Title
Dissimilarity Measures For Unconstrained Sorting Data

Journal
Multivariate Behavioral Research, 10(4)

ISSN
0027-3171

Author
Burton, Michael L

Publication Date
1975-10-01

DOI
10.1207/s15327906mbr1004_2

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: http://www.tandfonline.com/loi/hmbr20

Dissimilarity Measures For Unconstrained Sorting Data

Michael L. Burton

To cite this article: Michael L. Burton (1975) Dissimilarity Measures For Unconstrained Sorting Data, Multivariate Behavioral Research, 10:4, 409-423, DOI: 10.1207/s15327906mbr1004_2

To link to this article: http://dx.doi.org/10.1207/s15327906mbr1004_2

DISSIMILARITY MEASURES FOR UNCONSTRAINED SORTING DATA

MICHAEL L. BURTON
University of California, Irvine

ABSTRACT

In this article, three dissimilarity measures for the unconstrained sorting task are investigated. All three measures are metrics, but they differ in the kind of compensation which they make for differences in the sizes of cells within sortings. Empirical tests of the three measures are done with sorting data for occupations names and the names of behaviors, using the multidimensional scaling method.

INTRODUCTION

The unconstrained sorting task is one of several cognitive tests which can be used to obtain judgmental data about semantic organization. Studies of semantic organization using the unconstrained sorting task include the work of Rosenberg and his colleagues on implicit personality theory (Rosenberg, Nelson, and Vivekananthan, 1968), Burton's study of occupation names (1972) and Miller's work on English nouns (1969). In each of the above studies, a matrix of dissimilarity measures was computed from the sorting data, and these measures were analyzed either by multidimensional scaling (Shepard 1962, 1966; Kruskal, 1964) or by hierarchical clustering (Johnson, 1967) methods. In the unconstrained sorting task, subjects respond to verbal stimuli, which are written on cards. They sort the cards by the criterion of similarity of meaning, so that stimuli which appear to the subject to be similar are placed in the same pile. There is no restriction on the number of piles of cards or on the number of cards per pile. In the language of set theory, each subject, i, induces a partition, P_i, in the set S of stimulus elements.

Subjects may vary considerably in the kinds of partitions which they make. One useful distinction is between subjects who have large numbers of cells in their partition, sometimes referred to as "splitters", and subjects who have small numbers of cells in their partition, sometimes referred to as "lumpers." Boorman and Arabie (1972) define the concept of height of a partition on a scale from zero to one, so that a partition with a height of zero has one cell for each stimulus element, and a partition with a height of one has all stimulus elements in the same cell. The height of the partition is simply the number of pairs of elements which are placed together in cells divided by the total number of pairs of elements in the stimulus set. However, there is more to variability among subjects than simple differences in the height of the partition. Two partitions with the same height may differ with respect to the location of distinctions within the partitions. For example, the following two partitions of eight elements have the same height, but make fine distinctions within different halves of the stimulus set:

(1) (ABCD), (E), (F), (G), (H)

(2) (A), (B), (C), (D), (EFGH).

This paper discusses measures of dissimilarity among stimuli which compensate in different ways for differences in the sizes of the cells of partitions. It discusses a class of measures which are metrics and which are sums of dissimilarity measures for individual subjects. The dissimilarity measures for individual subjects vary in the kinds of adjustments which they make for the size of cell into which the individual places any two elements. An empirical study is done of three measures of dissimilarity, one of which computes increments to dissimilarity which are inversely related to cell size, the second of which computes increments to dissimilarity which are invariant under cell size, and the third of which computes increments to dissimilarity which have a positive relationship to cell size. The measures are inter-correlated, and comparisons are made among them in terms of the results of multidimensional scaling analysis.

Compensation for Differences in Cell Sizes

All of the measures proposed in this study are additive across subjects. The distance between x and y is the sum of distances between x and y for all subjects. For every subject, i, x and y are either in some cell, c_{ij} or are in different cells. If they are in c_{ij}, the size of c_{ij} may vary between 2 and N. If the cell size is small, the subject has made a relatively fine distinction between the members of c_{ij} and all other members of S. It seems reasonable to argue in this case that the average similarity among members of c_{ij} is relatively large, although some individual similarities may be small. If the cell size is large, the subject has made a relatively gross distinction between the members of c_{ij} and all other members of S. It seems reasonable to argue in this case that the average similarity among members of c_{ij} is relatively small, although some individual similarities may be large. This reasoning is best illustrated with a hierarchical model for organization of stimulus elements, although this model is only one of many which may be applicable to semantic organization. With a hierarchical model, a partition of the set is obtained by following the tree structure from its root to some node of each path. All elements below the node are placed in the same cell. This process can traverse different proportions of each path in the tree. Small cells are obtained by following the tree closer to the end of the branch. Large cells are obtained by cutting the tree closer to the beginning. A large number of partitions are consistent with a given tree structure, some of which have only large cells, some of which have only small cells, and some of which have a mixture of large and small cells. In any case, small cells contain only elements which are highly similar to each other, while large cells contain elements some of which have high similarity and some of which have relatively low similarity.

From this reasoning, one can conclude that a most accurate measure of similarity would compute a larger increment to similarity when two elements are in a small cell than when they are in a large cell, because the only possible estimate of the similarity of two elements is an estimate of the average similarity of all pairs of elements in the cell in which they are included. By the reasoning above, large cells have lower average similarity than small cells. It follows that an accurate distance measure should compute a small decrement to distance for small cells and a larger decrement to distance for larger cells.

It is also possible to argue that adjustments should be made in the case where x and y are in different cells. Here, the relevant fact is the total proportion of pairs of elements which are in different cells. If this proportion is low (height of the partition approaches one), one can argue that elements which are in different cells are more different on the average than if this proportion is high (height of partition approaches zero). Thus, one of the measures discussed below makes two adjustments, one in the case where x and y are in the same cell and the other in the case where x and y are in different cells.

Definition of a Set of Metrics

In this section are defined a set of metrics for sorting data.

In the following section, three dissimilarity measures are defined which belong to that set.

Let T be the number of subjects who do the sorting. Each subject induces a partition, P_i, of the set S of stimulus elements. Define

$$|P_i| = \text{cardinality of the partition for subject } i$$
$$c_{ij} = \text{cell } j \text{ of } P_i$$
$$N_{ij} = \text{size of } c_{ij} \ .$$

The proposed set of metrics takes the form

$$[1] \qquad D_{x,y} = \sum_{i=1}^{T} D_{(i)x,y}$$

where $D_{(i)x,y}$ is defined by the equation

$$[2] \qquad \begin{aligned} D_{(i)x,y} &= A_{ij} \text{ if subject } i \text{ placed } x \text{ and } y \text{ in } c_{ij} \\ &= B_i \text{ if subject } i \text{ placed } x \text{ and } y \text{ in different cells} \\ &= 0 \text{ if } x = y \ . \end{aligned}$$

Subject to the constraints

$$[3] \qquad B_i \geq \operatorname*{Max}_{j} A_{ij}$$

$$[4] \qquad A_{ij} > 0 \ .$$

It is intended that A_{ij} be a function of n_{ij}. It makes possible the first compensation discussed in the previous section. B_i can be a function of the number of pairs of elements which are not included in any cell, making possible the second compensation discussed in the previous section. Note that, if stimulus x is the only member of a cell, then $D_{(i)x,y} = B_i$ for all $y \neq x$.

In order to prove that $D_{x,y}$ is a metric it is necessary only to prove that $D_{(i)x,y}$ is a metric, since the sum of metrics is a metric.

To prove that $D_{(i)x,y}$ is a metric requires proof of positivity, symmetry and the triangle inequality. Positivity is ensured by equations [3] and [4]. Symmetry is obvious from equation [2]. Rather than prove the triangle inequality, I shall prove the ultrametric inequality, which implies the triangle inequality:

For any two points, x and y, and for any third point, z,

$$\text{[5]} \qquad d(x,y) \leq \max\,[d(x,z), d(y,x)].$$

Following Millers' strategy for proof of the triangle inequality (1969), it is possible to distinguish five possible situations for any x, y, and z.

1. x, y, and z are all in c_{ij}. In this case,

   $$D_{(i)x,y} = D_{(i)x,z} = D_{(i)y,z} = A_{ij}.$$

   Clearly the ultrametric inequality is satisfied since $D_{(i)x,y}$ is equal to the maximum of the other two distances.
2. x, y, and z are all in different cells. In the second case,

   $$D_{(i)x,y} = D_{(i)x,z} = D_{(i)y,z} = B_i.$$

   The ultrametric inequality is satisfied by the same reasoning as in the first case.
3. x and y, are in c_{ij} and z is in a different cell. In this case

   $$D_{(i)x,z} = D_{(i)y,z} = B_i$$
   $$D_{(i)x,y} = A_{ij}.$$

   Clearly, $D_{(i)x,y}$ is less than or equal to the maximum of the other two distances.
4. x and z are in c_{ij} and y is in a different cell. In this case,

   $$D_{(i)x,y} = D_{(i)y,z} = B_i$$
   $$D_{(i)x,z} = A_{ij}.$$

   Here $D_{(i)x,y}$ is equal to the maximum of the other two distances.
5. y and z are in c_{ij} and x is in a different cell. In this case

   $$D_{(i)x,y} = D_{(i)x,z} = B_i$$
   $$D_{(i)y,z} = A_{ij}.$$

   Here again $D_{(i)x,y}$ is equal to the maximum of the other two distances.

Although the distances $D_{(i)x,y}$ satisfy the ultrametric inequality, it can be shown by a counter-example that $D_{x,y}$ does not necessarily satisfy the ultrametric inequality. In this example, there are two subjects doing the sorting. The first subject sorts y and z together into c_{ik} and puts x in a third cell. The second subject sorts x and z together into c_{2j} and puts y in a third cell. The distance measures from these two sortings are

$$D_{(1)y,z} = A_{1k}$$
$$D_{(1)x,y} = D_{(1)x,z} = B_1$$
$$D_{(2)x,z} = A_{2j}$$
$$D_{(2)y,z} = D_{(2)x,y} = B_2.$$

Then

$$D_{x,y} = B_1 + B_2$$
$$D_{x,z} = B_1 + A_{2,j}$$
$$D_{y,z} = A_{1,k} + B_2.$$

Clearly, $D_{x,y}$ is greater than or equal to the maximum of $D_{x,z}$ and $D_{y,z}$.

Definition of Three Dissimilarity Measures

The first measure, $Z_{x,y}$, is related to the concept of height of a partition (Boorman and Arabie, 1972). The height of P_i, H_i, is defined by the formula

$$[6] \qquad H_i = \sum_{j=1}^{|P_i|} H_{ij} = \sum_{j=1}^{|P_i|} \frac{(N_{ij})!/[2!(N_{ij}-2)!]}{N!/[2!(N-2)!]}$$

H_i has a value of zero if all elements of S are in different cells and a value of one if all elements of S are in a single cell. H_{ij} is the contribution to H_i for c_{ij} and is the proportion of pairs of elements which are found within c_{ij}. $Q_i = 1 - H_i$ is the proportion of pairs of elements which are not in the same cell.

To define $Z_{x,y}$ we first define the similarity measure for each subject $S_{(i)x,y}$ by the equation

[7]
$$S_{(i)x,y} \begin{cases} = -\text{Log}_2 H_{ij} & \text{if } x \text{ and } y \text{ are in } c_{ij} \\ = \text{Log}_2 Q_i & \text{if } x \text{ and } y \text{ are in different cells for subject } i \\ = \text{Log}_2 \{N!/[2!(N-2)!] + \varepsilon\} & \text{if } x = y. \end{cases}$$

Here ε is any number greater than zero. For the purposes of this discussion, we allow ε to be one. A value greater than zero for ε ensures that the similarity of an element to itself will be greater than the maximum similarity of an element to any other element. Define

$$C = \text{Log}_2 \{N!/[2!(N-2)!] + 1\}.$$

Then

$$D_{(i)x,y} = C - S_{(i)x,y}.$$

Using this formula,

$$\begin{aligned} A_{ij} &= C + \text{Log}_2 H_{ij} \\ &= \text{Log}_2 \{N!/[2!(N-2)!] + 1\} + \\ &\quad \text{Log}_2 \{(N_{ij})!/[2!(N_{ij}-2)!]\} - \\ &\quad \text{Log}_2 \{N!/[2!(N-2)!]\} \\ &\cong \text{Log}_2 \{(N_{ij})!/[2!(N_{ij}-2)!]\}. \end{aligned}$$

$$B_i = C - \text{Log}_2 [1 - H_i]$$

B$_i$
$$\begin{aligned} &= C - \text{Log}_2 \left(1 - \frac{\Sigma\{(N_{ij})!/[2!(N_{ij}-2)!]\}}{N!/[2!(N-2)!]}\right) \\ &= C - \text{Log}_2 (\{N!/[2!(N-2)!]\} - \\ &\quad \Sigma\{(N_{ij})!/[2!(N_{ij}-2)!]\}) + \\ &\quad \text{Log}_2 \{N!/[2!(N-2)!]\} \\ &\cong 2\,\text{Log}_2 \{N!/[2!(N-2)!]\} - \\ &\quad \text{Log}_2 (\{N!/[2!(N-2)!]\} - \\ &\quad \Sigma\{(N_{ij})!/[2!(N_{ij}-2)!]\}). \end{aligned}$$

The maximum value of A_{ij} is C, and occurs when the height of P_i is one. The minimum value of B_i is also C and occurs when the height of P_i is zero. Since the minimum value of B_i is equal to the maximum value of A_{ij}, and since A_{ij} is greater than zero, equations 3 and 4 are satisfied, and $Z_{x,y}$ is a metric.

$S_{(i)x,y}$ has an information theoretic flavor. H_{ij} is the probability that any two elements, chosen at random, will be included in c_{ij}. Thus, when x and y are included in c_{ij}, $S_{(i)x,y}$ is the Log_2 of the probability that the two elements are in c_{ij}. Similarly, Q_i is the probability that any two elements, chosen at random, will be found to be in different cells from each other. Thus, when x and y are placed in different cells, $S_{(i)x,y}$ is a negative number equal to the Log_2 of the probability that any two elements will be in different cells. The magnitude of $\text{Log}_2\, Q_i$ decreases as Q_i increases. If the subject makes no discriminations among elements by placing them all in the same cell, or if the subject places each element in a separate cell, P_i provides no information about the internal structure of the set S. In either of these two cases $S_{(i)x,y}$ is the same for all x and y, and $D_{(i)x,y}$ is uniformly equal to C.

The second and third measures are two cases from a general formula

$$[8] \qquad \begin{aligned} S_{(i)x,y} &= (N_{ij})^{\alpha} && \text{if } x \text{ and } y \text{ are in } c_{ij} \\ &= 0 && \text{if } x \text{ and } y \text{ are in different cells for subject } i \\ &= \text{Max}\,(N_{ij})^{\alpha} + \varepsilon && \text{if } x = y, \end{aligned}$$

where α is any number and ε is any number greater than zero. Let $C = \text{Max}\,(N_{ij})^{\alpha} + \varepsilon$.

α measures the degree of compensation for differences in cell size:

For $\alpha < 0$, small cells make higher increments to similarity than large cells.

For $\alpha = 0$, all cells make the same increment to similarity.

For $\alpha > 0$, large cells make higher increments to similarity than small cells.

Define

$$D_{(i)x,y} = C - S_{(i)x,y}$$

Then

$$A_{ij} = C - (N_{ij})^{\alpha}$$
$$B_{ij} = C.$$

Since $B_i > \text{Max } A_{ij}$ and $A_{ij} > 0$, the constraints of equations [3] and [4] are satisfied.

The second measure, $F_{x,y}$, is the same as is used by Miller (1969). Here $\alpha = 0$. This measure is neutral with respect to compensation for cell sizes.

For the third measure, $G_{x,y}$, $\alpha = +1$. For this measure, if x and y are in c_{ij}, the similarity of x to y is increased by N_{ij}. Thus, $G_{x,y}$, assigns more weight to larger cells.

The Empirical Test

The data for the empirical test consists of two bodies of sorting data, one for names of behaviors and the other for names of occupations. In each case the number of names is 34. The 50 subjects who sorted the behavior names were students in an introductory psychology class at the University of California, Irvine, during the spring of 1971.[1] The 54 subjects who sorted the occupations names were people who responded to an advertisement at Harvard University during the spring of 1969, and were mostly Harvard undergraduates or staff.

The three measures were computed for all pairs of stimuli for each set of data, and were then scaled in three and two dimensions using the TORSCA multidimensional scaling program (Young and Torgerson, 1968). Stress figures for the computations are listed in Table 1.

Table 1
Stress Figures for the Multidimensional Scalings

	Z	F	G	Z	F	G
Occupations	.143	.136	.129	.217	.208	.187
Behavior	.132	.123	.099	.177	.160	.141
	3 dimensions			2 dimensions		

[1] I am indebted to Mr. Celestin Kemikimba, who assisted with this experiment.

Measure Z consistently produces scalings with highest stress and Measure G consistently produces scalings with lowest stress, for a given data set and given number of dimensions.

Table 2 lists the correlations among distances in the multidimensional configurations. As with the stress computations, the Measure F appears to be intermediate to Measure Z and Measure G. In all cases, the correlation of F to G is higher than the correlation of Z to G for the same data and same number of dimensions. Statistical tests using the r to Z transformation (Hays, 1963) produce significance levels of $p < .001$ for the three tests.

The correlation of Z to F is also higher in all cases than the correlation of Z to G ($p < .001$). Both the pattern of stresses and

Table 2
Correlations among Scaled Distances,
for Three and Two Dimensions

Behaviors

		Z		F		G	
		3D	2D	3D	2D	3D	2D
Z	3D	X	.952	.947	.903	.905	N.C.
	2D		X	.964	.935	.923	N.C.
F	3D			X	.948	.972	N.C.
	2D				X	.926	N.C.
G	3D					X	N.C.
	2D						X

Occupations

Z		F		G			
3D	2D	3D	2D	3D	2D		
X	.904	.985	.891	.857	.830	3D	Z
	X	.908	.986	.872	.918	2D	
		X	.916	.911	.884	3D	F
			X	.908	.960	2D	
				X	.944	3D	G
					X	2D	

the correlation patterns are consistent with the logic of the dissimilarity measures, for which F is intermediate to Z and G in its treatment of cell size differences. Since the Z measure places emphasis on small cells, one would predict that it would preserve fine distinctions which are made by only part of the subjects. By contrast, almost all of the variability in the G measure should be accounted for by the higher-level distinctions which are made by most subjects. If there are two major clusters within the stimulus set, the G measure will provide little information about the internal organization of those two clusters, whereas the Z measure could be expected to recover the internal organization of the clusters. Measures which assign higher weight to general categories reduce the complexity of the data relative to measures which assign higher weight to fine distinctions. By so doing, the G measure runs the risk of producing a degenerate scaling solution, in which the scaled configuration consists simply of two or more clusters which tend to collapse to single points. When there is a degenerate solution, stress will tend to approach zero. Thus, the lower stress figures for the G measure could be due to a tendency to collapse clusters, by losing most of the internal structure of clusters.

Figures 1, 2, and 3 represent the two dimensional scalings of behavior names for the Z, F, and G measures, respectively. For all

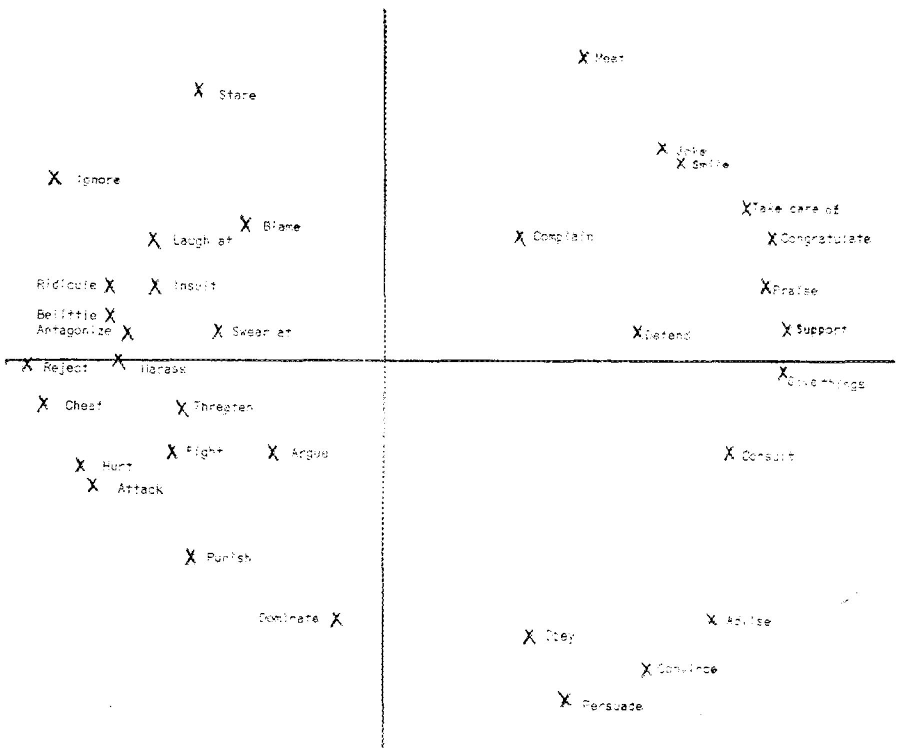

Figure 1
Two Dimensional Scaling of Behaviors, Z Measure Stress = .177

three scalings there is a clear first dimension, which is simply a distinction between friendly behaviors and unfriendly behaviors. For all three measures, the internal structure of the cluster of friendly behaviors is approximately the same, although "meet" moves from the top of the picture with the *Z* measure to the horizontal axis with the *F* measure, to the bottom of the picture for the *G* measure. However, the cluster of unfriendly behaviors changes radically from the *Z* measure to the *G* measure. With the *G* measure, it is clear that the cluster of unfriendly behaviors has begun to collapse. This trend is also apparent with the *F* measure, although the effect is much weaker. With the *Z* measure, it is also possible to make a tentative interpretation of the second dimension. Egalitarian behaviors appear to be at the top of the figure both for friendly and unfriendly behaviors, whereas behaviors which involve dominance and status differences appear to be near the bottom of the figure. "Punish," "dominate," "obey," "persuade," "convince," and "advise" take the most extreme negative values on di-

Figure 2
Two Dimensional Scaling of Behaviors, F Measure Stress = .160

mension 2, and all involve attempts to change the behavior of another person. The *G* measure pushes "punish" up towards the horizontal axis and assigns an extreme negative value to "meet", thereby destroying the interpretability of the second dimension.

The fact that the cluster of unfriendly behaviors collapses with the *G* measure, whereas the cluster of friendly behaviors remains intact, suggests that more subjects made fine distinctions among the friendly behaviors than did so among the unfriendly behaviors. Apparently a large number of subjects sorted all negative behaviors into a single pile, while making several distinctions among the positive behaviors. This effect can be predicted by the concept of marked and unmarked categories. In his discussion of this concept, Greenberg (1966) formulates the hypothesis that distinctions tend to occur for the unmarked category which become neutralized for the marked category. Between two categories, the marked category is formed from the unmarked category by the addition of a derivational affix. Thus, "friendly" plus the prefix "un" results in "unfriendly", which is the marked category. The concept of mark-

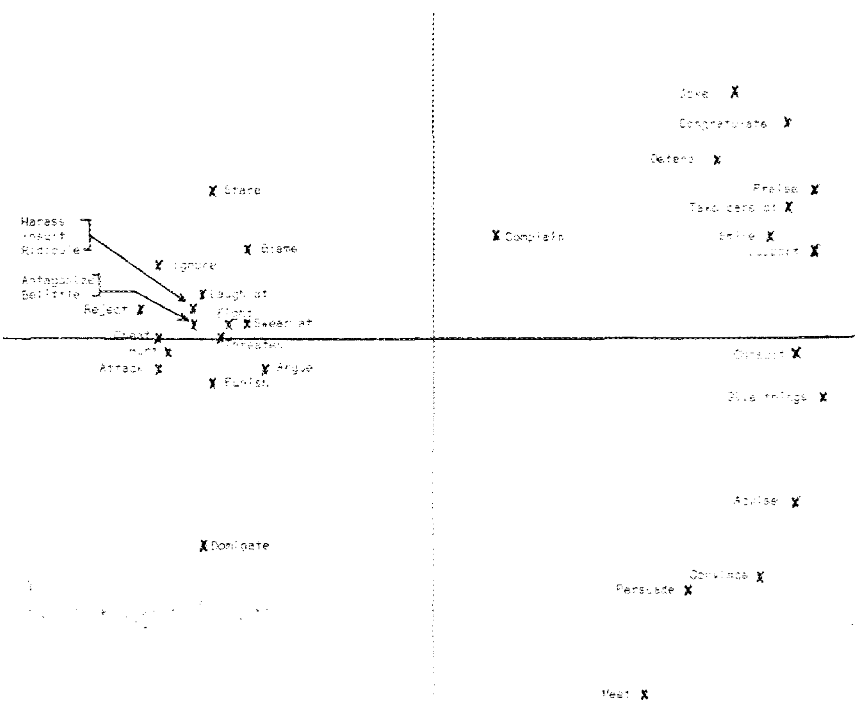

Figure 3
Two Dimensional Scaling of Behavior, G Measure Stress = .141

ing predicts that people will make fewer distinctions within the marked category than they do within the unmarked category. The present data tend to support this generalization.

The previous discussion has shown that the scaling solutions for the F measure are intermediate between those for the Z measure and those for the G measure, and that both the F and G measures produce less interpretable scaling solutions, with the G measure tending to collapse clusters. It is also relevant to ask whether the dissimilarity measure for the F measure are intermediate between those for the Z measure and those for the G measure; that is, whether the observed patterns of correlation are not simply an artifact of the scalng procedure. Table three lists the correlations among the dissimilarity measures. In both cases, the correlation of Z to G is lower than the correlation of Z to F and the correlation of F to G, and the differences in correlation are statistically significant ($p < .001$).

Table 3
Correlations among the Dissimilarity Measures

Behaviors

	Z	F	G
Z	X	.956	.822
F		X	.951
G			X

Occupations

	Z	F	G
Z	X	.932	.842
F		X	.926
G			X

Conclusions

The empirical investigation demonstrates statistically reliable differences between measure G and measure Z. These differences can be perceived in the multidimensional scalings of the behavior names. The G measure tends to collapse the cluster of unfriendly behaviors, thereby reducing the stress measure as a degenerate solution is approached. Although correlations among the three measures are all greater than .80, and correlations among distances in the multidimensional configurations are also all greater than .80, the Z measure is clearly the most satisfactory measure for multidimensional scaling purposes, because of the danger of degenerate solutions with measures which do not compensate for cell size differences by assigning a higher weight to small cells.

REFERENCES

Boorman, S. A. and P. Arabie. "Structural Measures and the Method of Sorting." In R. Shepard, A. K. Romney and S. B. Nerlove, eds., *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, Vol. I, Seminar Press, 1972.

Burton, M. L. "Semantic Dimensions of Occupation Names." In A. K. Romney, S. N. Shepard and S. B. Nerlove, eds., *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, Vol. II, Seminar Press, 1972.

Greenberg, Joseph. *Language Universals with Special Reference to Feature Hierarchies*. The Hague: Mouton, 1966.

Hays, W. L. *Statistics for Psychologists*, New York: Holt, Rinehart and Winston, 1963.

Johnson, S. C. "Hierarchical Clustering Schemes." *Psychometrika*, 1967, *32*, 241-54.

Kruskal, J. B. "Multidimensional Scaling by Optimizing Goodness of Fit to a Nometric Hypothesis." *Psychometrika*, 1964, *29*, 1-27.

Miller, George. "A Psychological Method to Investigate Verbal Concepts." *Journal of Mathematical Psychology*, 1969, *6*, 169-191.

Rosenberg, S., G. Nelson, and P. S. Vivekananthan. "A Multidimensional Approach to the Structure of Personality Impressions." *Journal of Personality and Social Psychology*, 1968, *9*, 283-294.

Shepard, Roger N. "The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function." *Psychometrika*, 1962, *27*, 125-139.

Shepard, Roger N. "Metric Structures in Ordinal Data." *Journal of Mathematical Psychology*, 1966, *3*, 287-315.

Young, F. W. and W. S. Torgerson. TORSCA, a FORTRAN IV Program for Shepard-Kruskal Multidimensional Scaling Analysis. *Behavioral Science*, 1969, *12*, 498.