

## The Analysis of Free-Sorting Data: Beyond Pairwise Cooccurrences

John T. Daws

New York University

**Abstract:** Free-sorting data are obtained when subjects are given a set of objects and are asked to divide them into subsets. Such data are usually reduced by counting, for each pair of objects, how many subjects placed both of them into the same subset. The present study examines the utility of a group of additional statistics, the cooccurrences of sets of three objects. Because there are dependencies among the pair and triple cooccurrences, adjusted triple similarity statistics are developed. Multidimensional scaling and cluster analysis — which usually use pair similarities as their input data — can be modified to operate on three-way similarities to create representations of the set of objects. Such methods are applied to a set of empirical sorting data: Rosenberg and Kim's (1975) fifteen kinship terms.

**Keywords:** Multidimensional scaling; Cluster analysis; Three-way proximity.

### 1. The Method of Free Sorting

A commonly used data-collection technique is the method of free sorting. In the most general case, a group of  $n$  subjects are presented with a set of  $N$  objects and are asked to place "similar" objects together and to separate "dissimilar" ones. The collection of exhaustive and disjoint subsets produced by each subject will be called a *sorting* or a *partition*.

---

The author thanks Phipps Arabie, Lawrence Hubert, Lawrence Jones, Ed Shoben, and Stanley Wasserman for their considerable contributions to this paper.

Author's Address: John T. Daws, Department of Psychology, New York University, 6 Washington Place, Room 304, New York, New York 10003, USA.

The method of free sorting has been widely used as a data collection strategy for several reasons. The task is simple and relatively easy for subjects to carry out. Researchers (e.g., Best and Ornstein 1986) have been able to use it with children as young as four. The method is cognitively engaging (Miller 1969), and subjects report enjoying the task. The method can be a quicker and less taxing way of collecting relational data than, for example, the method of pairwise comparisons, in which all distinct pairs of the objects are presented to subjects who are asked to rate the pairs on similarity. As Rosenberg (1982) noted, the advantage for sorting is greater with a larger set of objects, as the number of pairs increases with the square of the number of objects.

Free-sorting data can be aggregated in one of two ways: over subjects or over objects. Researchers who aggregate over objects may be interested either in the similarities among subjects with respect to how they partition the objects, or in the similarity of each subject's partition to some target partition of the objects. Such researchers would use partition comparison techniques (Hubert and Arabie 1985; Hubert and Levin 1976), which are not the topic of this paper. The researcher whose interest is in the relationships among the objects would aggregate the data over subjects. It is with this type of analysis that this paper is concerned.

### 1.1 Pairwise Cooccurrences

When data are aggregated over subjects, it is most common that they are reduced by counting the number of subjects who place each pair of objects into the same subset. The more frequently a pair of objects is sorted together, the more similar those two objects are taken to be. These pairwise cooccurrences provide well behaved information about the objects being sorted because their complements (the frequencies with which the pairs are **not** sorted together) satisfy the metric axioms of minimality, symmetry, and the triangle inequality (Miller 1969).

Clearly, information about the subjects' sorting responses is lost when sorting data are reduced to pairwise cooccurrences, relative to the much greater detail available in a frequency distribution of the possible partitions (Arabie and Boorman 1973). This type of loss is illustrated in the upper part of Table 1. There, two different groups of hypothetical subjects have sorted a set of four hypothetical objects in two quite different ways, yet the pairwise cooccurrences are identical for the two groups. A researcher who uses only the pairwise cooccurrences would be asserting, in effect, that there are no differences between these two groups of subjects. Any analysis based on the pairwise cooccurrences alone would produce identical results for Group 1 and Group 2, the differences between them notwithstanding.

Table 1

Results of Two Simulated Free-Sorting Experiments:  
Partition Frequencies and Summary Statistics

	Group 1		Group 2	
Frequency	ABCD	0	ABCD	0
Vectors	ABC-D	5	ABC-D	1
(F)	ABD-C	0	ABD-C	0
	ACD-B	0	ACD-B	0
	A-BCD	1	A-BCD	2
	AB-CD	0	AB-CD	1
	AC-BD	1	AC-BD	2
	AD-BC	0	AD-BC	0
	AB-C-D	1	AB-C-D	4
	AC-B-D	0	AC-B-D	3
	A-BC-D	1	A-BC-D	4
	AD-B-C	0	AD-B-C	0
	A-BD-C	2	A-BD-C	0
	A-B-CD	2	A-B-CD	0
	A-B-C-D	5	A-B-C-D	1
Pairwise	AB	6	AB	6
Cooccurrences	AC	6	AC	6
(S <sub>2</sub> )	BC	7	BC	7
	AD	0	AD	0
	BD	4	BD	4
	CD	3	CD	3
Triples	ABC	5	ABC	1
Cooccurrences	ABD	0	ABD	0
(S <sub>3</sub> )	ACD	0	ACD	0
	BCD	1	BCD	2
Adjusted	ABC	-1.33	ABC	-5.33
Triple	ABD	-3.33	ABD	-3.33
Similarities	ACD	-3.00	ACD	-3.00
(S <sub>3</sub> <sup>*</sup> )	BCD	- .67	BCD	-2.67
Pairwise	AB	-4.67	AB	-8.67
Similarities	AC	-4.33	AC	-8.33
Based on the	BC	-2.00	BC	-8.00
Adjusted Triple	AD	-6.33	AD	-6.33
Similarities	BD	-4.00	BD	-6.00
(S <sub>2</sub> <sup>*</sup> )	CD	-3.67	CD	-5.67

## 1.2 Alternatives to Pairwise Cooccurrences

Several researchers (including Burton 1975; Donderi 1988; Rosenberg and Kim 1975; and Takane 1980, 1984) have proposed other pairwise similarity measures for free-sorting data which are alternatives to the pairwise cooccurrence frequencies. Hojo's (1986) method for the analysis of sorting data does not reduce the data to cooccurrence frequencies at all, although the method is reportedly impractical with more than eight objects.

This paper describes a new approach to the analysis of sorting data which retains information about the cooccurrences of the  $\binom{N}{3}$  sets of three objects in addition to the  $\binom{N}{2}$  sets of pairs of objects. Just as a pairwise cooccurrence is used as a measure of the similarity of a pair of objects, the cooccurrence of three objects can be used as a measure of the similarity of those three objects.

## 2. Three-Way Proximity

In Tucker's (1964; Carroll and Arabie 1980) terminology, the cooccurrence frequencies of triples of objects are three-way, one-mode data (the single mode being objects). There has been some work extending the concept of distance to such data, although none of it has explicitly considered free-sorting data.

Hayashi (1972) addressed the problem of representing a set of three-way proximities ( $e_{ijk}$ ) by a set of points located in a two-dimensional Euclidean space. The configuration sought is the one that maximizes the sum of cross-products between  $e_{ijk}$  and the squared area of the triangle formed by the three points in the space representing objects  $i$ ,  $j$ , and  $k$ . The  $e_{ijk}$  are assumed to be symmetric over all six permutations of the objects, to be positive, and to increase in some way with the dissimilarity of the three objects. The problem is analogous to classical multidimensional scaling (Torgerson 1958).

Cox, Cox, and Branco (1991) extended nonmetric multidimensional scaling to the case of three-way proximities. The objects are to be represented by a configuration of points in multidimensional Euclidean space, in which the distance function for each set of three points is defined as

$$d_{ijk} = (d_{ij}^2 + d_{ik}^2 + d_{jk}^2)^{1/2},$$

where  $d_{ij}$  is the Euclidean distance between points  $i$  and  $j$  in the space. The configuration sought is the one for which the  $d_{ijk}$  have the best monotonic relationship with the proximity data  $\delta_{ijk}$ . The badness of fit between data and

configuration is measured by a straightforward generalization of Kruskal's (1964) stress statistic, in which  $d_{ijk}$  and  $\hat{d}_{ijk}$  replace  $d_{ij}$  and  $\hat{d}_{ij}$ , respectively.

Joly and Le Calvé (1995) describe three-way distances,  $t_{ijk}$ , using five axioms generalized from the familiar metric axioms for two-way distance measures. The authors discuss several alternative definitions of three-way distances: (a) the semi-perimeter (one-half of the sum of the lengths of the sides) of the triangle formed by three points in space, (b) the sum of the lengths of the paths from each of three points to some fourth point in the space (called the *star distance*), (c) the sum of the three squared Euclidean distances between the three pairs of points (i.e., the square of the Cox et al. 1991 measure), and (d) an ultrametric distance in which  $t_{ijk}$  is the height of the lowest node on a dendrogram at which objects  $i$ ,  $j$ , and  $k$  join. Joly and Le Calvé also discuss how these various distance relations may be represented in spatial configurations, dendrograms, or spanning trees.

It should be noted that the method of triads (M. W. Richardson 1938; Coombs 1964) has only a superficial connection to three-way proximity. In that method, although the objects are presented three at a time, the subjects are asked to judge the similarities of pairs of the objects. The subjects do not judge the three-way similarity of the objects.

### 3. Two- and Three-Way Similarity Measures in Free Sorting

The results of a free-sorting experiment can be summarized (with no loss of information) as a list of all possible partitions of a set of  $N$  objects, and the frequencies with which those partitions were generated by the  $n$  subjects. The number of possible partitions of a set of  $N$  objects is known as the *Bell number* (Moser and Wyman 1955) and will be denoted as  $T_N$ .

#### 3.1 The Simple Cooccurrence Frequencies

**The pairwise cooccurrences.** The cooccurrence frequency of any two objects is the sum of the frequencies for all partitions in which the two objects appear in the same subset. If we consider the partition frequencies as a  $T_N$ -element column vector,  $\mathbf{F}$ , then the  $\binom{N}{2}$ -element vector of pairwise cooccurrences,  $\mathbf{S}_2$ , can be defined as

$$\mathbf{S}_2 = \mathbf{Q}_2 \mathbf{F},$$

where  $\mathbf{Q}_2$  is a  $T_N$  by  $\binom{N}{2}$  binary indicator matrix, in which the  $(i,j)$  entry is 1 if in the  $i$ -th partition the  $j$ -th pair of objects is sorted together, and zero otherwise. The individual elements of the  $\mathbf{S}_2$  vector will be referred to as  $s_{ab}$ , for example, where the two subscripts identify the two objects.

Table 2  
Bell Number ( $T_N$ ) and Constant of Adjustment ( $c_N$ )  
for Various Numbers of Objects

$N$	$T_N$	$c_N$	$\frac{1-c_N}{c_N}$
1	1	--	--
2	2	--	--
3	5	--	--
4	15	.3333	2.00
5	52	.3000	2.33
6	203	.2703	2.70
7	877	.2450	3.08
8	4,140	.2240	3.46
9	21,147	.2066	3.84
10	115,975	.1919	4.21
11	678,570	.1793	4.58
12	4,213,597	.1686	4.93
13	27,644,437	.1591	5.28
14	190,899,322	.1509	5.63
15	1,382,958,545	.1435	5.97
20	51,724,158,235,372	.1163	7.59
25	4,638,590,332,229,999,353	.0987	9.13
30	846,749,014,511,809,332,450,147	.0861	10.60
35	$281,600,203,019,560,266,563 \times 10^9$	.0767	12.03
40	$157,450,588,391,204,931,289 \times 10^{15}$	.0694	13.40
45	$139,258,505,266,263,669,602 \times 10^{21}$	.0635	14.75
50	$185,724,268,771,078,270,438 \times 10^{27}$	.0586	16.08

**The triple cooccurrences.** Similarly, the  $\binom{N}{3}$ -element vector of triple cooccurrences is

$$\mathbf{S}_3 = \mathbf{Q}_3 \mathbf{F},$$

where  $\mathbf{Q}_3$  is a  $T_N$  by  $\binom{N}{3}$  binary indicator matrix, in which the  $(i, j)$  entry is 1 if the  $i$ -th partition contains the  $j$ -th triple, and zero otherwise. The elements of the  $\mathbf{S}_3$  vector will be referred to as  $s_{abc}$ , for example. The number of subscripts distinguishes the pair statistics from the triple statistics.

The complements of the  $\mathbf{S}_3$  statistics (the numbers of subjects who do not sort the object triples together) form a three-way distance metric, according to the Joly and Le Calvé (1995) definition.

### 3.2 The Adjusted Triple Similarity Measure

In free-sorting data, the cooccurrence of any triple of objects is related to the three pairwise cooccurrences of those three objects. The pair cooccurrences set both upper and lower bounds on the triple cooccurrence:

$$\frac{1}{2} (s_{ij} + s_{ik} + s_{jk} - n) \leq s_{ijk} \leq \min (s_{ij}, s_{ik}, s_{jk}).$$

In many cases, this lower bound will be negative, and therefore of no consequence.

To obtain a measure of triple similarity that is not predictable from the pairwise relationships, a new statistic can be defined that describes the similarity of each triple, corrected for the similarities of the three corresponding pairs. This statistic, called the *adjusted triple similarity*, is

$$s_{abc}^* = s_{abc} - c_N (s_{ab} + s_{ac} + s_{bc}),$$

where

$$c_N = \frac{T_{N-2} - T_{N-3}}{T_{N-1} - T_{N-2}}.$$

Why this statistic takes this form will be explained below. For now, note that the adjusted triple similarity measure is the simple triple frequency ( $s_{abc}$ ), minus a fraction of the sum of the three corresponding pair frequencies ( $s_{ab}$ ,  $s_{ac}$ , and  $s_{bc}$ ). As  $N$  increases,  $c_N$  decreases, as shown in Table 2.

An adjusted triple similarity statistic will be large when the three objects are frequently sorted all together, with few subjects sorting two of them together without including the third. The maximum value of  $(1 - 3c_N)n$  is reached when all  $n$  subjects sort all three objects together. An adjusted triple similarity statistic will be small when the objects are frequently sorted together in pairs but rarely or never sorted all together. The minimum value,

$-c_N n$ , is obtained when no subjects sort the three objects together, but all subjects sort exactly two of the three objects together.

Just as the simple pairwise and triple statistics can be defined as the product of a  $\mathbf{Q}$  matrix and the  $\mathbf{F}$  vector, the adjusted triple similarities can be defined as the product

$$\mathbf{S}_3^* = \mathbf{Q}_3^* \mathbf{F},$$

where  $\mathbf{Q}_3^*$  is a  $T_N$  by  $\binom{N}{3}$  matrix in which the  $(i,j)$  entry is  $(1 - 3c_N)$  if the  $i$ -th partition contains the  $j$ -th triple,  $-c_N$  if the  $i$ -th partition contains exactly one of the pairs corresponding to the  $j$ -th triple, or zero if the three objects in the  $j$ -th triple are sorted into three different subsets.

### 3.3 Sampling Distributions Under a Null Model

We will assume there are  $n$  subjects, each of whom contributes exactly one partition of the  $N$  objects. The  $T_N$ -element vector of partition frequencies,  $\mathbf{F}$ , can be defined as the sum of  $n$  independently and identically distributed random vectors, each vector containing a single one and  $T_N - 1$  zeroes. Thus  $\mathbf{F}$  has a multinomial distribution, with parameters  $n$  and  $\mathbf{P}$ , where  $\mathbf{P}$  is a  $T_N$ -element vector containing the probabilities of the various partitions. The covariance matrix of  $\mathbf{F}$  is

$$\Sigma_{\mathbf{F}} = n [ \text{diag}(\mathbf{P}) - \mathbf{P} \mathbf{P}' ],$$

where  $\text{diag}(\mathbf{P})$  is the diagonal matrix whose  $(i,i)$  element is  $p_i$ , the probability that any one of the  $n$  subjects will produce the  $i$ -th partition.

To choose the  $\mathbf{P}$  vector is to choose a null model. There are various possibilities. In the most general type of null model, all partitions having the same number of subsets and the same numbers of objects in each subset would be equally likely. For example, the partitions  $abc-de$ ,  $abe-cd$ ,  $ae-bcd$ , and all other partitions of five objects into subsets of size two and three would occur with equal probability. In such a model, the objects are mutually substitutable. Whatever is true for one object is true for all the other objects, whatever is true for one pair is true for all other pairs, and so on.

There is a special case of this null model in which all of the  $T_N$  possible partitions are equally likely. This special case will be called the *random-partitioning model*. All elements of  $\mathbf{P}$  are  $1/T_N$ . Each of the  $T_N$  elements of the  $\mathbf{F}$  vector has expected value of  $n/T_N$  (i.e., the expected frequency of any partition is the number of subjects divided by the number of partitions) and variance  $n(T_N - 1)/T_N^2$ . The covariance between any two elements is  $n/T_N^2$ .

The  $\mathbf{S}_2$ ,  $\mathbf{S}_3$ , and  $\mathbf{S}_3^*$  statistics are different linear combinations of the  $\mathbf{F}$  vector, where the linear combinations are given by the columns of the  $\mathbf{Q}_2$ ,



$\mathbf{Q}_3$ , and  $\mathbf{Q}_3^*$  matrices. Thus the expectation of each  $\mathbf{S}$  vector is the product of the appropriate  $\mathbf{Q}$  matrix and the expectation of the  $\mathbf{F}$  vector, and their covariance matrices are produced by pre- and post-multiplying the covariance matrix of  $\mathbf{F}$  by the appropriate  $\mathbf{Q}$  matrix or matrices.

**The pairwise cooccurrences.** Under the random-partitioning null model, the expected value of each of the  $\mathbf{S}_2$  statistics is  $n T_{N-1}/T_N$ . Their variances are  $n(T_N T_{N-1} - T_{N-1}^2)/T_N^2$ . The covariance between any two  $\mathbf{S}_2$  statistics is  $n(T_N T_{N-2} - T_{N-1}^2)/T_N^2$ . This covariance is the same for all pairs of  $\mathbf{S}_2$  statistics regardless of whether the two pairs have an object in common. For example,  $\text{cov}(s_{ab}, s_{ac}) = \text{cov}(s_{ab}, s_{cd})$ .

**The triple cooccurrences.** Under the random-partitioning null model, each of the  $\mathbf{S}_3$  statistics has an expected value of  $n T_{N-2}/T_N$  and a variance of  $n(T_N T_{N-2} - T_{N-2}^2)/T_N^2$ . The covariance between two  $\mathbf{S}_3$  statistics **does** depend on whether they involve a common pair of objects. If the two statistics have a common pair (as for  $s_{abc}$  and  $s_{abd}$ ), the covariance is  $n(T_N T_{N-3} - T_{N-2}^2)/T_N^2$ . If there is no pair in common (as for  $s_{abc}$  and  $s_{ade}$ , or  $s_{abc}$  and  $s_{def}$ ), the covariance is  $n(T_N T_{N-4} - T_{N-2}^2)/T_N^2$ .

The covariance between an  $\mathbf{S}_2$  statistic and an  $\mathbf{S}_3$  statistic similarly depends on whether the triple includes the pair (as, for example,  $s_{ab}$  and  $s_{abc}$ ) or not (as for  $s_{ab}$  and  $s_{acd}$ , or  $s_{ab}$  and  $s_{cde}$ ). In the former case, the covariance is  $n(T_N T_{N-2} - T_{N-1} T_{N-2})/T_N^2$ ; in the latter case, it is  $n(T_N T_{N-3} - T_{N-1} T_{N-2})/T_N^2$ . Thus, because  $T_{N-2}$  is greater than  $T_{N-3}$ ,

$$\text{cov}(s_{ab}, s_{abc}) > \text{cov}(s_{ab}, s_{acd}) = \text{cov}(s_{ab}, s_{cde}).$$

The inequality of these covariances motivated the development of the adjusted triple similarity statistics.

**The adjusted triple similarity statistics.** A triple similarity statistic was sought which would adjust the triple cooccurrences to remove pairwise similarity effects and have, under the random-partitioning null model, constant covariances with all the  $\mathbf{S}_2$  statistics. The new statistic was desired to be of the form

$$s_{abc}^* = s_{abc} + f(s_{ab}, s_{ac}, s_{bc}),$$

where  $f$  is some symmetric function on the pairwise cooccurrences, selected so that

$$\text{cov}(s_{ab}, s_{abc}^*) = \text{cov}(s_{ab}, s_{acd}^*) = \text{cov}(s_{ab}, s_{cde}^*).$$

The equality of these three covariances, along with the property of substitutability of the objects, would imply that all the entries of the covariance matrix of  $\mathbf{S}_2$  and  $\mathbf{S}_3^*$  are equal.

If we restrict the search for  $f$  to weighted sums of the three pair cooccurrences, then the problem is to find a weight,  $w$ , that implies equality for the

three types of covariance. We can solve for  $w$  algebraically and find that

$$w = \frac{T_{N-3} - T_{N-2}}{T_{N-1} - T_{N-2}} = -c_N,$$

which gives us the  $S_3^*$  statistics as defined above.

Each of the adjusted triple similarity statistics has expected value  $n(T_{N-2} - 3c_N T_{N-1})/T_N$  and variance  $\frac{n}{T_N^2} [(T_N T_{N-2} - T_{N-2}^2) - 6c_N(T_N T_{N-2} - T_{N-1} T_{N-2}) + 3c_N^2(T_N T_{N-1} + 2T_N T_{N-2} - 3T_{N-1}^2)]$ . The covariance of any  $S_3^*$  statistic with any  $S_2$  statistic is

$$\frac{n}{T_N^2} [(T_N T_{N-3} - T_{N-1} T_{N-2}) - 3c_N(T_N T_{N-2} - T_{N-1}^2)].$$

By design, this covariance does not depend on the particular objects involved but is constant over all pair-triple combinations.

### 3.4 Sampling Distributions Under a Hierarchical Model

Another way of understanding the adjusted triple similarity statistics is to see how they behave under a non-null model. The particular non-null model to be used here is a hierarchical one, in which the only partitions which have nonzero probabilities of occurring are those consistent with a certain hierarchical structure of the objects. For example, the set of partitions  $abcd$ ,  $ab-cd$ ,  $ab-c-d$ , and  $a-b-c-d$  would form a hierarchically consistent set. A psychological interpretation of this model is that the subjects share a common hierarchical arrangement of the objects. Each subject will sort the objects into one of the hierarchically consistent partitions, depending on which level of the hierarchy is most salient to him or her.

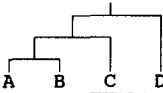
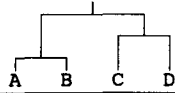
**The pairwise cooccurrences.** From Martin (1970), we know that under such conditions the pairwise cooccurrences will be ultrametric. That is, for any set of three objects  $i, j$ , and  $k$ , the two smaller pair similarities will be equal:

$$s_{ij} \geq \min(s_{ik}, s_{jk}).$$

**The triple cooccurrences.** To describe the implications of a hierarchical model on a three-way similarity measure, we must consider subsets of four objects. Furthermore, there are two possible cases which must be considered. (See Table 3.) In the first case, the objects agglomerate one at a time. In the second, the objects agglomerate into two clusters of two objects each before forming the global cluster.

Table 3

Relationships Among the Adjusted Triple Similarity Statistics Under a Hierarchical Non-Null Model

		Case 1	Case 2
			
<b>F</b>	ABCD	$f_{11}$	$f_{21}$
	ABC-D	$f_{12}$	0
	ABD-C	0	0
	ACD-B	0	0
	A-BCD	0	0
	AB-CD	0	$f_{22}$
	AC-BD	0	0
	AD-BC	0	0
	AB-C-D	$f_{13}$	$f_{23}$
	AC-B-D	0	0
	A-BC-D	0	0
	AD-B-C	0	0
	A-BD-C	0	0
	A-B-CD	0	0
	A-B-C-D	$f_{14}$	$f_{24}$
<b>S<sub>2</sub></b>	AB	$f_{11}+f_{12}+f_{13}$	$f_{21}+f_{22}+f_{23}$
	AC	$f_{11}+f_{12}$	$f_{21}$
	BC	$f_{11}+f_{12}$	$f_{21}$
	AD	$f_{11}$	$f_{21}$
	BD	$f_{11}$	$f_{21}$
	CD	$f_{11}$	$f_{21}+f_{22}$
<b>S<sub>3</sub></b>	ABC	$f_{11}+f_{12}$	$f_{21}$
	ABD	$f_{11}$	$f_{21}$
	ACD	$f_{11}$	$f_{21}$
	BCD	$f_{11}$	$f_{21}$
<b>S<sub>3</sub><sup>*</sup></b>	ABC	$f_{11}+f_{12}-c_N(3f_{11}+3f_{12}+f_{13})$	$f_{21}-c_N(3f_{21}+f_{22}+f_{23})$
	ABD	$f_{11}-c_N(3f_{11}+f_{12}+f_{13})$	$f_{21}-c_N(3f_{21}+f_{22}+f_{23})$
	ACD	$f_{11}-c_N(3f_{11}+f_{12})$	$f_{21}-c_N(3f_{21}+f_{22})$
	BCD	$f_{11}-c_N(3f_{11}+f_{12})$	$f_{21}-c_N(3f_{21}+f_{22})$
	SUMMARY	$s_{abc}^* \geq s_{abd}^*$	$s_{acd}^* = s_{bcd}^* \geq s_{abc}^* = s_{abd}^*$
		$s_{acd}^* = s_{bcd}^* \geq s_{abd}^*$	

In Case 1, the objects can be labeled so that the four possible partitions are  $abcd$ ,  $abc-d$ ,  $ab-c-d$ , and  $a-b-c-d$ . The pair similarities will have the relation

$$s_{ab} \geq s_{ac} = s_{bc} \geq s_{ad} = s_{bd} = s_{cd} ,$$

and the three smaller triple similarities (the ones involving object  $d$ , the last to join the cluster) will be equal to each other:

$$s_{abc} \geq s_{abd} = s_{acd} = s_{bcd} .$$

In Case 2, the objects can be labeled so that the four possible partitions are  $abcd$ ,  $ab-cd$ ,  $ab-c-d$ , and  $a-b-c-d$ . The pair similarities will have the relation

$$s_{ab} \geq s_{cd} \geq s_{ac} = s_{ad} = s_{bc} = s_{bd} ,$$

and all four of the triple similarities will be equal:

$$s_{abc} = s_{abd} = s_{acd} = s_{bcd} .$$

In both cases, for any set of four objects, at least three of the four triple cooccurrences will be equal. The fourth one will be greater than or equal to the other three (Joly and Le Calvé 1995). For any labeling of the objects, it will always be true that

$$s_{ijk} \geq \min (s_{ijl}, s_{ikl}, s_{jkl}) .$$

**The adjusted triple statistics.** The relations among the  $S_3^*$  statistics under the hierarchical model can be derived from the relations among the pairwise and triple cooccurrences. In Case 1, the result is that there are two possible patterns for the  $S_3^*$  statistics. Either

$$s_{abc}^* \geq s_{acd}^* = s_{bcd}^* \geq s_{abd}^* ,$$

or

$$s_{acd}^* = s_{bcd}^* > s_{abc}^* \geq s_{abd}^* .$$

The second relation (in which the largest similarity statistic does not belong to the most similar triple) will hold only when objects  $a$  and  $b$  cooccur much more often with each other than they do with object  $c$ . Precisely, it holds if and only if

$$\frac{s_{ab} - s_{ad}}{s_{ac} - s_{ad}} > \frac{1 - c_N}{c_N} .$$

The value of the ratio on the right-hand side (for which there is a column in Table 2) is always greater than or equal to two, and it increases as  $N$  increases, making it more difficult for  $s_{abc}^*$  to be smaller than  $s_{acd}^*$  and  $s_{bcd}^*$  with larger numbers of objects.

In Case 2, the  $S_3^*$  statistics always have the pattern:

$$s_{acd}^* = s_{bcd}^* \geq s_{abc}^* = s_{abd}^* .$$

The adjusted triple statistics involving both members of the more similar pair (here,  $a$  and  $b$ ) will be less than those involving both members of the less similar pair ( $c$  and  $d$ ).

## 4. Uses of the Adjusted Triple Similarities

### 4.1 Methods of Representation

To extend a pairs-based method of structural representation to handle triple proximities, two problems must be solved. One is to modify the fitting algorithm, so that it operates on triples instead of (or in addition to) pairs. The other is to define how three-way distances are to be recovered from the solution.

**Hierarchical clustering.** In a hierarchical clustering solution, the recovered three-way distance can be defined as the height of the lowest node in the hierarchy at which the three objects join. Alternatively, the recovered three-way distances could be adjusted for the corresponding recovered pairwise distances by subtracting from each one some function of the three pair distances.

The extension of a hierarchical clustering algorithm from a pair-based (e.g., S. C. Johnson 1967) to a triples-based procedure mirrors the generalization of graphs to hypergraphs (Berge 1973). Where graphs comprise a set of nodes (i.e., objects) and a set of edges (i.e., relations) between pairs of nodes, hypergraphs comprise a set of nodes and a set of relations among groups of more than two nodes. We will use what are called *uniform hypergraphs of rank three*, hypergraphs in which the relations are defined on sets of exactly three nodes. The relation in this case is the similarity of three objects. A group of  $M$  nodes for which all  $\binom{M}{3}$  possible edges are present is called a *complete sub-hypergraph*.

A complete-link hierarchical clustering algorithm for three-way similarities creates clusters which correspond to complete sub-hypergraphs. (For example, to create a cluster containing the four objects,  $a$ ,  $b$ ,  $c$ , and  $d$ , the four similarities  $s_{abc}$ ,  $s_{abd}$ ,  $s_{acd}$ , and  $s_{bcd}$  must all be large.) The first step in the clustering process is to form a cluster of the three most similar objects. Each subsequent step would either (a) form a new cluster of the next most similar three objects, (b) add a new object to an existing cluster if there is such an object with sufficiently high similarities, or (c) merge two existing clusters. The algorithm continues until there is but one cluster containing all  $N$  objects.

A single-link hierarchical clustering algorithm for three-way similarities creates clusters which correspond to connected sub-hypergraphs. There are two different ways to define connectivity for rank-three uniform hypergraphs. One could require that nodes be connected via edges which have two nodes in common, or it could be sufficient that the edges have only one node in common. Under either definition, a four-object cluster,  $abcd$ , could be created if  $s_{abc}$  and  $s_{abd}$  were large, even if  $s_{acd}$  and  $s_{bcd}$  were not. Under the one-node definition of connectivity (but not under the two-node definition), the five-object cluster  $abcde$  could be created if both  $s_{abc}$  and  $s_{ade}$  were large.

These three hierarchical clustering algorithms (complete-link, two-node single-link, and one-node single-link) will successfully recover a dendrogram from a set of triple cooccurrence frequencies (the  $S_3$ ), if those frequencies are obtained from a set of partitions all consistent with that dendrogram. For any arbitrary set of four objects, either (a) all four triple cooccurrences are equal, in which case the objects join at the same level of the dendrogram, or (b) three of the triple cooccurrences are equal and the fourth is larger, in which case the three objects with the larger cooccurrence form a cluster without the fourth object.

For the adjusted triple statistics (the  $S_3^*$ ), however, these algorithms might fail to recover the correct dendrogram from a set of error-free data, because the largest adjusted triple similarity does not necessarily belong to the three most similar objects in each set of four. It would be possible to devise an algorithm which could always recover the correct dendrogram from error-free  $S_3^*$  data, but — because it would have to rely on equalities as well as inequalities among the statistics — such an algorithm is unlikely to be at all robust when used with less-than-perfect data. Instead, we recommend using one of the three-way algorithms for the hierarchical clustering of adjusted triple similarities. The anomalous case in which the largest of a set of four  $S_3^*$  statistics does not belong to the three most similar objects should be relatively rare, especially if  $N$  is large.

With empirical data, the three algorithms could produce different dendrograms, just as S. C. Johnson's (1967) two-way single- and complete-link algorithms can. The single-link methods, especially the one-node method, would tend to agglomerate objects much more quickly than would the complete-link method. The gulf between completeness and connectivity is wider for rank-three hypergraphs than for ordinary (rank-two) graphs.

An alternative to the single- and complete-link algorithms is a modification of De Soete's (1984) least-squares algorithm for directly fitting a hierarchical tree to a set of proximity data. This algorithm searches for the dendrogram from which the recovered interobject distances account for as much of the variance in the input dissimilarities as possible.

A method of hierarchical clustering of triple similarities, different from the ones described here, has been developed by Joly and Le Calvé (1995).

**Multidimensional scaling.** The multidimensional scaling methods developed by Hayashi (1972) or by Cox et al. (1991) can be used to analyze either the triple cooccurrences or the adjusted triple similarities. Other techniques, such as R. M. Johnson's (1973) pairwise multidimensional scaling, could also be adapted for three-way proximity data.

Three distance functions which have been used or proposed for multidimensional scaling are the squared area of the triangle formed by the three points which represent the three objects in space (Hayashi 1972), the square root of the sum of the squared lengths of the sides of that triangle (Cox et al. 1991), and the semi-perimeter (Joly and Le Calvé 1995).

One would expect these three distance functions to be related to each other, because all three measure the separation of three points in space. The functions would diverge for triangles in which two of the vertices were close but the third was farther away. Such triangles would have small area, but large semi-perimeter and large sum of squared sides. Thus, the area definition of three-way distance implies that if two objects are very similar (i.e., are located very close to each other in space), then all triples containing those two objects must also be similar, all but disregarding the location of the third object in the space. For the semi-perimeter and sum of squared sides definitions, conversely, the dissimilarity of a triple must be at least as great as the dissimilarity of the two most dissimilar pair of objects; a triple containing two dissimilar objects cannot itself have a high similarity. It must be the researcher's responsibility to choose a three-way distance function that is appropriate for the data being scaled.

There is a fundamental problem in representing three-way proximity in a geometric model. In Euclidean space, the relationship among any three points is completely specified by the distances between the three pairs of points: To know the lengths of the three sides of a triangle is to know everything about the triangle except orientation. Implicit in the use of triples in psychological models, however, is the expectation that the triple relationships are — or at least can be — something over and above the three pair relationships. One would want to use the three-way similarities because one expected them to carry information which the two-way similarities do not. One might have, for example, a set of three objects for which the three-way similarity is low but the two-way similarities are high. It will be difficult if not impossible to represent those objects as points in a spatial configuration.

#### 4.2 Reduction of the Adjusted Triples Measure to Pair Statistics

It may be useful to define a pairwise similarity measure from the adjusted triple similarities. This pairwise measure would summarize the distribution of the  $S_3^*$  values for the  $N - 2$  different triples which contain the pair in question. One advantage of having a pairwise similarity measure is that it permits the immediate use of any of the standard pairwise similarity scaling techniques.

There are several ways in which such a measure could be defined. It could be defined as the maximum (or as the minimum) of the corresponding  $S_3^*$  statistics. Another possibility, the one which will be used here, is to define the pair similarity as the sum of the  $N - 2$  different  $S_3^*$  measures which involve the given pair. This measure will be called  $S_2^*$ , with its  $\binom{N}{2}$  individual elements defined as  $s_{ab}^* = \sum_k s_{abk}^*$ . If the pair has, on average, large adjusted triple similarities with the other objects, then the  $S_2^*$  measure for that pair will be large.

Like the  $S_3^*$  statistics, these  $S_2^*$  statistics have constant covariances with the pairwise cooccurrences under the null assumption that the subjects are randomly partitioning the objects. For example,

$$\text{cov}(s_{ab}, s_{ab}^*) = \text{cov}(s_{ab}, s_{ac}^*) = \text{cov}(s_{ab}, s_{cd}^*).$$

#### 4.3 Confirmatory Uses

Because the  $S_3^*$  and the  $S_2^*$  have no particular relationships with the  $S_2$  under the random-partitioning null model, they can be used in confirmatory analyses. If the subjects are partitioning the objects randomly, then whatever relationships happen to exist among the  $S_2$  tell us nothing about the  $S_3^*$  or the  $S_2^*$ . Observing, for example, that  $s_{ab}$  is one of the larger cooccurrences says nothing about the size of  $s_{abc}^*$  or  $s_{ab}^*$ . If instead, the subjects are partitioning the objects nonrandomly, we should expect correspondences between the  $S_2$  and the  $S_3^*$  or the  $S_2^*$ .

One general strategy in the confirmatory use of the adjusted triple similarities (following Hubert, Golledge, Kenny, and G. D. Richardson 1981) is first to analyze the pairwise cooccurrences, using multidimensional scaling, hierarchical clustering, or another technique; then recover some measure of three-way similarity from the results of that analysis; and finally correlate those recovered similarities with the  $S_3^*$  statistics calculated from the same data. If this correlation is not significantly greater than zero, it may be true that (a) the subjects sorted the objects randomly, (b) the analysis of the  $S_2$  is faulty (perhaps a poor local minimum was obtained), or (c) the analysis



Table 4  
 Pairwise Similarity Matrices  
 for the Rosenberg and Kim Kinship Data

$S_2$	GrF	GrM	GrS	GrD	Bro	Sis	Fat	Mot	Son	Dau	Nep	Nie	Unc	Aun
GrMother	74													
GrSon	47	37												
GrDaughter	37	47	72											
Brother	9	2	12	3										
Sister	2	9	3	12	75									
Father	12	3	11	2	30	22								
Mother	3	12	2	11	22	30	72							
Son	11	2	16	5	33	24	51	42						
Daughter	2	11	5	16	24	33	42	51	71					
Nephew	7	1	12	5	11	4	8	2	11	4				
Niece	0	6	6	13	2	9	0	6	4	11	73			
Uncle	7	1	6	0	8	2	13	6	6	0	43	36		
Aunt	1	7	0	6	2	8	6	13	0	6	36	43	75	
Cousin	1	1	4	3	8	7	1	1	3	2	32	32	46	47

$S_2^*$	GrF	GrM	GrS	GrD	Bro	Sis	Fat	Mot	Son	Dau	Nep	Nie	Unc	Aun
GrMother	4 <sup>a</sup>													
GrSon	73	52												
GrDaughter	52	73	18											
Brother	60	38	56	34										
Sister	38	60	34	55	22									
Father	60	38	51	28	94	73								
Mother	38	60	28	51	72	94	49							
Son	53	30	59	37	95	73	95	71						
Daughter	30	53	37	59	73	95	72	95	52					
Nephew	51	30	54	32	60	39	45	25	48	26				
Niece	25	45	34	56	30	51	16	37	25	46	22			
Uncle	51	30	41	21	50	30	43	24	35	14	83	59		
Aunt	30	50	21	41	29	50	23	42	14	35	64	79	34	
Cousin	39	39	39	38	45	44	30	30	31	30	78	71	59	58

<sup>a</sup> To eliminate negative values, a constant of 90 has been added to each  $S_2^*$  value.

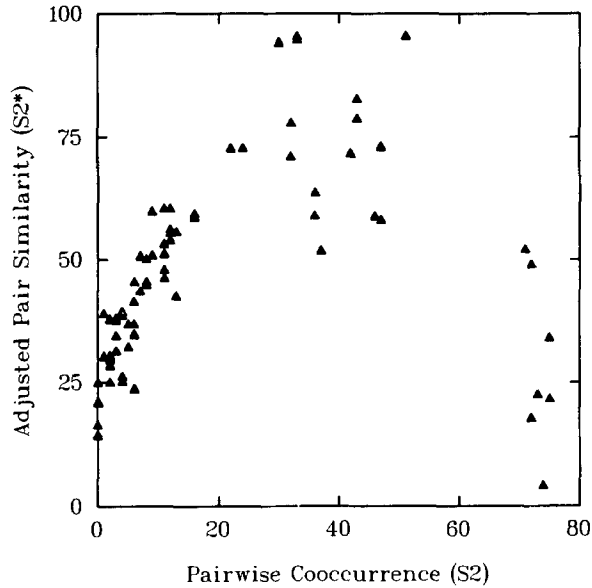


Figure 1. The two pairwise similarity measures for the Rosenberg and Kim kinship data. The pair cooccurrences ( $S_2$ ) are on the horizontal axis; the pair measure based on the adjusted triples ( $S_2^*$ ) is on the vertical. (To eliminate negative values, a constant of 90 has been added to each  $S_2^*$  value.)

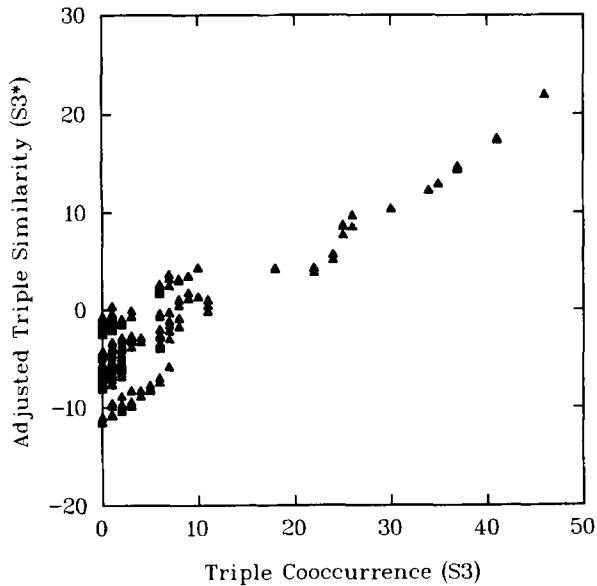


Figure 2. The two three-way similarity measures for the Rosenberg and Kim kinship data. The triple cooccurrences ( $S_3$ ) are on the horizontal axis; the adjusted triples measure ( $S_3^*$ ) is on the vertical.

represents the pairwise cooccurrences, but does not adequately represent the three-way similarities of the objects. An application of this confirmatory strategy can be found in Daws (1993).

### 5. Application to the Rosenberg and Kim Kinship Data

Rosenberg and Kim (1975) used a free-sorting task with several groups of Rutgers undergraduates. The 15 objects to be sorted were the most common kinship terms in English: *grandfather*, *grandmother*, *grandson*, *granddaughter*, *brother*, *sister*, *father*, *mother*, *son*, *daughter*, *nephew*, *niece*, *uncle*, *aunt*, and *cousin*. The subjects were told to “sort the 15 words into categories on the basis of some aspect of meaning” (p. 491). The data used here are those from the 85 female subjects in the single-sort condition which were published in Rosenberg (1982).

The pairwise cooccurrences (the  $S_2$  statistics) for the 15 objects are shown in lower-triangular form in the upper part of Table 4. The pairwise measure based on the adjusted triple similarities (the  $S_2^*$ ) is in the lower part of the table. The largest  $S_2$  values belong to the seven pairs of relatives who differ only in their sex (e.g., *grandfather* and *grandmother*). In the  $S_2^*$  matrix, however, those seven cross-sex pairs have some of the smallest values, which indicates that those relatives were very often sorted together as pairs, but were not consistently sorted together with some third object(s). The largest  $S_2^*$  statistics belong to same-sex pairs such as *grandfather-grandson* and *sister-daughter*. The differences between these two matrices are shown graphically in Figure 1, a plot of  $S_2^*$  against  $S_2$ . The seven points at the extreme right in this plot represent the cross-sex pairs. The plot also shows that the relationship between the two statistics is approximately linear for pairs of objects which were infrequently sorted together.

Figure 2 shows the analogous scatterplot for the triples statistics,  $S_3$  and  $S_3^*$ . The relationship is nearly monotonic for the triples which were frequently sorted together, but not for the infrequent ones.

#### 5.1 Multidimensional Scaling.

The two two-way similarity matrices were each analyzed with the KYST-2A multidimensional scaling program (Kruskal, Young, and Seery 1977). For the pairwise cooccurrences, a degenerate solution with zero stress was obtained. KYST-2A placed the 15 objects into three discrete groups at the vertices of an equilateral triangle: the grand-relatives, the immediate family, and the collateral relatives. (See Figure 3.)

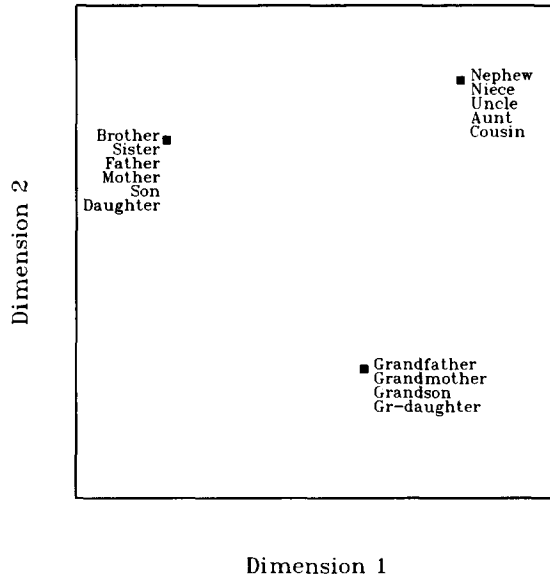


Figure 3. The spatial configuration from the multidimensional scaling of the Rosenberg and Kim pair cooccurrences ( $S_2$ ). This two-dimensional solution has zero stress.

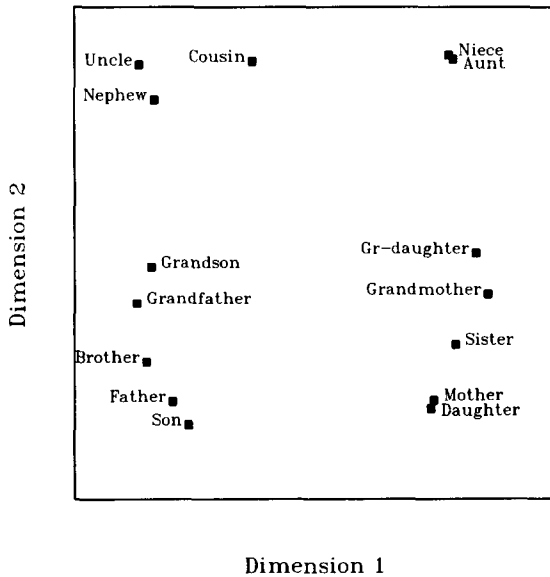


Figure 4. The spatial configuration of the kinship terms from the multidimensional scaling of the pair measure based on the adjusted triples ( $S_2'$ ). This figure is the plane formed by the first two axes in a three-dimensional solution. Stress is .098.

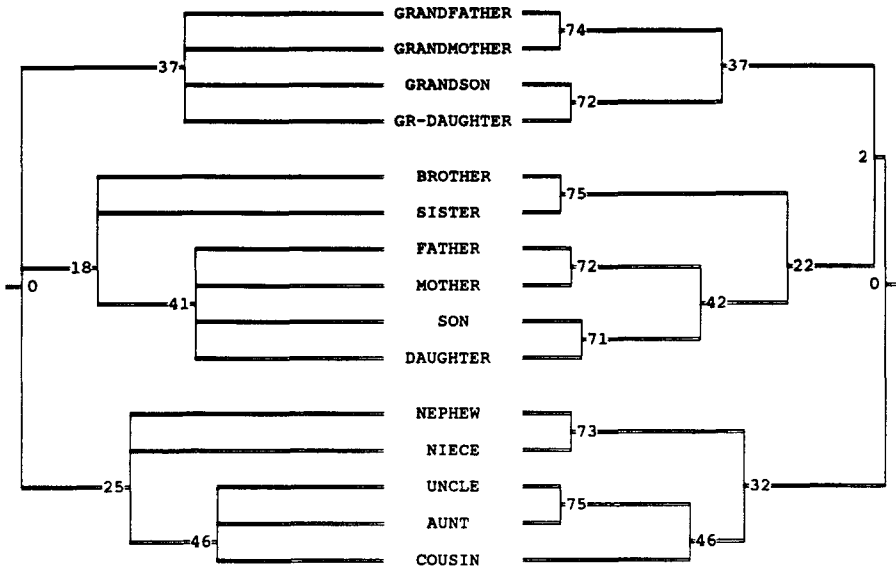


Figure 5. Results of complete-link hierarchical clustering of the kinship terms. The dendrogram on the left was obtained from the triple cooccurrences ( $S_3$ ); the one on the right, from the pair cooccurrences ( $S_2$ ). The numbers are Johnson's (1967)  $\alpha$  values, which indicate the minimum within-cluster similarity for each cluster.

For the  $S_2^*$  matrix, the multidimensional scaling was more informative. Using a quasi-metric starting configuration (KYST-2A's TORSCA option), a three-dimensional solution with a stress of .098 was found. This solution separated the objects by sex on the first dimension (with the ambiguous *cousin* located more centrally than the others) and by genetic closeness on the second dimension. The plane of these two dimensions is presented in Figure 4. The third dimension elevates *grandfather*, *grandmother*, *grandson*, and *granddaughter*, thus separating the relatives into the three groups identified in the  $S_2$  analysis. (The three groups are visible in the plane of the second and third dimensions.)

Examination of the  $S_2$  and the  $S_2^*$  matrices helps reveal why the two analyses differ so. The cross-sex pairs have the largest  $S_2$  values, but their  $S_2^*$  values are among the smallest. For the  $S_2$  analysis to produce a sex dimension, each of these seven highly similar pairs of relatives would have to be separated. In the degenerate solution, each of the seven pairs is represented by a pair of coincident points.

## 5.2 Hierarchical Clustering

Complete-link hierarchical clustering was performed on the  $S_2$  and on the  $S_3$  statistics. The resulting dendrograms are shown in Figure 5, with the triples-based one on the left and the pairs-based one on the right. The two dendrograms are quite consistent, given that the three-way algorithm cannot form two-object clusters.

## 6. Discussion

There is a tradition — and indeed a convenience — of defining similarity relationships and distance functions on pairs of objects. This paper has shown that, for free-sorting data, useful information about the similarity of the objects can be gleaned from the cooccurrences of triples of objects.

Three-way proximity data can come from sources other than free-sorting. Sociometric interaction data, *pick any* data, brand-switching data, confusion data — all can yield information about the similarities of sets of three objects. A researcher could even ask subjects to make direct judgments of how similar or dissimilar a set of three objects appear. In addition to these **direct** definitions of three-way proximity, there exist various possibilities for coefficients of association or correlation which can be used to **derive** three-way proximity from two-way, two-mode objects-by-attributes data (cf. Shepard 1972).

The mathematical and psychological properties of three-way proximity certainly merit further study.

## References

- ARABIE, P., and BOORMAN, S. A. (1973), "Multidimensional Scaling of Measures of Distance Between Partitions," *Journal of Mathematical Psychology*, 10, 148-203.
- BERGE, C. (1973), *Graphs and Hypergraphs*, Amsterdam: North Holland.
- BEST, D. L., and ORNSTEIN, P. A. (1986), "Children's Generation and Communication of Mnemonic Organizational Strategies," *Developmental Psychology*, 22, 845-853.
- BURTON, M. L. (1975), "Dissimilarity Measures for Unconstrained Sorting Data," *Multivariate Behavioral Research*, 10, 409-423.
- CARROLL, J. D., and ARABIE, P. (1980), "Multidimensional Scaling," *Annual Review of Psychology*, 31, 607-649.
- COOMBS, C. H. (1964), *A Theory of Data*, New York: Wiley.
- COX, T. F., COX, M. A. A., and BRANCO, J. A. (1991), "Multidimensional Scaling for  $n$ -Tuples," *British Journal of Mathematical and Statistical Psychology*, 44, 195-206.
- DAWS, J. T. (1993), "The Analysis of Free-Sorting Data: Beyond Pairwise Cooccurrences," Ph.D. Thesis, University of Illinois at Urbana-Champaign.
- DE SOETE, G. (1984), "A Least Squares Algorithm for Fitting an Ultrametric Tree to a Dissimilarity Matrix," *Pattern Recognition Letters*, 2, 133-137.

- DONDERI, D. C. (1988), "Information Measurement of Distinctiveness and Similarity," *Perception & Psychophysics*, 44, 576-584.
- HAYASHI, C. (1972), "Two Dimensional Quantification Based on the Measure of Dissimilarity Among Three Elements," *Annals of the Institute of Statistical Mathematics*, 24, 251-257.
- HOJO, H. (1986), "Response Threshold Models for Binary Ranking and Sorting," *Behaviormetrika*, 20, 1-12.
- HUBERT, L. J., and ARABIE, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193-218.
- HUBERT, L. J., GOLLEDGE, R. G., KENNY, T., and RICHARDSON, G. D. (1981), "Unidimensional Seriation: Implications for Evaluating Criminal Justice Data," *Environment and Planning A*, 13, 309-320.
- HUBERT, L. J., and LEVIN, J. R. (1976), "Evaluating Object Set Partitions: Free-sort Analysis and Some Generalizations," *Journal of Verbal Learning and Verbal Behavior*, 15, 459-470.
- JOHNSON, R. M. (1973), "Pairwise Nonmetric Multidimensional Scaling," *Psychometrika*, 38, 11-18.
- JOHNSON, S. C. (1967), "Hierarchical Clustering Schemes," *Psychometrika*, 32, 241-254.
- JOLY, S., and LE CALVÉ, G. (1995), *Three-Way Distances*, *Journal of Classification*, 12, 191-205.
- KRUSKAL, J. B. (1964), "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," *Psychometrika*, 29, 1-27.
- KRUSKAL, J. B., YOUNG, F. W., and SEERY, J. B. (1977), *How To Use KYST-2A, A Very Flexible Program To Do Multidimensional Scaling and Unfolding*, Murray Hill, NJ: AT&T Bell Labs.
- MARTIN, E. (1970), "Toward an Analysis of Subjective Phrase Structure," *Psychological Bulletin*, 74, 153-166.
- MILLER, G. A. (1969), "A Psychological Method to Investigate Verbal Concepts," *Journal of Mathematical Psychology*, 6, 169-191.
- MOSER, L., and WYMAN, M. (1955), "An Asymptotic Formula for the Bell Numbers," *Transactions of the Royal Society of Canada*, 49, 49-53
- RICHARDSON, M. W. (1938), "Multidimensional Psychophysics," (Summary), *Psychological Bulletin*, 35, 659-660.
- ROSENBERG, S. (1982), "The Method of Sorting in Multivariate Research with Applications Selected from Cognitive Psychology and Person Perception," in *Multivariate Applications in the Social Sciences*, Eds., N. Hirschberg and L.G. Humphreys, Hillsdale, NJ: Erlbaum, 117-142.
- ROSENBERG, S., and KIM, M. P. (1975), "The Method of Sorting as a Data-Gathering Procedure in Multivariate Research," *Multivariate Behavioral Research*, 10, 489-502.
- SHEPARD, R. N. (1972), "A Taxonomy of Some Principal Types of Data and of Multidimensional Methods for Their Analysis," in *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences: Vol. 1. Theory*, Eds., R.N. Shepard, A.K. Romney, and S. Nerlove, New York: Seminar Press, 21-47.
- TAKANE, Y. (1980), "Analysis of Categorizing Behavior by a Quantification Method," *Behaviormetrika*, 8, 75-86.
- TAKANE, Y. (1984), "Multidimensional Scaling of Sorting Data," in *Topics in Applied Statistics*, Eds., Y.P. Chaubey and T.D. Dwivedi, Montreal: Concordia University Press, 659-666.
- TORGERSON, W. S. (1958), *Theory and Methods of Scaling*, New York: Wiley.

TUCKER, L. R (1964), ‘‘The Extension of Factor Analysis to Three-Dimensional Matrices,’’ in *Contributions to Mathematical Psychology*, Eds., N. Frederiksen and H. Gulliksen, New York: Holt, Rinehart and Winston, 109-127.