (A closely related discussion of the deterministic vector model as a special case of the deterministic unfolding model can be found in Carroll, 1972, 1980. In the latter, this is demonstrated geometrically, as well as algebraically.)

Stochastic Tree Unfolding Models

The third family of models we consider are discrete tree-structure models of choice. Here, we no longer assume that the stimuli can be represented spatially by means of a small number of continuous dimensions. Rather, it is assumed that the choice objects can be characterized by means of discrete stimulus features that are organized in a particular way, namely a tree. These stochastic tree unfolding (STUN) models either assume a hierarchically organized tree—the type often associated with hierarchical clustering—or a so-called *free* or unrooted tree in which no (necessary) hierarchical organization is implied. Associated with these two types of tree structures are two types of metrics that are most naturally associated with each. Just like in the spatial unfolding model on which the WIP model is based, the subjects and the stimuli are jointly represented in a single structure in the same way. Whereas in the spatial unfolding model the subjects and the stimuli are presented by points in a Euclidean space, in the STUN models the subjects and the stimuli are represented by terminal nodes of a tree on which a metric is defined. The distance between the nodes representing a subject and a stimulus determines the utility of that stimulus for that subject.

A general tree with stimuli and subjects both represented as terminal nodes, is illustrated in Figure 3. This tree can be viewed either as a hierarchical tree or as another type of nonhierarchically organized tree commonly

Figure 3
Illustrative Two-Set Tree



Note. The terminal nodes labeled 1 and 2 represent subjects, whereas the nodes labeled A–F represent stimuli.

called a *free tree, additive tree,* or *path-length tree.* In the case of hierarchical trees, the most natural metric or distance function is the ultrametric. It can be viewed as being defined by associating heights with internal nodes of the hierarchical tree and defining the distance between two terminal nodes as the height of their least common ancestor node (i.e., the first internal or ancestral node at which the two meet as one ascends the hierarchical tree toward the root or universal ancestral node in that tree). This is illustrated in Figure 3 for a hierarchical tree with eight terminal nodes. The tree in Figure 3 is what is sometimes called a *two-set tree,* in which the terminal nodes (those at the bottom of this inverted tree—with the root at the top) are from two sets of entities. The nodes labeled 1 and 2 can be thought of as corresponding to subjects (individuals making preferential choice judgments), and the nodes labeled by letters correspond to the stimuli about which the preference judgments are made. The heights in an ultrametric tree must satisfy a partial order; that is, the height of a superordinate node must be greater or equal to that of one subordinate to it. Thus, h_5 must be $\geq h_1$ and h_2, and $h_6 \geq h_3$ and h_4, whereas h_7 must be $\geq$ any of the other six heights. Therefore, in this illustrative example, $d_{1A} = h_1$, $d_{1B} = d_{1C} = h_5 \geq h_1$, whereas $d_{1D} = d_{1E} = d_{1F} = h_7$ ($\geq h_5 \geq h_1$), so that if this were a tree unfolding representation, Subject 1 would be predicted to prefer A to any of the other stimuli, to prefer either of stimuli B or C to any of D, E, or F, but to be indifferent between the pair A and B or among the set D, E, and F. (Analogous statements could be inferred about the quite different preferences of Subject 2 in Figure 3). We can view such a hierarchical tree as representing the stimuli in terms of discrete features that happen to be hierarchically organized in a fashion consistent with such a tree; then the internal nodes correspond to these features. The terminal node at which a subject appears would represent that subject's most preferred stimulus, characterized in terms of these features. A subject's ideal stimulus may correspond to one of the actual stimuli represented in the same feature tree or may correspond to a set of features not corresponding to any actual stimulus.

One-set trees have frequently been used to represent the conceptual structure of stimuli as reflected in similarity data (cf. Carroll, 1976; Carroll & Chang, 1973; Carroll & Pruzansky, 1980; De Soete, 1983, 1984; Johnson, 1967; Sattath & Tversky, 1977). In the case of a one-set ultrametric tree, a very strong condition, called the *ultrametric inequality,* can be shown to hold for the distances. The ultrametric inequality states that for any three points A, B, and C, the following inequality holds

$$d_{AC} \leq \max(d_{AB}, d_{BC}),$$

which can easily be shown to be equivalent to the condition that every triangle must be acute isosceles (i.e., isosceles with the two largest distances equal). The ultrametric inequality is a special case of, and thus a much stronger condition than, the triangle inequality

$$d_{AC} \leq d_{AB} + d_{BC},$$

which is one of the metric axioms that hold for any metric (e.g., the Euclidean metric or distance function most often assumed in one form or other in multidimensional scaling of similarity data). As shown by Furnas (1980), given a two-set tree, for which only the between-set distances are known, a somewhat more general condition holds, called the *two-set ultrametric condition.* This can be stated in terms of the following inequality:

$$d_{1A} \leq \max(d_{1B}, d_{2A}, d_{2B}),$$

1 and 2 being any two objects from the first set (say, subjects) and A and B any two from the second set (say, stimuli). This two-set ultrametric condition was used by De Soete, DeSarbo, Furnas, and Carroll (1984a, 1984b) to implement a method of fitting an ultrametric tree unfolding model to rating scale judgments of preference, in the form of a subject by stimuli matrix of preference ratings (or other "dominance" values). In unfolding theory, such a rectangular preference score matrix can be interpreted as an off-diagonal proximity matrix, so that the entries can be viewed as linearly or monotonically related to distances between these two sets (with no data relating to distances within either set—among subjects or among stimuli).

The second form of metric associated with a tree is a path-length or additive metric, in which each branch, or link, in the tree has associated with it a length (sometimes called a *weight*), and the distance between two terminal nodes (or two internal nodes, for that matter) is the length of the unique path connecting them. In Figure 3 we have indicated lengths for the subtree of the larger tree that falls below Internal Node 5. In this case, the distance from Subject 1 to Stimulus A is $d_{1A} = \lambda_1 + \lambda_2$, whereas $d_{1B} = \lambda_2 + \lambda_5 + \lambda_6 + \lambda_3$, and $d_{1C} = \lambda_2 + \lambda_5 + \lambda_6 + \lambda_4$. There is no constraint on these lengths (except nonnegativity), so the order of these three distances could be anything whatever (d_{1B} can be greater than or less than d_{1C} depending on whether λ_3 is greater than or less than λ_4, whereas each of the two distances could be either

larger or smaller than d_{1A} depending on the values of these branch lengths). A two-set path-length (or additive) tree does impose structure on the possible preference ordering, however, when a larger number of objects (or stimuli) and subjects is involved. In fact, going back to the case of a one-set tree structure, there is a certain four-point condition that must hold on the distances for them to be consistent with a path-length tree. We shall not state that four-point condition here, but we refer the reader to Carroll (1976) for a discussion of this. It turns out that a certain six-point condition derived by Brossier (1986) obtains for a two-set path-length (or additive) tree. Thus, there must be at least six points (three from each set) for nontrivial constraints to be imposed on preference orderings by this model. Although ultrametric and path-length trees may seem to be quite different, there are some well-defined relationships between them. Every ultrametric tree can be easily converted into a path-length tree, whereas every additive tree can be decomposed into the sum of an ultrametric tree and a tree with a single internal node (often called a *star* tree by graph theorists). For a further discussion of these relationships, see Carroll (1976) and Carroll and Pruzansky (1980).

The STUN models were first introduced by Carroll, DeSarbo, and De Soete (1987), with further developments published by Carroll, DeSarbo, and De Soete (1988, 1989); Carroll and De Soete (1990); and DeSarbo, De Soete, Carroll, and Ramaswamy (1988). The name given to the most general family of STUN models is GSTUN. The GSTUN model can begin with either an ultrametric or a path-length tree as the structural component, entailing a two-set tree with n stimuli and m (>1) subjects. In GSTUN the continuous parameters—the heights of the internal nodes in the case of an ultrametric tree or the lengths of the branches in the case of a path-length tree—are assumed to vary independently from one paired-comparison trial to another, following a multivariate normal distribution. This leads to a form for the model that is remarkably similar to the WV or WIP models, namely

$$p_{ijk} = \Phi\left(\frac{d_{ik} - d_{ij}}{\delta_{ijk}}\right), \qquad (3)$$

where d_{ij} is the expected tree distance between subject i and stimulus j and δ_{ijk} a generalized Euclidean distance defined in terms of a matrix called the *path matrix*. The path matrix is an indicator matrix that defines the heights or branch lengths, as the case may be, that are included in the definition of the distance for a particular subject–stimulus pair.

More recently, Carroll and De Soete (1990) proposed a particularly interesting special case of GSTUN, assuming a path-length tree and a very special structure on the covariance matrix of the multivariate normal distribution of the branch lengths. First, we assume that the (off-diagonal) covariances are all zero, so that the covariance matrix is diagonal. We then assume that these diagonal values, the variances, are equal to the means of the corresponding variables. A simpler way to describe this model is to say that it is a path-length tree model (on n stimuli and m subjects) in which the lengths L_b of the $2(n+m)-3$ branches b are assumed to be independently univariate normally distributed with mean μ_b and variance ${\sigma_b}^2 = \mu_b$; that is,

$$L_b \sim N(\mu_b, \mu_b).$$

We initially called this model a *quasi-Poisson* case of the GSTUN model, as the distribution of each branch length simulates an independent Poisson process that has a mean equal to its variance. A Poisson process, although leading to a discrete distribution, is a not unnatural process to assume for such branch lengths. Hence, this model seems particularly appealing as a first approximation to a GSTUN model replacing normal with Poisson distributions. Because the variables that directly affect the p_{ijk}'s, and thus are of most importance in this model, are sums of these independently distributed random variables, the central limit theorem would tend to guarantee that the results of our quasi-Poisson model would not differ too much from a true Poisson model, in which the branch-length distributions follow precisely a Poisson distribution. There is, after all, a theoretical problem in assuming branch lengths, which should be positive, to be normally distributed, as this implies at least the possibility of a negative branch length occurring. But the robustness implied by the central limit theorem would seem to spare us this embarrassment, at least asymptotically! More recently, we have noted that a more general model, assuming the variances to be merely proportional, not equal, to the respective branch length means, leads to exactly the same form of the choice probabilities p_{ijk}. We have also renamed this model the Probabilistic Additive Tree Unfolding (PATU) model (see De Soete & Carroll, in press, for a further discussion).

Thus, the PATU model is a special case of GSTUN, assuming a path-length or additive tree, in which the branch lengths are independent random variables, L_b, with

$$L_b \sim N(\mu_b, \alpha\mu_b),$$

for some positive constant α. The choice probabilities predicted by the PATU model are

$$p_{ijk} = \Phi\left(\frac{d_{ik} - d_{ij}}{\sqrt{d_{jk}}}\right),$$

where d_{ij}, d_{ik}, and d_{jk} are all path-length distances between the relevant pair of entities—subject and stimulus pair for the numerator distances; stimulus pair for the distance appearing in the denominator of the expression to which the standard normal distribution function is applied. (As proved in the article by Carroll & De Soete, 1990, the generalized Euclidean distance δ_{ijk} that appears in the denominator of Equation 3 becomes the square root of the additive tree distance between stimuli j and k and is independent of i under the distributional assumptions made in this PATU model. Thus, the PATU model is a

moderate utility model leading to pairwise choice probabilities that are a function of only distances in a single two-set path-length tree, leading to the probability that subject i prefers stimulus j to stimulus k being a strictly increasing function of the difference in tree distance between the subject and the two stimuli divided by the square root of the tree distance between the stimuli.)

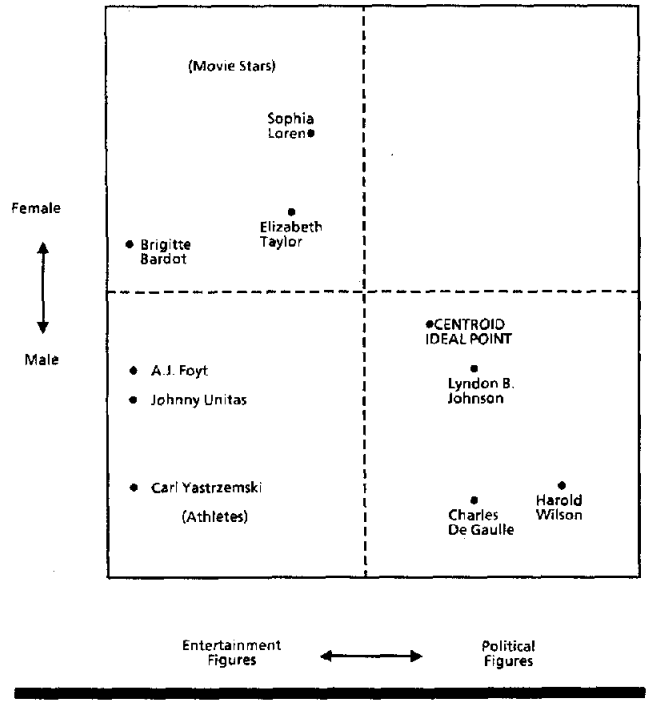
Illustrative Applications

We now illustrate the models and associated (maximum likelihood) methods of fitting them to paired comparisons data on some data collected by Rumelhart and Greeno (1971). Rumelhart and Greeno collected pairwise preference judgments for interviews—each lasting 15 minutes—with various celebrities. The nine celebrities consisted of three politicians (Charles De Gaulle, Lyndon B. Johnson, and Harold Wilson), three female movie stars (Brigitte Bardot, Sophia Loren, and Elizabeth Taylor), and three male athletes (A. J. Foyt, Johnny Unitas, and Carl Yastrzemski). The subjects, 234 undergraduates, were treated as replications of each other, so that only the aggregate-level paired comparisons data are available. Each subject was asked, for each pair of celebrities, whom he or she wanted to interview for 15 minutes. At the time the data were collected, all of the celebrities were alive and both considerably younger and more relevant to these student subjects than they would be today. The actual 9×9 paired comparisons matrix, whose general entry is the relative frequency for the corresponding pair for which the row stimulus (celebrity) was preferred to the column stimulus, is listed in the article by Rumelhart and Greeno (1971).

Spatial Analyses

Both the WV and WIP models were applied to these data. Because the fitting procedure uses a maximum likelihood procedure, it is possible to use the associated statistical test of significance to infer both the appropriate dimensionality and the comparative fit of the models. It was found that the WIP model fit the data better than did the WV model. Thus, we were able to draw the rather precise conclusion, statistically speaking, that out of this particular class of spatial multidimensional models for pairwise-choice data, the two-dimensional WIP model represented these data most accurately. Not only that, but we could conclude that the WIP model with a general covariance matrix did not fit these data statistically better than did the standard WIP model, in which the covariance matrix is constrained to be an identity matrix. Thurstone's (1927) model of comparative judgment Case V was also fit to these data, and it was found that this strong utility did not fit the data very well. The fact that a two-dimensional WIP model, rather than the Thurstone Case V model, was selected is consistent with the conclusion reached by Rumelhart and Greeno (1971) that the condition of strong stochastic transitivity was violated in these data. Not only are we able to infer a specific two-dimensional model, but we can construct, via our analysis, the two dimensions on which this model is based, as well as the position of the centroid of the distribution of ideal points. Because the standard WIP model was fit, both the orientation of the coordinate axes in this two-dimensional space and its origin are arbitrary. The solution we obtained is shown in Figure 4, with a rotation of coordinates and location of the origin chosen so as to optimize interpretability of the resulting configuration. The first axis can be interpreted as political versus entertainment figures, with the three politicians on the right and the various movie stars and sports figures on the left. (Reflection of the axes is permissible, of course, so we could have reversed the left–right polarity of these two types of celebrities had we desired.) We have placed the origin of this two-dimensional configuration so that we can further divide it into four quadrants. The lower right quadrant contains the three political figures (who all happen to be male). The lower left quadrant contains all of the athletes (who also happen to be all male), and the upper left quadrant contains the movie stars (all female). Thus, the second (vertical) dimension could be interpreted simply as gender (male vs. female) or as contrasting—for the entertainment figures, those who are movie stars with those who are sports figures. Unfortunately, because the gender attribute is completely confounded with the movie stars versus athlete distinction, while there are no female political figures included, we cannot use these data to conclude with certainty whether the second dimension is gender, actors versus athletes, or (most likely) a composite of the two. It would be interesting to replicate this study

Figure 4
Representation of the Rumelhart and Greeno (1971) Data According to the Two-Dimensional Wandering Ideal Point Model



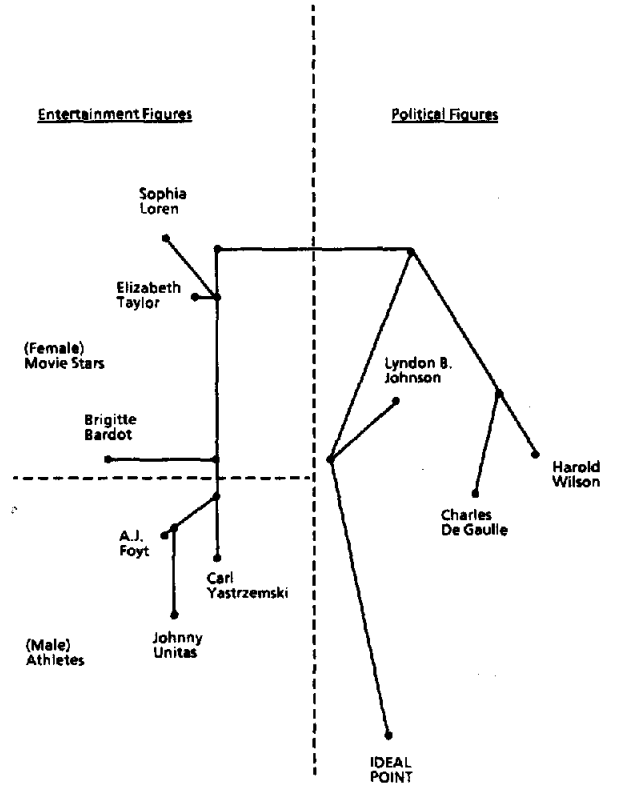
with a larger group of celebrities including male movie stars, female athletes, and such female political leaders as Margaret Thatcher, Indira Ghandi, or Golda Meir. We would speculate that, if some of the latter (none of whom, with the exception of Golda Meir, were politically prominent when these data were actually collected, however) were included, they would at least fill the now empty fourth upper right quadrant in Figure 4—or, even more likely, would serve to unconfound the gender and the actor versus athlete dimensions.

The location of the centroid of the ideal point distribution (centroid ideal point in Figure 4) indicates that the average of these 234 subjects tended to prefer to interview Lyndon B. Johnson over other figures, and generally preferred to interview politicians over either movie stars or athletes. It should be kept in mind that the two dimensions displayed in Figure 4 were inferred entirely from a single (aggregated) matrix of paired comparisons preference data (although we did have the freedom to rotate the axes and shift the origin of the space to optimize interpretability). This is remarkable in view of the fact that it was once the popular view among most psychologists that the most one could hope to infer from such a matrix of paired comparisons preferences would be a single best order or unidimensional preference scale. That we can infer information not only about the utility of the choice alternatives but also about the similarity structure of the alternatives, seems astounding when viewed from this perspective, which was prevalent as recently as, say, 25 or 30 years ago. Both our sophistication in modeling such complex data and (perhaps even more importantly) the enormous power provided by today's high-speed computers, coupled with appropriate statistical methodology and elegant numerical analysis techniques, have made what once seemed a virtual impossibility a relatively straightforward routine analytical tool.

Tree-Structure Analysis

Next, we illustrate the application of a particular version of a stochastic tree unfolding model that was discussed in the previous section, namely the PATU model, to the same preferences for interviewing these nine celebrities. The resulting additive tree structure is shown in Figure 5. This specific analysis was first reported in Carroll and De Soete (1990), although the tree was drawn in a somewhat different way. Here, we have drawn the tree structure not only to emphasize, as well as possible, how analogous conclusions can be drawn about the preference judgments of these subjects, but also to emphasize the differences necessarily implied by these rather different structural models (a tree structure vs. a two-dimensional spatial structure). We see that the longest internal branch in this unrooted tree separates the politicians from the entertainment figures—consistent with the distinction implied by the first dimension in Figure 4. Although, as drawn here, there is a branch separating the (female) movie stars from the (male) athletes, it is not, by any means, one of the larger ones in the tree. In fact, we can see that Brigitte Bardot is closer to the male athletes in this tree-structure

Figure 5
Representation of the Rumelhart and Greeno (1971) Data According to the Probabilistic Additive Tree Unfolding Model



representation than she is to either of the other female movie stars. Furthermore, the political figures (all men in this example) are all closer in the tree representation to the (female) movie stars than to the (male) athletes. This is somewhat inconsistent with the interpretation of the second dimension in Figure 4 as gender. It suggests, in fact, that perhaps the second dimension really reflects an actors versus athletes distinction and correlates with gender only because of the accidental confounding of those two attributes. In the PATU representation in Figure 5, we can see that the ideal point node is, again, closest to the node representing Lyndon B. Johnson and generally closer to the politicians than to the entertainment figures. Thus, we can conclude that the two different types of representation of these paired comparisons data lead to highly similar conclusions in many respects, although there are some obvious differences between the two representations. Hence, the two types of representations should not be regarded as competitive but rather as complementary, as we argued elsewhere for the case of proximities data (Carroll, 1976; Carroll & Pruzansky, 1980).

Conclusion

As we have seen quite clearly in the two rather different analyses of the same data set in the previous section, the

class of models and associated methods discussed in this article provide potentially powerful procedures for inferring underlying continuous dimensions or discrete features from paired-comparisons data. It is important to note that these dimensions or features (the latter in the form of a discrete tree structure) are inferred entirely from these preference data alone, without any a priori information or theoretical assumptions as to what these dimensions or features are or might be. Not only is this a powerful methodology for the study of mature human choice behavior, it also provides a potential methodology for inferring perceptual (or even conceptual) dimensions or features underlying the behavior of infants, members of primitive cultures lacking the verbal skills required to make such relatively complex similarity judgments, or even of lower organisms. For, although judgments of similarity or dissimilarity usually require some degree of verbal or at least symbolic ability, choice behavior, particularly the simple behavior exhibited in making a forced choice between two alternatives, is about the most fundamental form of behavior observable on the part of either human or nonhuman organisms, at all stages of maturation or educational attainment. Thus, models and methods of the type illustrated in this article could very well form the basis for a new paradigm in which the essential multidimensionality of the behavior of organisms at all levels of evolutionary and maturational scales can ultimately be discovered.

We believe we have only begun to see the emergence of such a new paradigm for the scientific study of the multidimensional structure of preferential choice. We eagerly await the manifold and as yet totally unanticipated shapes this new paradigm may assume within the next decades.

REFERENCES

Bennett, J. F., & Hays, W. L. (1960). Multidimensional unfolding: Determining the dimensionality of ranked preference data. *Psychometrika, 25,* 27–43.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: 1. The method of paired comparisons. *Biometrika, 39,* 324–345.

Brossier, G. (1986). Etude des matrices de proximité rectangulaires en vue de la classification [The study of rectangular proximity matrices for classification]. *Revue de Statistique Appliquées, 35,* 43–68.

Carroll, J. D. (1972). Individual differences multidimensional scaling. In R. N. Shepard, A. K. Romney, & S. B. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences* (Vol. 1, pp. 105–155). New York: Seminar Press.

Carroll, J. D. (1976). Spatial, non-spatial and hybrid models for scaling. *Psychometrika, 41,* 439–463.

Carroll, J. D. (1980). Models and methods for multidimensional analysis of preferential choice (or other dominance) data. In E. D. Lantermann & H. Feger (Eds.), *Similarity and choice* (pp. 234–289). Bern, Switzerland: Huber.

Carroll, J. D., & Chang, J.-J. (1973). A method for fitting a class of hierarchical tree structure models to dissimilarities data and its application to some "body parts" data of Miller's. *Proceedings of the 81st Annual Convention of the American Psychological Association, 8,* 1097–1098.

Carroll, J. D., DeSarbo, W. S., & De Soete, G. (1987). Stochastic tree unfolding (STUN) models. *Communication & Cognition, 20,* 63–76.

Carroll, J. D., DeSarbo, W. S., & De Soete, G. (1988). Stochastic tree unfolding (STUN) models: Theory and application. In H. H. Bock (Ed.), *Classification and related methods of data analysis* (pp. 421–430). Amsterdam: North-Holland.

Carroll, J. D., DeSarbo, W. S., & De Soete, G. (1989). Two classes of stochastic tree unfolding models. In G. De Soete, H. Feger, & K. C. Klauer (Eds.), *New developments in psychological choice modeling* (pp. 161–176). Amsterdam: North-Holland.

Carroll, J. D., & De Soete, G. (1990). Fitting a quasi-Poisson case of the GSTUN (General Stochastic Tree Unfolding) model and some extensions. In M. Schader & W. Gaul (Eds.), *Knowledge, data and computer-assisted decisions* (pp. 93–102). Berlin, Germany: Springer-Verlag.

Carroll, J. D., & Pruzansky, S. (1980). Discrete and hybrid scaling models. In E. D. Lantermann & H. Feger (Eds.), *Similarity and choice* (pp. 108–139). Bern, Switzerland: Hans Huber.

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review, 57,* 145–158.

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

DeSarbo, W. S., De Soete, G., Carroll, J. D., & Ramaswamy, V. (1988). A new stochastic ultrametric unfolding methodology for assessing competitive market structure and deriving market segments. *Applied Stochastic Models and Data Analysis, 4,* 185–204.

De Soete, G. (1983). A least squares algorithm for fitting additive trees to proximity data. *Psychometrika, 48,* 621–626.

De Soete, G. (1984). A least squares algorithm for fitting an ultrametric tree to a dissimilarity matrix. *Pattern Recognition Letters, 2,* 133–137.

De Soete, G., & Carroll, J. D. (1983). A maximum likelihood method for fitting the wandering vector model. *Psychometrika, 48,* 553–566.

De Soete, G., & Carroll, J. D. (1986). Probabilistic multidimensional choice models for representing paired comparisons data. In E. Diday, Y. Escoufier, L. Lebart, J. Pagès, Y. Schektman, & R. Tommasone (Eds.), *Data analysis and informatics IV* (pp. 485–497). Amsterdam: North-Holland.

De Soete, G., & Carroll, J. D. (in press). Probabilistic multidimensional models of pairwise choice. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition.* Hillsdale, NJ: Erlbaum.

De Soete, G., Carroll, J. D., & DeSarbo, W. S. (1986). The wandering ideal point model: A probabilistic multidimensional unfolding model for paired comparisons data. *Journal of Mathematical Psychology, 30,* 28–41.

De Soete, G., Carroll, J. D., & DeSarbo, W. S. (1989). The wandering ideal point model for analyzing paired comparisons data. In G. De Soete, H. Feger, & K. C. Klauer (Eds.), *New developments in psychological choice modeling* (pp. 123–137). Amsterdam: North-Holland.

De Soete, G., DeSarbo, W. S., Furnas, G. W., & Carroll, J. D. (1984a). The estimation of ultrametric and path length trees from rectangular proximity data. *Psychometrika, 49,* 289–310.

De Soete, G., DeSarbo, W. S., Furnas, G. W., & Carroll, J. D. (1984b). Tree representations of rectangular proximity matrices. In E. Degreef & J. Van Buggenhaut (Eds.), *Trends in mathematical psychology* (pp. 377–392). Amsterdam: North-Holland.

Furnas, G. W. (1980). *Objects and their features: The metric representation of two class data.* Unpublished doctoral dissertation, Stanford University.

Halff, H. M. (1976). Choice theories for differentially comparable alternatives. *Journal of Mathematical Psychology, 14,* 244–246.

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

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York: Wiley.

Rumelhart, D. L., & Greeno, J. G. (1971). Similarity between stimuli: An experimental test of the Luce and Restle choice models. *Journal of Mathematical Psychology, 8,* 370–381.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika, 42,* 319–345.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273–286.

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