Section II
THE BASIC MODEL

3 Multidimensional scaling by optimizing goodness-of-fit to a non-metric hypothesis*

Joseph B. Kruskal

The problem of multidimensional scaling, broadly stated, is to find n points whose interpoint distances match in some sense the experimental dissimilarities of n objects. Instead of dissimilarities the experimental measurements may be similarities, confusion probabilities, interaction rates between groups, correlation coefficients, or other measures of proximity or dissociation of the most diverse kind. Whether a large value implies closeness or its opposite is a detail and has no essential significance. What is essential is that we desire a monotone relationship, either ascending or descending, between the experimental measurements and distances in the configuration.

We shall refer only to dissimilarities and similarities, but we explicitly include in these terms all the varied kinds of measurement indicated above. We also note that similarities can always be replaced by dissimilarities (for example, replace s_{ij} by $k - s_{ij}$). Since our procedure uses only the rank ordering of the measurements, such a replacement does no violence to the data.

According to Torgerson ([17], p. 250), the methods in use up to the time of his book follow the general two-stage procedure of first using a one-dimensional scaling technique to convert the dissimilarities or similarities into distances, and then finding points whose interpoint distances have approximately these values. The statistical question of goodness of fit is treated separately, not as an integral part of the procedure. Despite the success these methods have had, their rationale is not fully satisfactory. Due to the nature of the one-dimensional scaling techniques available, these methods either accept the averaged dissimilarities or some fixed transformation of them as distances or else use the variability of the data as a critical element in forming the distances.

A quite different approach to multidimensional scaling may be found in Coombs [5]. However, its rationale is also subject to certain criticisms.

A major advance was made by Roger Shepard [15a, b], who introduced two major innovations. First, he introduced as the central feature the goal of obtaining a monotone relationship between the experimental dissimilarities or similarities and the distances in the configuration. He clearly indicates that the satisfactoriness of a proposed solution should be judged by the degree to

*reprinted from *Psychometrika*, **29**, 1964, pp. 1–27

which this condition is approached. Monotonicity as a goal was proposed earlier [for example, see Shepard ([14], pp.333–334) and Coombs ([5], p. 513)], but never so strongly. Second, he showed that simply by requiring a high degree of satisfactoriness in this sense and without making use of variability in any way, one obtains very tightly constrained solutions and recovers simultaneously the form of the assumed but unspecified monotone relationship. In other words, he showed that the rank order of the dissimilarities is itself enough to determine the solution. (In a later section we state a theorem which further clarifies this situation.) Thus his technique avoids all the strong distributional assumptions which are necessary in variability-dependent techniques, and also avoids the assumption made by other techniques that dissimilarities and distances are related by some fixed formula. In addition, it should be pointed out that Shepard described and used a practical iterative procedure for finding his solutions with the aid of an automatic computer.

However, Shepard's technique still lacks a solid logical foundation. Most notably, and in common with most other authors, he does not give a mathematically explicit definition of what constitutes a solution. He places the monotone relationship as the central feature, but points out ([15a], p. 128) that a low-dimensional solution cannot be expected to satisfy this criterion perfectly. He introduces a measure of departure δ from this condition [15a, pp. 136–137] but gives it only secondary importance as a criterion for deciding when to terminate his iterative process. His iterative process itself implies still another way of measuring the departure from monotonicity.

In this paper we present a technique for multidimensional scaling, similar to Shepard's, which arose from attempts to improve and perfect his ideas. Our technique is at the same statistical level as least-squares regression analysis. We view multidimensional scaling as a problem of statistical fitting—the dissimilarities are given, and we wish to find the configuration whose distances fit them best.

"To fit them best" implies both a goal and a way of measuring how close we are to that goal. Like Shepard, we adopt a monotone relationship between dissimilarity and distance as our central goal. However, we go further and give a natural quantitative measure of nonmonotonicity. Briefly, for any given configuration we perform a monotone regression of distance upon dissimilarity, and use the residual variance, suitably normalized, as our quantitative measure. We call this the *stress*. (A complete explanation is given in the next section.) Thus for any given configuration the stress measures how well that configuration matches the data.

Once the stress has been defined and the definition justified, the rest of the theory follows without further difficulty. The solution is defined to be the best-fitting configuration of points, that is, the configuration of minimum stress.

There still remains the problem of computing the best-fitting configuration. However, this is strictly a problem of numerical analysis, with no psychological implications. (The literature reflects considerable confusion between the main problem of definition and the subsidiary problem of compu-

tation.) In a companion paper [12] we present a practical method of computation, so that our technique should be usable on many automatic computers. (A program which should be usable at many large computer installations is available on request.)

In our two papers we extend both theory and the computational technique to handle missing data and certain non-Euclidean distances, including the city-block metric. It would not be difficult to extend the technique further so as to reflect unequal measurement errors.

We wish to express our gratitude to Roger Shepard for his valuable discussions and for the free use of his extensive and valuable collection of data, obtained from many sources. All the data used in this paper come from that collection.

The Stress

In this section we develop the definition of stress. We remark in advance that since it will turn out to be a "residual sum of squares," it is positive, and the smaller the better. It will also turn out to be a dimensionless number, and can conveniently be expressed as a percentage. Our experience with experimental and synthetic data suggests the following verbal evaluation.

Stress	Goodness of fit
20%	poor
10%	fair
5%	good
$2\frac{1}{2}$%	excellent
0%	"perfect"

By "perfect" we mean only that there is a perfect monotone relationship between dissimilarities and the distances.

Let us denote the experimentally obtained dissimilarity between objects i and j by δ_{ij} . We suppose that the experimental procedure is inherently symmetrical, so that $\delta_{ij} = \delta_{ji}$. We also ignore the self-dissimilarities δ_{ii} . Thus with n objects, there are only $n(n-1)/2$ numbers, namely δ_{ij} for $i < j; i = 1, \cdots, n-1; j = 2, \cdots, n$. We ignore the possibility of ties; that is, we assume that no two of these $n(n-1)/2$ numbers are equal. Later in the paper we will be able to abandon every one of the assumptions given above, but for the present they make the discussion much simpler. Since we assume no ties, it is possible to rank the dissimilarities in strictly ascending order:

$$\delta_{i_1 j_1} < \delta_{i_2 j_2} < \delta_{i_3 j_3} < \cdots < \delta_{i_M j_M} .$$

Here $M = n(n-1)/2$.

We wish to represent the n objects by n points in t-dimensional space. Let us call these points $x_1, \cdots, x_n$. We shall suppose for the present that we know what value of t we should use. Later we discuss the question of determining the appropriate value of t. (Formally and mathematically, it is possible to use any number of dimensions. The appropriate value of t is a matter of scientific judgment.)

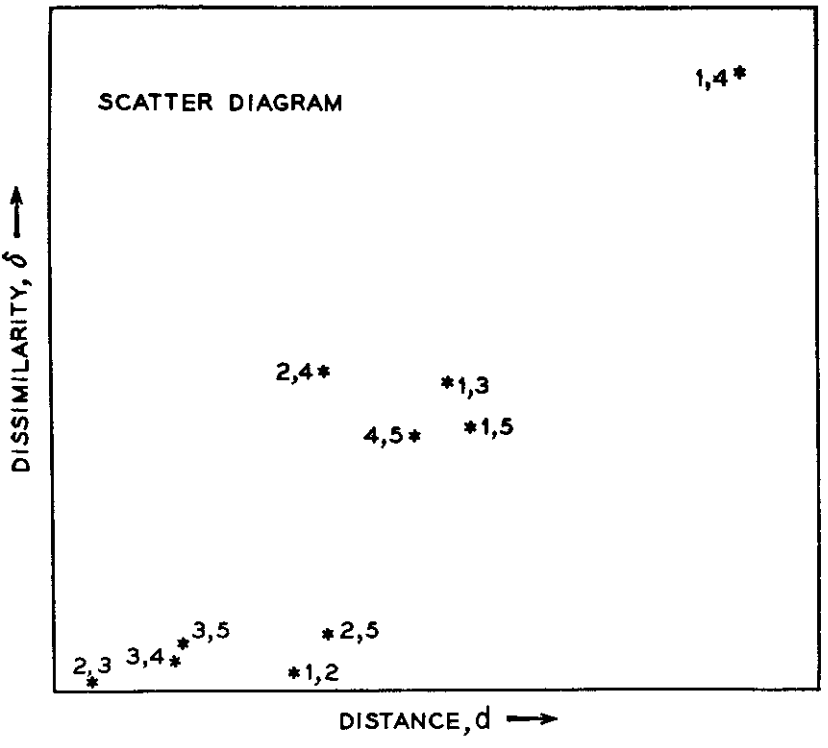

FIGURE 1

Let us suppose we have n points in t-dimensional space. We call this a *configuration.* Our first problem is to evaluate how well this configuration represents the data. Later on we shall want to find the configuration which represents the data best. At the moment, however, we are only concerned with constructing the criterion by which to judge configurations. To do so, let d_{ij} denote the distance from x_i to x_j . If x_i is expressed in orthogonal coordinates by

$$x_i = (x_{i1}, \cdots, x_{is}, \cdots, x_{it}),$$

then we have

$$d_{ij} = \sqrt{\sum_{s=1}^{t} (x_{is} - x_{js})^2}.$$

In order to see how well the distances match the dissimilarities, large with large and small with small, let us make a scatter diagram (Fig. 1). There are M stars in the diagram. Each star corresponds to a pair of points, as shown. Star (i, j) has abscissa d_{ij} and ordinate δ_{ij} . This diagram is fundamental to our entire discussion. We shall call it simply the *scatter diagram.*

Let us first ask "What should perfect match mean?" Surely it should mean that whenever one dissimilarity is smaller than another, then the corresponding distances satisfy the same relationship. In other words, perfect match should mean that if we lay out the distances d_{ij} in an array

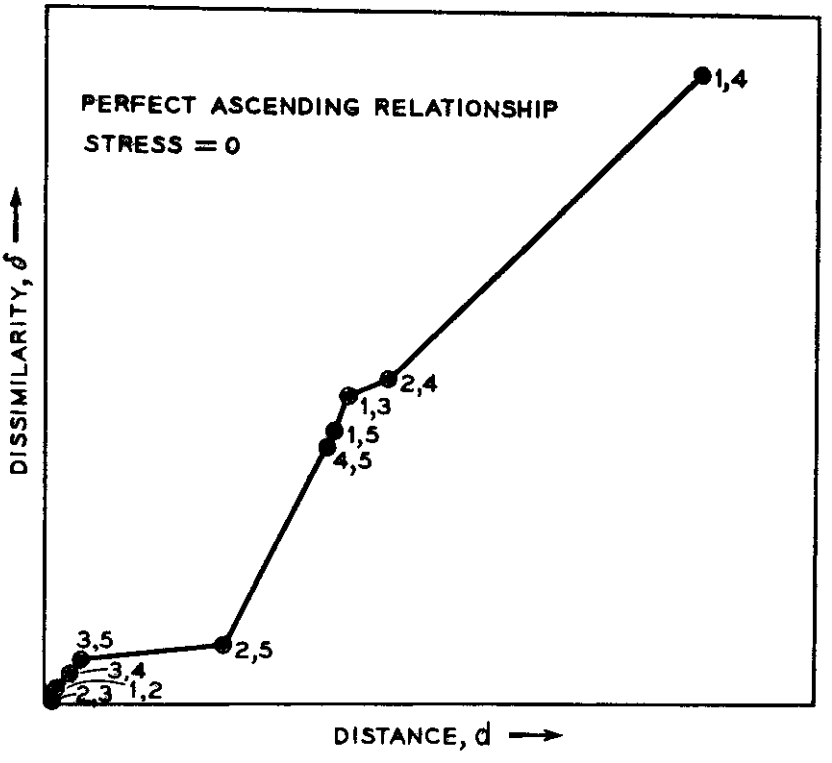

FIGURE 2

$$d_{i_1 j_1}\ ,\ d_{i_2 j_2}\ ,\ d_{i_3 j_3}\ ,\ \cdots\ ,\ d_{i_M j_M}$$

corresponding to the array of dissimilarities given above, then the smallest distance comes first, and the other distances follow in ascending order. In terms of the scatter diagram, this means that as we trace out the stars one by one from bottom to top, we always move to the right, never to the left. This fails in Fig. 1, but holds in Fig. 2.

To measure how far a scatter diagram such as Fig. 1 departs from the ideal of perfect fit, it is natural to fit an ascending curve to the stars as in Fig. 3 and then to measure the deviation from the stars to the curve. This is precisely what we do. However, the details are of critical importance.

Should we measure deviations between the curve and stars along the distance axis or along the dissimilarity axis? The answer is "along the distance axis." For if we measure them along the dissimilarity axis, we shall find ourselves doing arithmetic with dissimilarities. This we must not do, because we are committed to using only the rank ordering of the dissimilarities. To say the same thing in a different way, we wish to measure goodness of fit in such a way that monotone distortion of the dissimilarity axis will not have any effect. This clearly prevents us from measuring deviations along the dissimilarity axis.

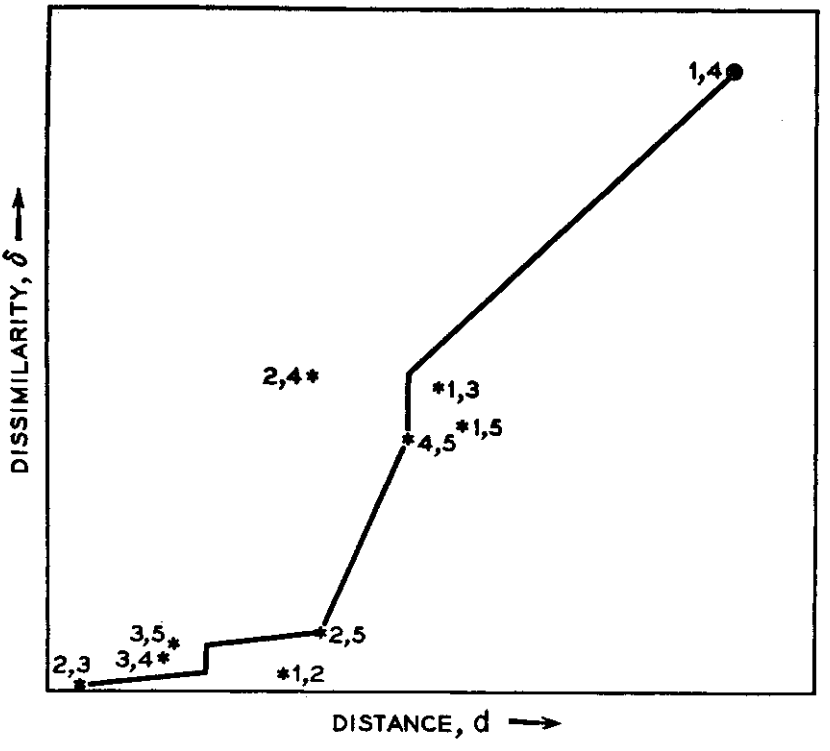

FIGURE 3

Having decided to measure the deviations along the distance axis, we next see that we do not actually need the whole curve, but only M points on it, as shown in Fig. 4. The rest of the curve does not enter into the calculation of deviations. We may continue to talk of fitting a curve, but all we mean is fitting the points.

Each point we fit shares the value of δ with the corresponding star, but has its own value of d. If a star is located at $(d_{ij}\ ,\ \delta_{ij})$, then we denote the corresponding point by $(\hat{d}_{ij}\ ,\ \delta_{ij})$. Thus fitting the curve means no more than fitting the values of $\hat{d}_{ij}$.

We realize of course that the numbers $\hat{d}_{ij}$ are not distances. There is no configuration whose interpoint distances are $\hat{d}_{ij}$. The $\hat{d}_{ij}$ are merely a monotone sequence of numbers, chosen as "nearly equal" to the d_{ij} as possible, which we use as a reference to measure the nonmonotonicity of the numbers d_{ij} . To simplify the discussion, we delay the precise definition of $\hat{d}_{ij}$ for a little while.

The fitted curve was of course intended to be ascending. Phrased in terms of the M points $(\hat{d}_{ij}\ ,\ \delta_{ij})$ which in effect constitute the curve, this means that as we trace out these points from bottom to top, we never move to the left but only to the right. Phrased in terms of the numbers $\hat{d}_{ij}$, it means that when they are arranged in the standard order

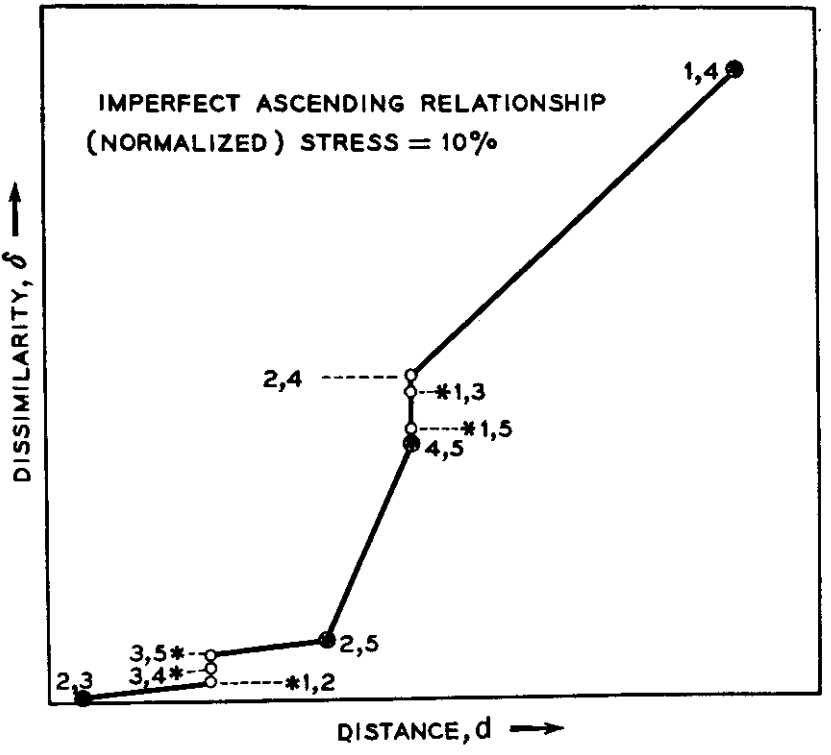

FIGURE 4

$$\hat{d}_{i_1 j_1}, \hat{d}_{i_2 j_2}, \hat{d}_{i_3 j_3}, \cdots, \hat{d}_{i_M j_M},$$

then each $\hat{d}_{ij}$ is greater than or equal to the one before it, namely

$$\hat{d}_{i_1 j_1} \leqq \hat{d}_{i_2 j_2} \leqq \hat{d}_{i_3 j_3} \leqq \cdots \leqq \hat{d}_{i_M j_M} \qquad \text{(Mon)}.$$

Whenever any numbers satisfy these inequalities, we shall say that they are *monotonically related* to the d_{ij}.

Now suppose we have the fitted values $\hat{d}_{ij}$, which satisfy (Mon) of course. Then the horizontal deviations are $d_{ij} - \hat{d}_{ij}$. How shall we combine these many individual deviations into a single overall deviation? Following a time-honored tradition of statistics, we square each deviation and add the results:

$$\text{raw stress} = S^* = \sum_{i<j} (d_{ij} - \hat{d}_{ij})^2.$$

Except for normalization, this is our measure of goodness of fit. It measures how well the given configuration represents the data. And very prosaic looking it is too—nothing more than the old familiar "residual sum of squares" associated with so many fitting techniques. It is special in only two ways: first, in the use of distance axis deviations; second, because of the fact that the fitted curve is chosen not from a "parametric" family of curves, such

as polynomials or trigonometric series, but from a "nonparametric" family of curves, namely, all monotone ascending curves.

The raw stress still lacks certain desirable properties. Most notably, while it is clearly invariant under rigid motions of the configuration (rotation, translations, and reflections), it is not invariant under uniform stretching and shrinking of the configuration. In other words, if we stretch the configuration $x_1, \cdots, x_n$ by the factor k to the configuration $kx_1, \cdots, kx_n$, that is, replace each point $(x_{i1}, \cdots, x_{it})$ by $(kx_{i1}, \cdots, kx_{it})$, then the raw stress changes. In fact, it changes from S^* to k^2S^* because the numbers $\hat{d}_{ij}$ also change by the factor k. Surely sheer enlargement of a configuration should not change how well it fits the data, for the relationships between the distances do not change. An obvious way to cure this defect in the raw stress is to divide it by a scaling factor, that is, a quantity which has the same quadratic dependence on the scale of the configuration that raw stress does. Such a scaling factor is easily found. We use

$$T^* = \sum_{i<j} d_{ij}^2 .$$

Thus

$$\frac{S^*}{T^*} = \frac{\sum_{i<j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}$$

is a measure of goodness of fit which has all the desirable properties of S^*, and in addition is invariant under change of scale, that is, uniform stretching or shrinking. This is the normalization. (Another plausible scaling factor is the variance of the numbers d_{ij}. We plan to compare these scaling factors elsewhere.)

Finally, it is desirable to use the square root of this expression, which is analogous to choosing the standard deviation in place of the variance. Thus our definition of the normalized stress is

$$\text{stress} = S = \sqrt{\frac{S^*}{T^*}} = \sqrt{\frac{\sum_{i<j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}}.$$

Again we emphasize that this measures how well the given configuration represents the data. Smaller stress means better fit. Zero stress means "perfect" fit in our special sense.

Now it is easy to define the $\hat{d}_{ij}$. They are the numbers which minimize S (or equivalently, S^*) subject to the constraint (Mon). Thus we may condense our entire definition of stress into the following formula.

$$S(x_1, \cdots, x_n) = \text{stress of the fixed configuration } x_1, \cdots, x_n$$

$$= \min_{\substack{\text{numbers } \hat{d}_{ij} \\ \text{satisfying (Mon)}}} \sqrt{\frac{\sum (d_{ij} - \hat{d}_{ij})^2}{\sum d_{ij}^2}}.$$

We point out that this minimization is accomplished not by varying a trial

set of values for the $\hat{d}_{ij}$, but rather by a rapid, efficient algorithm which is described in detail in the companion paper [12].

Now that we have defined the stress, we have a quantitative way of evaluating any configuration. Clearly the configuration we want is the configuration whose stress is a minimum, for this is the configuration which best fits the data. Thus we define

$$\text{stress in } t \text{ dimensions} = \min_{\substack{\text{all } t\text{-dimensional} \\ \text{configurations}}} S(x_1 , \cdots , x_n),$$

and we define the best-fitting configuration to be the one which achieves this minimum stress.

How do we find the minimum-stress configuration? We may answer this question at three levels. At the intuitive level, we may describe the procedure as one of successive approximation. We start with an arbitrary configuration, move all the points a little so as to improve it a bit, and then repeat this procedure until we reach the configuration from which no improvement is possible. Typically, anywhere from 15 to 100 such steps are necessary to reach the final configuration. Roughly speaking, we move points x_i and x_j closer together if $\hat{d}_{ij} < d_{ij}$, and apart in the opposite case, so as to make d_{ij} more like $\hat{d}_{ij}$. Of course, each point x_i is subject to many such motions at once, and usually these will be in partial conflict.

At the theoretical level, we see that our problem is to minimize a function of many variables, namely $S(x_1 , \cdots , x_n)$. Actually the stress S is a function of nt variables, as each vector x_i has t coordinates. The problem of minimizing a function of many variables is a standard problem in numerical analysis, and to solve it we adopt a widely used iterative technique known as the "method of gradients" or the "method of steepest descent."

Finally, at the practical level, we give in a companion paper [12] all the important details necessary to perform this iterative technique successfully.

An Example

To illustrate these ideas, we use synthetic data based on a 15-point configuration in the plane. Our configuration is shown by the + signs in Fig. 11. It was used by Shepard ([15b], p. 221) and taken by him from Coombs and Kao ([6], p. 222). To create the 105 dissimilarities we applied a monotone distortion to the interpoint distances, and then added independent random normal deviates to the distorted distances. Specifically,

$$\delta_{ij} = -(0.9) \exp [-(1.8)d_{ij}] - 0.1 + \eta_{ij} ,$$

where η_{ij} is normal with mean 0 and standard deviation 0.01.

We analyze these synthetic data in two dimensions ($t = 2$). The arbitrary starting configuration is shown by numbered circles in Fig. 5. (This and many later figures were created automatically by the computer with the aid of the General Dynamics Electronics Model SC-4020 Highspeed Microfilm Printer.) The lines show the motion of the first iteration to the next,

slightly better configuration. The stress of the first configuration is 47.3%. After one iteration it is down to 44.3%. After ten iterations the configuration has become that in Fig. 6, with stress 2.92%. (For most practical purposes the calculation could stop here, as the configuration hardly changes after this.) After fifty iterations the minimum-stress configuration shown in Fig. 7 is reached; its stress is 2.48%. The scatter diagrams of these three configura-

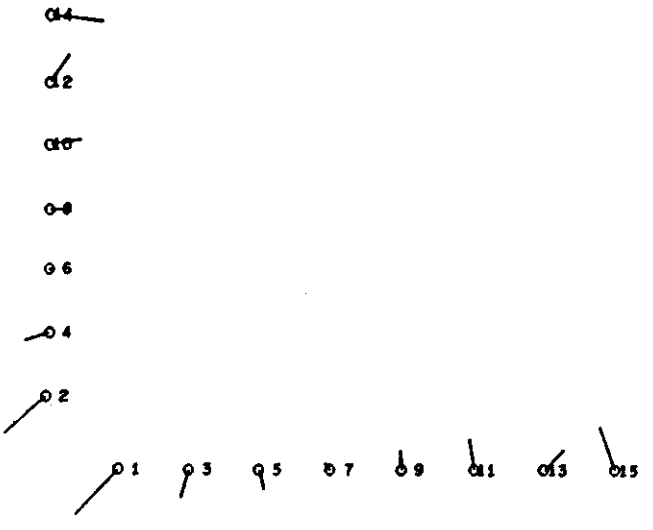

FIGURE 5
Initial Configuration (Coombs and Kao Data)

tions are shown in Figs. 8, 9, and 10. The monotone distorting function has been accurately recovered, and is displayed in the last of these scatter diagrams.

To show how accurately the original configuration has been recovered, we display in Fig. 11 the recovered configuration together with the original configuration of Coombs and Kao. The recovered configuration has been reflected and rotated by eye into best apparent agreement with the original configuration for this purpose. Since the angular position of the recovered configuration is quite arbitrary, this is entirely legitimate.

Another obvious way of measuring how nearly alike the two configurations are is to compare the distances $d_{ij}^{(1)}$ within one configuration with the distances $d_{ij}^{(2)}$ within the other. Corresponding distances differ typically by 3.16%. More precisely, the expression

$$\sqrt{\frac{\sum_{i<j} (d_{ij}^{(1)} - d_{ij}^{(2)})^2}{\sum_{i<j} \left(\frac{d_{ij}^{(1)} + d_{ij}^{(2)}}{2}\right)^2}}$$ has the value 0.0316.

How Many Dimensions?

So far we have assumed that the number of dimensions to be used is fixed and known. In practice, this is seldom the case. The final determination of how many coordinates to recover from the data rests ultimately with the

scientific judgment of the experimenter. However, we can suggest certain aids.

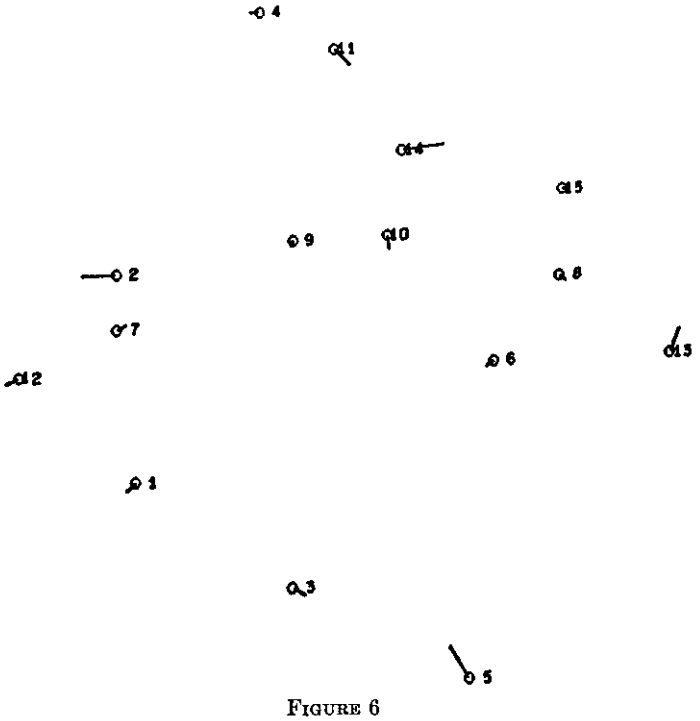

FIGURE 6
Configuration After 10 Iterations (Coombs and Kao Data)

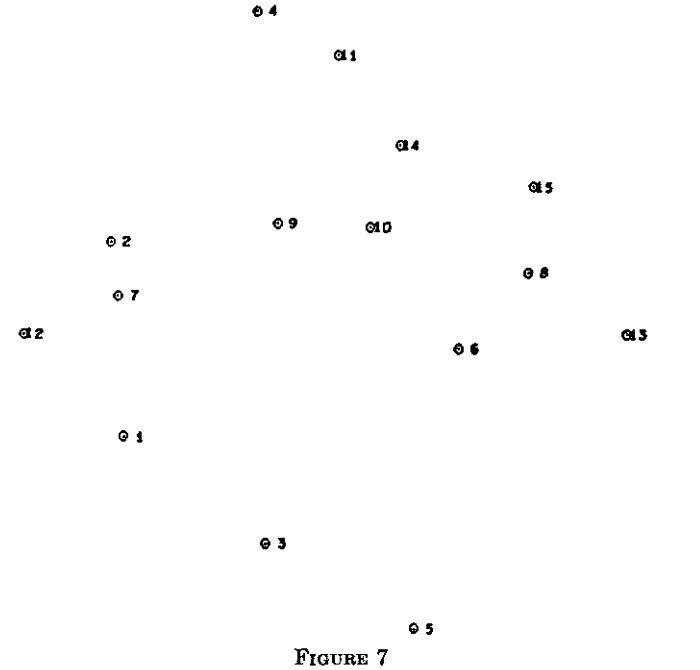

FIGURE 7
Configuration After 50 Iterations (Coombs and Kao Data)

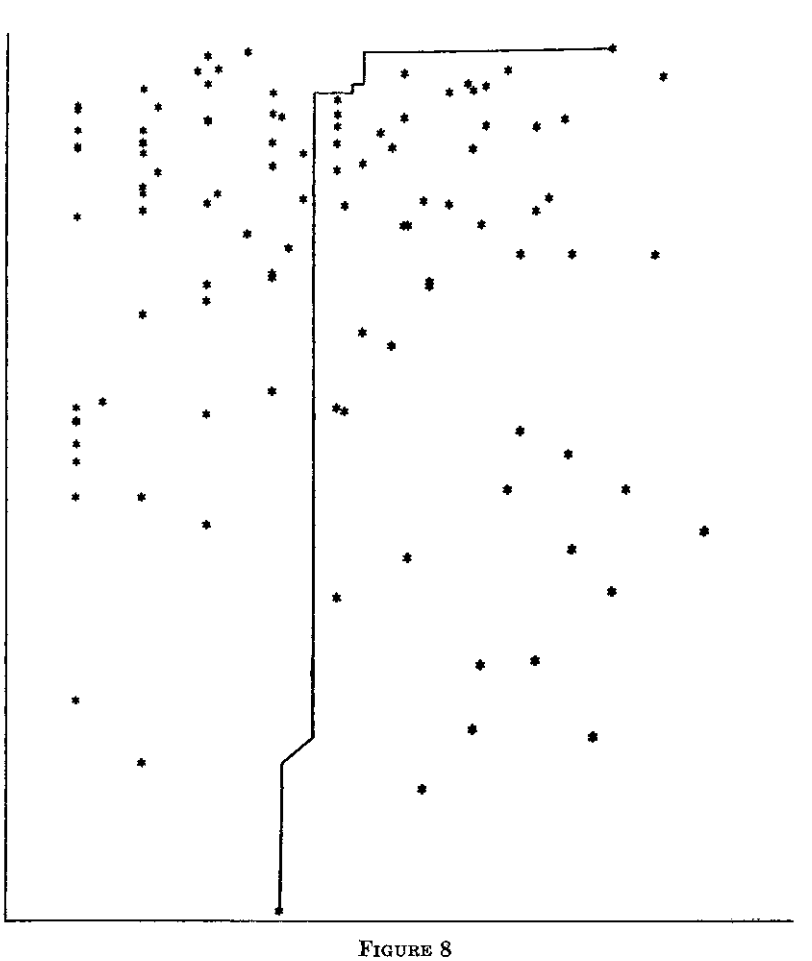

FIGURE 8
Initial Scatter Diagram (Coombs and Kao Data)

The analysis should be done in several dimensions, and a graph plotted to show the dependence of minimum stress on dimension. Of course, as t increases, minimum stress decreases. For $t \geqq n - 1$, the minimum stress is always 0. (Perfect match can always be managed with n points in $n - 1$ dimensions.) It is reasonable to choose a value of t which makes the stress acceptably small, and for which further increase in t does not significantly reduce stress. Good data sometimes exhibit a noticeable elbow in the curve, thus pointing to the appropriate value of t.

A second criterion lies in the interpretability of the coordinates. If the t-dimensional solution provides a satisfying interpretation, but the $(t + 1)$-dimensional solution reveals no further structure, it may be well to use only the t-dimensional solution. A third criterion can be used if there is an independent estimate of the statistical error of the data. The more accurate the data, the more dimensions one is entitled to extract.

To study the question of dimensionality, we first use synthetic data. Separate sets of ten, fifteen, and twenty random points in six dimensions were

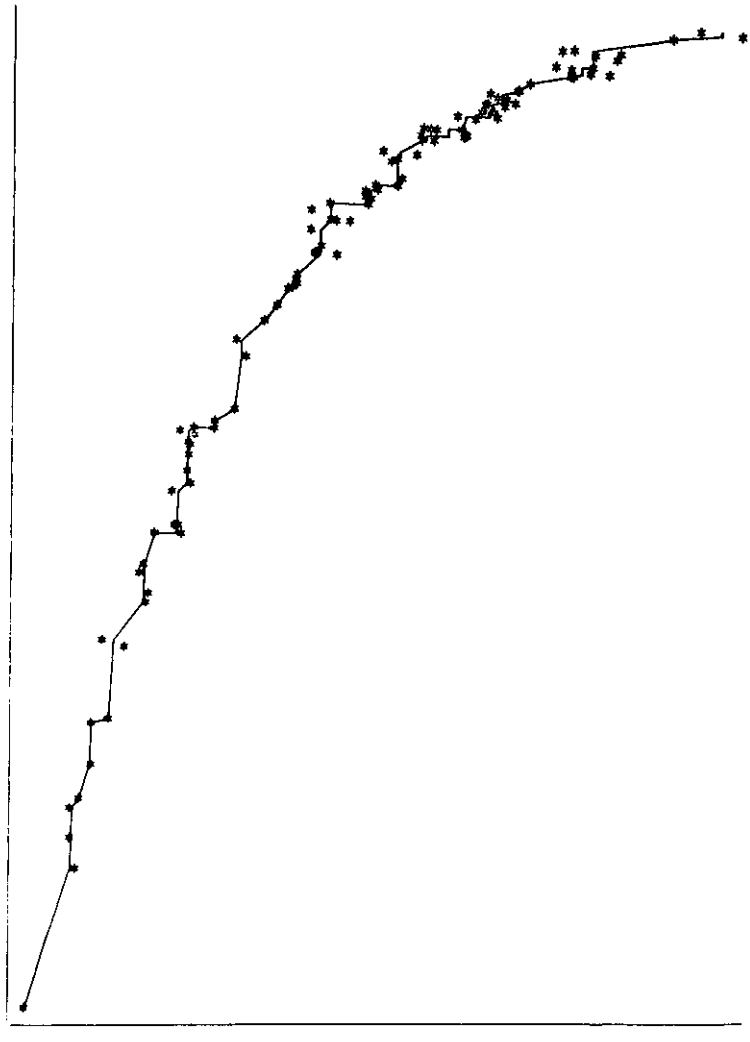

FIGURE 9

Scatter Diagram After 10 Iterations (Coombs and Kao Data)

chosen. The actual distances were used as dissimilarities δ_{ij} . Fig. 12 shows how stress varies with dimension for these three sets of data. A perfect match is obtained in six dimensions. The ten-point curve displays a distinct elbow, which strongly suggests the use of three dimensions. Of course, with error-free synthetic data, further coordinates may be successfully extracted, but even with excellent experimental data this curve would make the use of more than three dimensions quite dubious. (Examination of the original configuration of ten points shows that by chance it lies very nearly in a three-dimensional subspace.) The fifteen- and twenty-point curves are much less clear. If we obtained curves similar to these but without perfect fit in six dimensions from real data, then three dimensions would seem advisable, four would also seem reasonable, and five might be justified by other considerations, such as

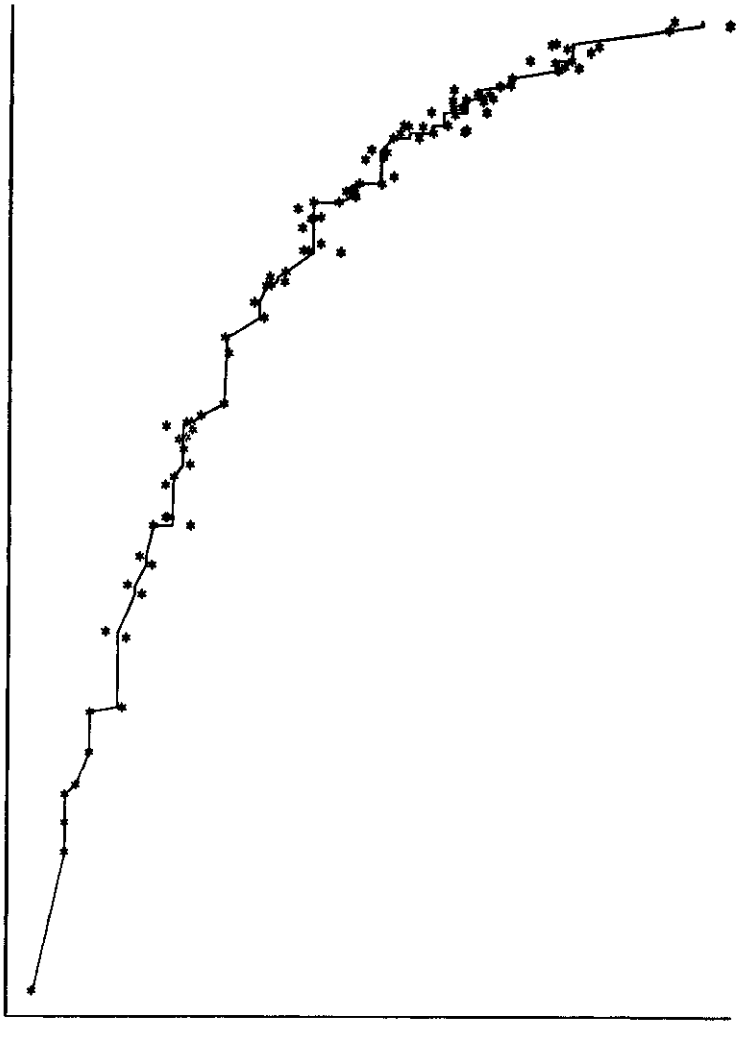

FIGURE 10

Scatter Diagram After 50 Iterations (Coombs and Kao Data)

good interpretability or independent indications of very low variability in the data.

Let us illustrate these ideas with data from Indow and Uchizono [9]. (The dissimilarities themselves did not appear in the paper. We thank Professor Indow for providing them.) They obtained direct judged dissimilarities between 21 colors of constant brightness, using an ingenious technique. It may seem obvious that the analysis should be done in two dimensions. However, there is the possibility that colors of constant brightness may be best described as lying on a *curved* two-dimensional surface. If this should be the case, we would want $t = 3$. In any case, it is instructive to see what happens. Fig. 13 shows the dependence of stress on dimension. The elbow in the curve at dimension 2 confirms our natural expectation that two dimensions

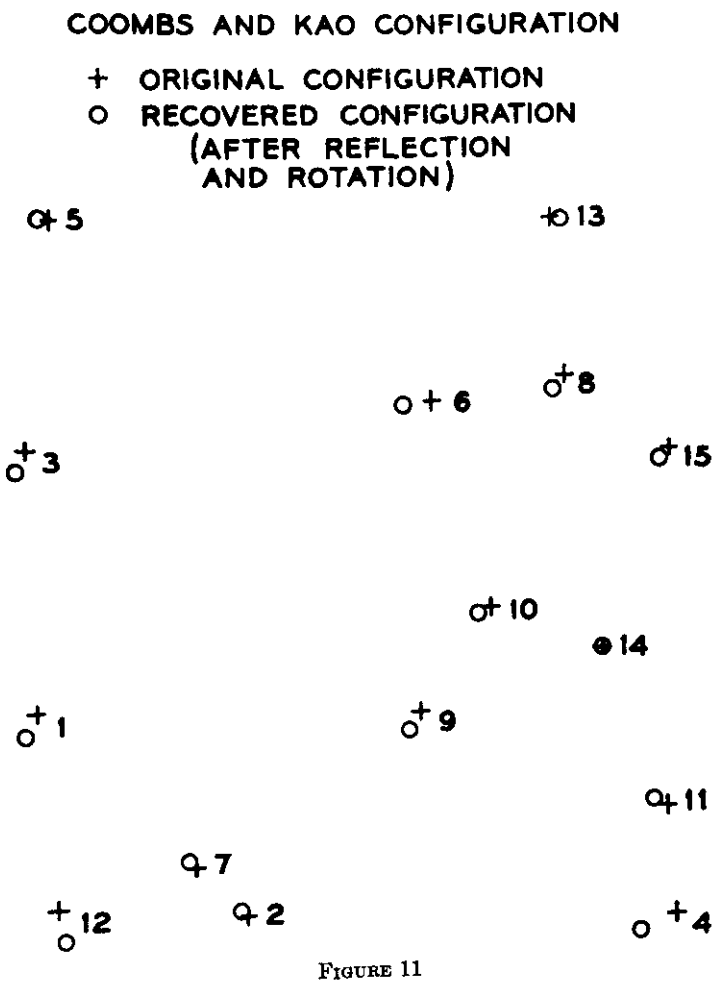

FIGURE 11

are appropriate, but does not completely rule out the possibility that three dimensions might become appropriate with more comprehensive data of the same sort. Figs. 14 and 15 show the configuration and the scatter diagram when the dimension is two. The configuration, which resembles the one given by Indow and Uchizono, corresponds roughly to the Munsell diagram for the 21 colors, but with considerable stretching and shrinking in various places. The scatter diagram, with a stress of 7.27%, would be classified as fair-to-good.

A very similar experiment by Indow and Kanazawa [10] supplies a

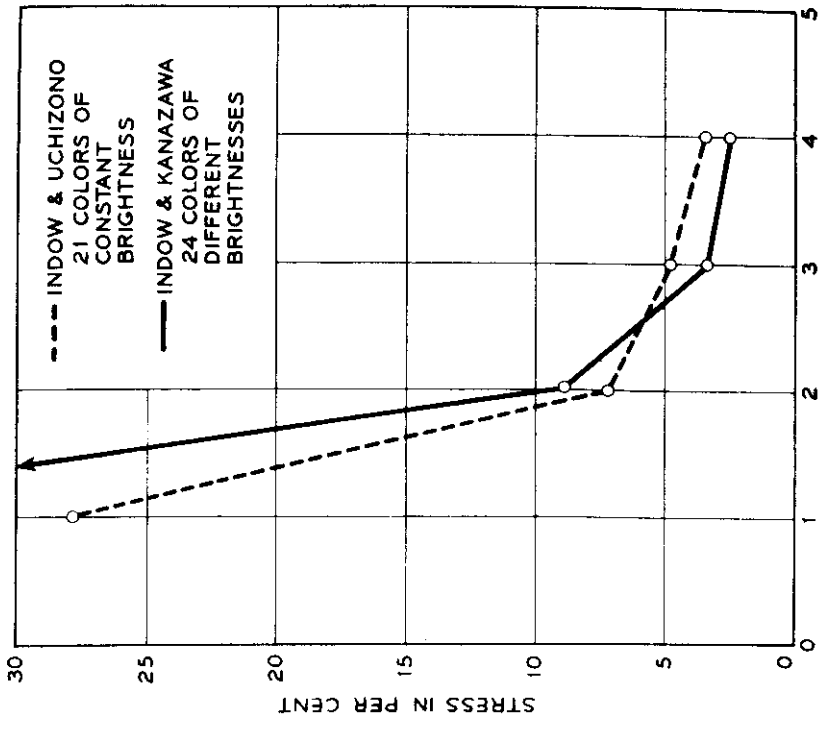

FIGURE 13

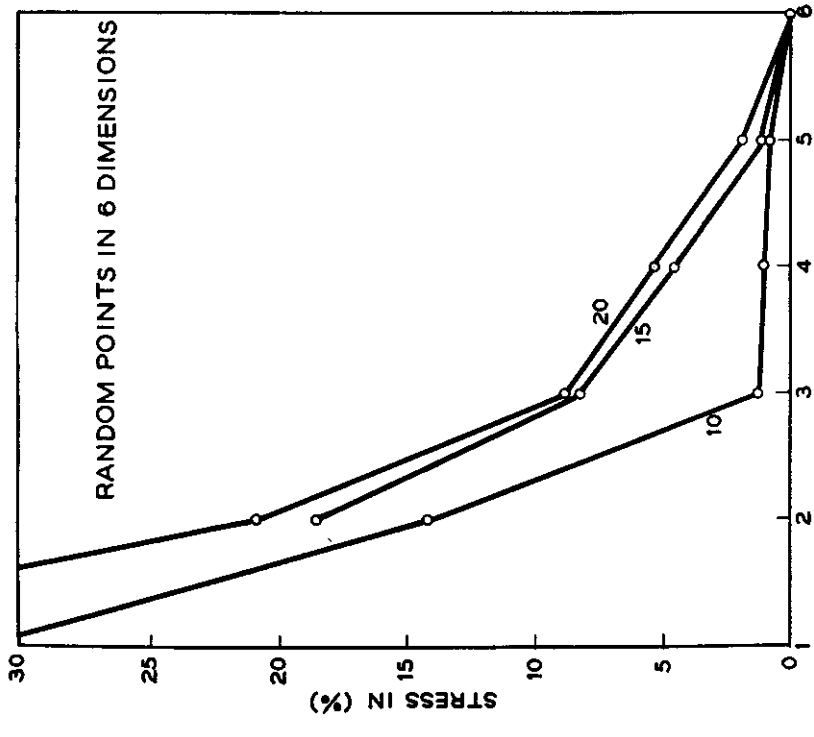

FIGURE 12

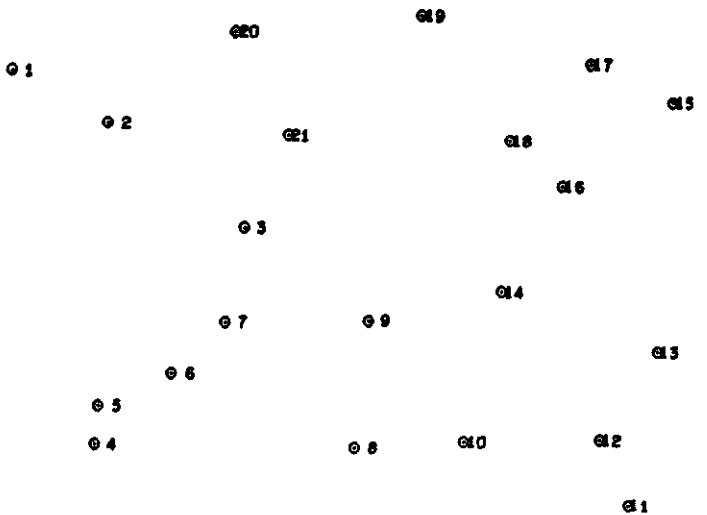

FIGURE 14
Configuration for 21 Colors of Constant Brightness
(Indow and Uchizono)

FIGURE 15
Scatter Diagram for 21 Colors of Constant Brightness
(Indow and Uchizono Data)

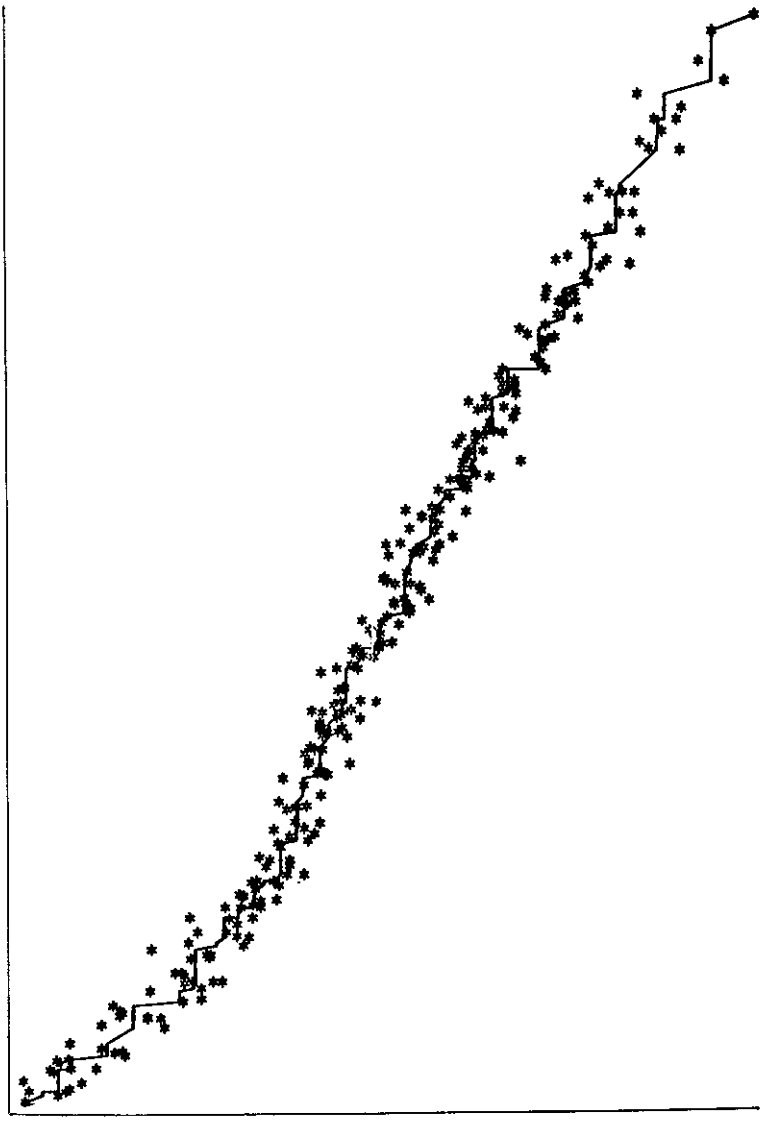

FIGURE 16
Scatter Diagram for 24 Colors of Varying Brightness (Indow and Kanazawa Data)

second illustration. In this experiment 24 colors of differing brightness were used. Fig. 13 fits well with our expectation that three dimensions are appropriate. The reason that the stress is fairly small in two dimensions is that after rotation to principal axes the third recovered coordinate varies over only half the range of the first two coordinates. This third coordinate corresponds approximately to brightness. The scatter diagram in three dimensions (Fig. 16) has a stress of 3.67%, and would be classified as fair-to-excellent. Our configuration in three dimensions resembles that obtained by Indow and Kanazawa.

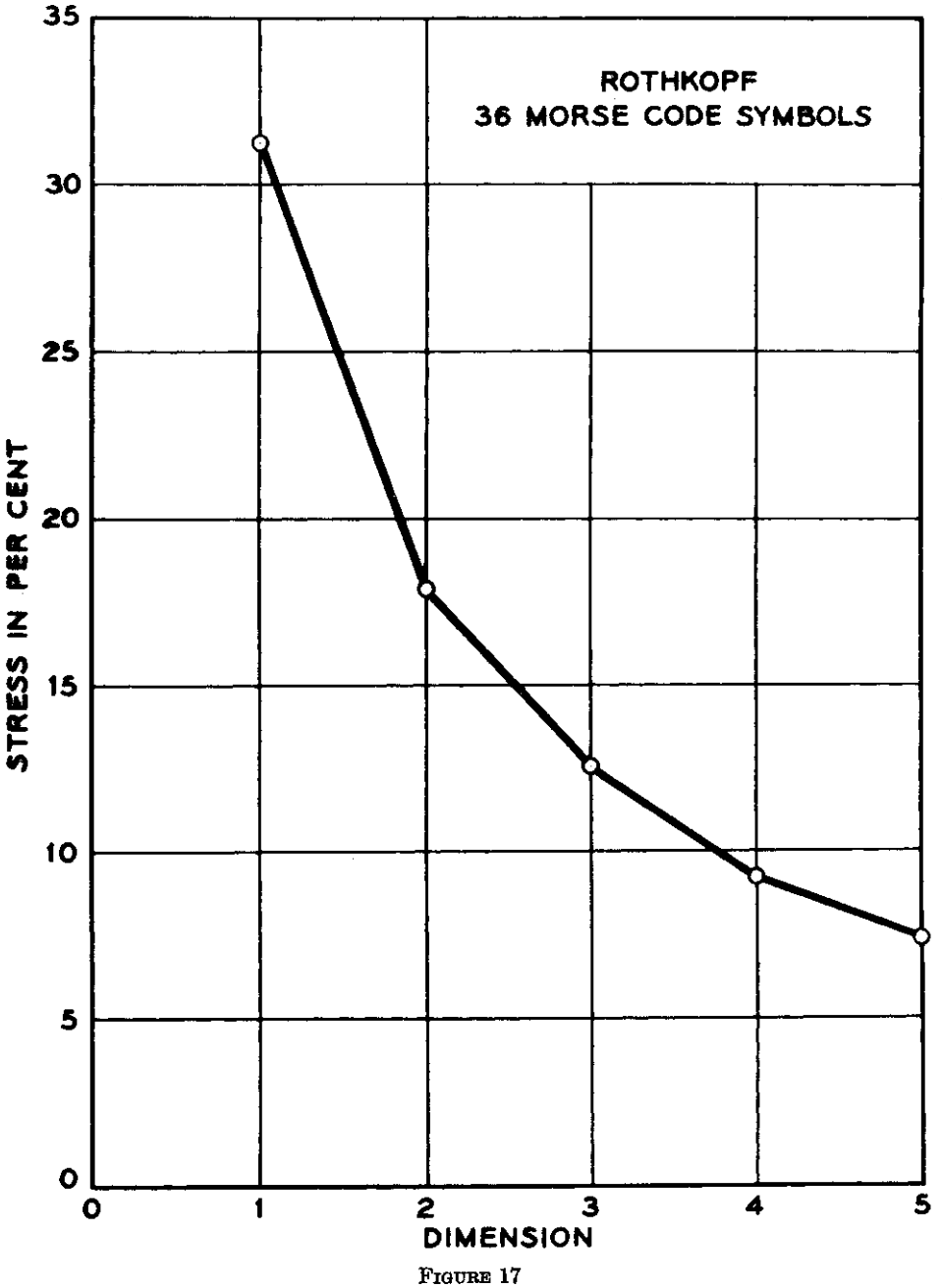

FIGURE 17

Our third illustration is based on the confusions between 36 Morse Code symbols from E. Rothkopf [13]. An analysis of these and other data, using our technique and our computer program, appears in Shepard [16]. We have calculated the stress of the best-fitting configuration in one, two, three, four, and five dimensions (Fig. 17). The figure does not clearly show the number of

dimensions needed, but suggests that two is the minimum and four the maximum. However, Shepard [16] found a very lucid and convincing interpretation for the two-dimensional solution, while he could extract no further structure from the three-dimensional solution. Thus he successfully extracted two coordinates, but expressed some doubt about the value of extracting a third.

Missing Data, Nonsymmetry, and Ties

Suppose some of the dissimilarities are missing, either by error or by design. (When n is large, say $n = 50$ or 60, there are a great many dissimilarities. It may be adequate and economical to obtain data covering only some of them.) How shall we measure stress? It seems natural to generalize the definition given before by simply omitting, both in the numerator S^* and the denominator T^*, the terms which correspond to the missing dissimilarities. We accept this generalization, and incorporate it throughout the rest of the paper.

This idea may be considered simply as a special case of weights being attached to the various dissimilarities to reflect varying uncertainties of measurement. However, we shall not in this paper further pursue this notion of weights, nor certain still more general weighting schemes which come easily to mind.

Suppose that the measurement procedure is not inherently symmetrical, so that $\delta_{ij} \neq \delta_{ji}$. If we are willing to assume that δ_{ij} and δ_{ji} are measurements of the same underlying quantity, and differ only because of statistical fluctuation, then two natural procedures are open to us. One is to form symmetrical measurements by averaging δ_{ij} and δ_{ji}. A more interesting procedure is to generalize the definition of stress by letting the summations for S^* and T^* extend over all $i \neq j$ (rather than $i < j$). Also in some situations the self-dissimilarities δ_{ii} may be meaningful, and one may wish to let the summations include the cases $i = j$.

Suppose there are ties, that is, dissimilarities which by chance are precisely equal to one another. The reader will recall that the numbers $\hat{d}_{ij}$, used in our formula for the stress, were defined as those numbers which minimize S^* subject to the constraint that they are monotonely related to the dissimilarities δ_{ij}. How shall we interpret this constraint in the presence of ties?

There are two approaches. One, which we call the primary approach because it seems preferable, is to say that when $\delta_{ij} = \delta_{kl}$ we do not care which of d_{ij} and d_{kl} is larger nor whether they are equal or not. Consequently we do not wish to downgrade the configuration if $d_{ij} \neq d_{kl}$, and hence do not wish the stress to reflect the inequality. The way we accomplish this is by not constraining $\hat{d}_{ij}$ and $\hat{d}_{kl}$. Consequently the terms $(d_{ij} - \hat{d}_{ij})^2$ and $(d_{kl} - \hat{d}_{kl})^2$ are permitted to be zero, except as prevented by other constraints. Thus in case of the primary approach our only constraints on the $\hat{d}_{ij}$ are these, which are equivalent to (Mon).

(I) Whenever $\delta_{ij} < \delta_{kl}$, then $\hat{d}_{ij} \leqq \hat{d}_{kl}$.

The secondary approach is to say that $\delta_{ij} = \delta_{kl}$ is evidence that d_{ij} ought to equal d_{kl} , and to downgrade a configuration if this is not so. Consequently the stress ought to reflect this inequality. The way we accomplish this is by imposing the constraint $\hat{d}_{ij} = \hat{d}_{kl}$. Then if $d_{ij} \neq d_{kl}$, the terms $(d_{ij} - \hat{d}_{ij})^2$ and $(d_{kl} - \hat{d}_{kl})^2$ cannot be zero and reflect our displeasure at the inequality of d_{ij} and d_{kl} . Thus in the secondary approach to ties, the constraints on the $\hat{d}_{ij}$ are as follows.

(II)
$$\begin{cases} \text{Whenever} \quad \delta_{ij} < \delta_{kl} , & \text{then} \quad \hat{d}_{ij} \leqq \hat{d}_{kl} . \\ \text{Whenever} \quad \delta_{ij} = \delta_{kl} , & \text{then} \quad \hat{d}_{ij} = \hat{d}_{kl} . \end{cases}$$

The place in which the difference between these two approaches actually takes effect is deep inside the algorithm for finding the $\hat{d}_{ij}$. Details are given in the companion paper [12]. We remark that it is very simple to build optional use of both approaches into a computer program, and we have done this.

Non-Euclidean Distance

We plan to discuss elsewhere the full degree to which our procedure may be generalized. In principle, there appears to be no reason why the definition of stress could not be used with almost any kind of distance function at all. However, computing the minimum-stress configuration with more general distance functions may offer difficulties.

For a certain class of non-Euclidean distance functions our procedure is quite practical, and has been fully implemented in our computer program. The numerical techniques we describe below fully cover this generalization.

We refer to distance functions generally known in mathematics as the L_p-norms or l_p-norms, but occasionally referred to as Minkowski r-metrics. For any $r > 1$, define the r-distance between points $x = (x_1 , \cdots , x_t)$ and $y = (y_1 , \cdots , y_t)$ to be

$$d_r(x, y) = \left[\sum_{s=1}^{t} | x_s - y_s |^r \right]^{1/r} .$$

This is just like the ordinary Euclidean formula except that rth power and rth root replace squaring and square root. Then d_r is a genuine distance. In particular, it satisfies the triangle inequality, namely

$$d_r(x, z) \leqq d_r(x, y) + d_r(y, z).$$

[For proof of this fact, see for example Kolmogorov and Fomin ([11], pp. 19–22) or Hardy, Littlewood, and Polya ([8], pp. 30–33).] If $r = 2$, then d_r is ordinary Euclidean distance. If $r = 1$, then d_r is the so-called "city block" or "Manhattan metric" distance.

The Minkowski r-metrics share several properties with ordinary Euclidean distance. In particular, if we displace two points x and y by the same vector z, then the distance between them does not change. In symbols,

$$d_r(x, y) = d_r(x + z, y + z).$$

If we stretch vectors x and y by a scalar factor k, then the distance stretches by a factor k. In symbols,

$$d_r(kx, ky) = kd_r(x, y).$$

However, the Minkowski r-metrics differ sharply from Euclidean distance when rotations are involved. Any rigid rotation leaves Euclidean distances unchanged. The only rigid rotations which leave d_r unchanged in general are those rotations which transform coordinate axes into coordinate axes.

The numerical significance of these properties is brought out in another section. However, we point out here that while a configuration may be freely rotated when Euclidean distances are being used, it may not be when more general distances are used. We do not need to worry explicitly about finding the preferred angular orientation of the configuration, since the iterative minimization process automatically does this for us. However, we must be aware that the coordinate axes have a significance for d_r that they do not have for Euclidean distance.

As an illustration we use experimental data by Ekman [7]. He obtained direct judged similarities of 14 pure spectral colors. We have analyzed his data for several values of r. In every case we obtain the familiar color circle, very similar to the configuration obtained by Shepard [15a], though the precise shape, spacing, and angular orientation varies with r. Fig. 18 shows the stress of the best-fitting configuration as a function of r. We see that a value of 2.5 for r gives the best fit. We do not feel that this demonstrates any significant fact about color vision, though there is the hint that subjective distance between colors may be slightly non-Euclidean. However, it illustrates an approach to non-Euclidean distance that could be of significance in various situations.

Miscellaneous Remarks

The idea of recovering metric information from nonmetric information is not new. A quite different application of this idea, as well as a theoretical discussion, can be found in two papers by Aumann and Kruskal [2, 3]. (See particularly pp. 118–120 in the earlier paper.) Though the situation is not presented there as a psychological one, it does not differ from psychological situations in any essential way. The "subjects," called there "The Board" and consisting of Naval officers, are assumed to make certain comparisons, e.g., which of two simple logistic allocations is superior, as a result of some hypothetical quantitative process of which they are not aware. By using a fairly small number of such comparisons, the experimenter determines with limited uncertainty the numerical values which enter into this quantitative process.

Another very interesting discussion of converting nonmetric information into metric information may be found in Abelson and Tukey [1].

In this paper we assume that there is a true underlying configuration of points in Euclidean t-dimensional space, that we can ascertain only the linear ordering of the interpoint distances, and that we wish from this nonmetric information to recover the configuration. Of course, perfect recovery can at

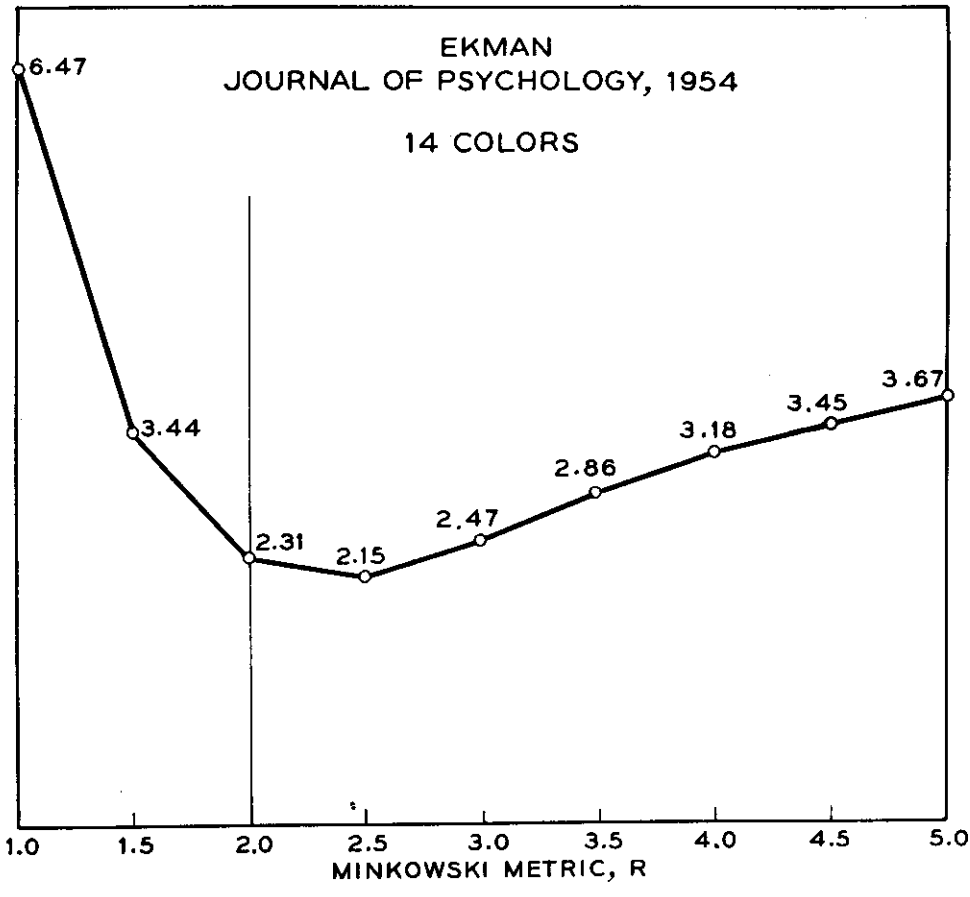

FIGURE 18

best mean construction of a configuration which differs from the original by rigid motions and uniform expansions, for such transformations leave the linear ordering of distances unchanged. Such transformations are called "similarities," and by a known geometrical theorem any transformation in which every distance is multiplied by a fixed constant is a similarity. Thus perfect recovery means construction of a configuration which is geometrically similar to the original.

If the configuration has only a finite number of points, then of course perfect reconstruction is not possible. However, if the number of points is large compared to the number of dimensions, then usually the reconstructed configuration must closely resemble the original. (We note that Shepard was the first to give a practical demonstration that in several dimensions a reasonable number of points are usually tightly constrained.) If the configuration is infinite, perfect recovery may very well be possible. In particular it is possible to prove that if A and B are subsets of Euclidean t-dimensional space (that is, configurations), and if f is a 1-to-1 mapping from A to B which preserves both strict inequality and equality of distances, then f must be a similarity if only

A is big enough. A is big enough if it is all of t-space, or if it is a truly t-dimensional convex subset, or even if it is merely a dense subset of the latter.

It is interesting to compare our technique with Shepard's. His iterative procedure closely resembles ours. Indeed, this whole paper is the outcome of the author's attempt to rationalize Shepard's successful iterative procedure. It is possible to describe his procedure in our terms thus. If d_{ij} is the mth largest distance, define $\hat{\delta}_{ij}$ to be the mth largest dissimilarity; instead of making the influence of x_j on x_i proportional to $d_{ij} - \hat{d}_{ij}$ as we do, he makes it proportional to $\delta_{ij} - \hat{\delta}_{ij}$. It does not appear possible to describe his procedure as one which minimizes some particular measurement of nonmonotonicity.

As far as results go, both procedures yield very similar configurations. Shepard's technique yields smoother-looking curves for dissimilarity versus distance. As actually programmed our procedure is substantially faster than Shepard's, but this probably reflects programming improvements rather than anything more fundamental.

It is interesting to read Bartholomew [4], who is concerned with testing whether parameters are equal, subject to the assumption that they are linearly ordered. (See especially p. 37.) His maximum-likelihood estimate of these parameters bears essentially the same relationship to the observations that our $\hat{d}_{ij}$ bear to d_{ij}. Furthermore, his expression U_k, which plays an important role in his paper and in the likelihood ratio, is essentially the same as our raw stress S^*. In fact it might be possible to interpret our minimum-stress configuration as being a maximum-likelihood estimate in some natural sense.

Summary

To give multidimensional scaling a firm theoretical foundation, we have defined a natural goodness of fit measurement which we call the stress. The stress measures how well any given configuration fits the data. The desired configuration is the one with smallest stress, which we find by methods of numerical analysis. The stress of this best-fitting configuration is a measure of goodness of fit.

Shepard first brought out clearly that what we *should* be looking for in multidimensional scaling is a monotone relation between the experimental data and the distances in the configuration. The stress is no more than a quantitative measurement of how well this holds.

REFERENCES

[1] Abelson, R. P. and Tukey, J. W. Efficient conversion of nonmetric information into metric information. *Proc. Amer. statist. Ass. Meetings, Social statist. Section,* 1959, 226–230.

[2] Aumann, R. J. and Kruskal, J. B. The coefficients in an allocation problem. *Naval Res. Logistics Quart.,* 1958, **5**, 111–123.

[3] Aumann, R. J. and Kruskal, J. B. Assigning quantitative values to qualitative factors in the Naval electronics problem. *Naval Res. Logistics Quart.,* 1959, **6**, 1–16.

[4] Bartholomew, D. J. A test of homogeneity for ordered alternatives. *Biometrika,* 1959, **46**, 36–48.

[5] Coombs, C. H. An application of a nonmetric model for multidimensional analysis of similarities. *Psychol. Rep.*, 1958, **4**, 511–518.
[6] Coombs, C. H. and Kao, R. C. On a connection between factor analysis and multidimensional unfolding. *Psychometrika*, 1960, **25**, 219–231.
[7] Ekman, G. Dimensions of color vision. *J. Psychol.*, 1954, **38**, 467–474.
[8] Hardy, G. H., Littlewood, J. E., and Polya, G. *Inequalities.* (2nd ed.) Cambridge, Eng.: Cambridge Univ. Press, 1952.
[9] Indow, T. and Uchizono, T. Multidimensional mapping of Munsell colors varying in hue and chroma. *J. exp. Psychol.*, 1960, **59**, 321–329.
[10] Indow, T. and Kanazawa, K. Multidimensional mapping of colors varying in hue, chroma and value. *J. exp. Psychol.*, 1960, **59**, 330–336.
[11] Kolmogorov, A. N. and Fomin, S. V. *Elements of the theory of functions and functional analysis.* Vol. 1. *Metric and normed spaces.* Translated from the first (1954) Russian edition by Leo F. Boron. Rochester, N. Y.: Graylock Press, 1957.
[12] Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, (accepted for publication, June, 1964).
[13] Rothkopf, E. Z. A measure of stimulus similarity and errors in some paired-associate learning tasks. *J. exp. Psychol.*, 1957, **53**, 94–101.
[14] Shepard, R. N. Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 1957, **32**, 325–345.
[15] Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika*, 1962, **27**, 125–139, 219–246.
[16] Shepard, R. N. Analysis of proximities as a technique for the study of information processing in man. *Human Factors*, 1963, **5**, 19–34.
[17] Torgerson, W. S. *Theory and methods of scaling.* New York: Wiley, 1958.

Manuscript received 4/11/63
Revised manuscript received 7/16/63

4 Non-metric multidimensional scaling: a numerical method*

Joseph B. Kruskal

1. *Introduction*

We describe a numerical method for multidimensional scaling. In a companion paper [7] we describe the rationale for our approach to scaling, which is related to that of Shepard [9]. As the numerical methods required are largely unfamiliar to psychologists, and even have elements of novelty within the field of numerical analysis, it seems worthwhile to describe them.

In [7] we suppose that there are n objects $1, \cdots, n$, and that we have experimental values δ_{ij} of dissimilarity between them. For a configuration of points $x_1, \cdots, x_n$ in t-dimensional space, with interpoint distances d_{ij}, we defined the *stress* of the configuration by

$$S = \sqrt{\frac{S^*}{T^*}} = \sqrt{\frac{\sum (d_{ij} - \hat{d}_{ij})^2}{\sum d_{ij}^2}},$$

where the values of $\hat{d}_{ij}$ are those numbers which minimize S subject to the constraint that the $\hat{d}_{ij}$ have the same rank order as the δ_{ij}. More precisely, the constraints are that $\hat{d}_{ij} \leqq \hat{d}_{i'j'}$ whenever $\delta_{ij} < \delta_{i'j'}$.

The stress is intended to be a measure of how well the configuration matches the data. More fully, it is supposed that the "true" dissimilarities result from some unknown monotone distortion of the interpoint distances of some "true" configuration, and that the observed dissimilarities differ from the true dissimilarities only because of random fluctuation. The stress is essentially the root-mean-square residual departure from this hypothesis.

By definition, the best-fitting configuration in t-dimensional space, for a fixed value of t, is that configuration which minimizes the stress. The primary computational problem is to find that configuration. A secondary computational problem, of independent interest, is to find the values of $\hat{d}_{ij}$ from the fixed given values of d_{ij}; this is the computational problem of "monotone regression." This latter computation constitutes one step of the main computation.

2. *Missing Entries*

In some cases not all dissimilarities will be observed. Frequently the self-dissimilarities δ_{ii} are either meaningless or unobserved. Sometimes there is no distinction experimentally between δ_{ij} and δ_{ji} so that only a half-matrix of dissimilarities is obtained. Sometimes certain individual dissimilarities may simply fail to be observed. If the number n of objects is large (say 40 or 50), the experimenter may very wisely decide in advance

*reprinted from *Psychometrika*, **29**, 1964, pp. 115–129

to observe only a fraction of the dissimilarities for reasons of cost. Whatever the reason, we adapt to the situation by a very simple change in the definition of the stress: namely, both sums which appear in that definition are restricted to run over those pairs (i, j) for which δ_{ij} is observed. Thus we accommodate missing observations without loss of elegance. Throughout this paper all similar sums will be understood in the same sense unless otherwise indicated.

Of course, if there are not enough dissimilarities observed our method will break down. What this means is that there will be a zero-stress configuration which has no real relationship to the data. One important case in which this occurs is when the objects are split into two groups and the only dissimilarities observed are those between objects in different groups. In this case there is a simple zero-stress configuration in one dimension, namely two distinct points, where each point represents all the objects in one group.

On the other hand, it is not merely a question of how many dissimilarities are observed, but depends on which ones are observed. In many cases of practical importance, one-half or one-quarter of the dissimilarities or fewer are quite sufficient if they are properly distributed in the matrix of all possible dissimilarities.

3. *Non-Euclidean Distance*

Of major interest is the ordinary case in which the distances are Euclidean. If the point x_i has (orthogonal) coordinates $x_{i1}, \cdots, x_{it}$, then the Euclidean (or Pythagorean) distance from x_i to x_j is given by

$$d_{ij} = \left[\sum_{l=1}^{t} (x_{il} - x_{jl})^2 \right]^{1/2}.$$

However, the theory is applicable to much more general distance functions. The numerical methods and formulas given in this paper cover a class of distance functions most often called the L_p or l_p metrics, but occasionally known as Minkowski r-metrics (the term we use). The Minkowski r-metric distance is given by

$$d_{ij} = \left[\sum_{l=1}^{t} |x_{il} - x_{jl}|^r \right]^{1/r}.$$

For $r \geqq 1.0$, this metric is a genuine distance function because it satisfies the triangle inequality. (For proof see ([6], pp. 19–22) or ([4], pp. 30–33).) We restrict ourselves to these cases.

For $r = 2.0$, this metric becomes Euclidean distance. For $r = 1.0$, this metric becomes the so-called "city-block distance" or "Manhattan metric"

$$d_{ij} = \sum_{l=1}^{t} |x_{il} - x_{jl}|.$$

For $r = \infty$, it becomes a familiar metric,

$$d_{ij} = \max_{l} |x_{il} - x_{jl}|,$$

which is sometimes called the l_∞-metric but is widely referred to in colloquial mathematics as the "sup" metric ("sup" is short for supremum).

4. *The Method of Steepest Descent*

Let us now focus on the computational problem that faces us. We restrict our attention to a fixed number of dimensions t and a fixed metric, that is, a fixed value of r. The determination of their values is a matter of judgment, and in many cases can be properly decided only after making analyses with several different values. Thus for the computational question we may assume fixed values of t and r.

An entire configuration can be described as a single vector (or a single point—we use the terms interchangeably) in nt-dimensional space, whose coordinates x_{il} for $i = 1$ to n and $l = 1$ to t are all the coordinates of all the points of the configuration. We refer to nt-dimensional space as "configuration space" and to t-dimensional space as "model space." We emphasize the fact that while a configuration has previously been viewed as n points in model space, we may with equal validity consider it as a single point

$$(x_{11}, \cdots, x_{1t}, \cdots, x_{n1}, \cdots, x_{nt})$$

in configuration space.

Suppose now that the values of δ_{ij} are given. Then for any point in configuration space, that is, for any configuration, there is a definite stress value S. In other words, S is a function

$$S = S(x_{11}, \cdots, x_{1t}, \cdots, x_{n1}, \cdots, x_{nt})$$

defined on the points of configuration space. Our problem is to find that point which minimizes S. Thus we are faced with a standard problem of numerical analysis: to minimize a function of several variables.

For a general review of methods used to solve this problem, see Spang [10], which includes a good bibliography.

To solve this problem we adopt a widely used method of numerical analysis. It is called the "method of steepest descent" or the "method of gradients." We start by picking a more or less arbitrary point in configuration space. In other words, we start with an arbitrary configuration. We then wish to improve the configuration a bit by moving it around slightly. The method of steepest descent calls for this to be done by ascertaining in which direction in configuration space S is decreasing most quickly, and moving a short step in that direction. This direction is called the (negative) gradient and is determined by evaluating the partial derivatives of the function S. In fact,

$$\left(-\frac{\partial S}{\partial x_{11}}, \cdots, -\frac{\partial S}{\partial x_{1t}}, \cdots, -\frac{\partial S}{\partial x_{nt}}\right)$$

is the (negative) gradient. After arriving at a new, slightly better point in configuration space, we again determine the gradient, which is different at different points, and move along it. After many repetitions we arrive at a

point from which no improvement is possible, in other words, at a minimum value of S. This is what we are looking for. We can tell when this happens, for at a minimum all the partial derivatives are zero, that is, the gradient vector is zero.

5. *The Difficulty of Local Minima*

A point in configuration space from which no small movement is an improvement is by definition a local minimum. However, a local minimum may or may not be an overall minimum. Fig. 1 shows a function of one

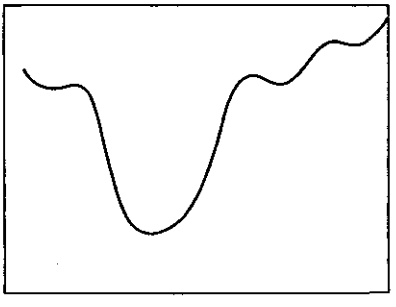

FIGURE 1

variable with four local minima. Only one of them is the overall minimum. If we seek a minimum by the method of steepest descent or by any other method of general use, there is nothing to prevent us from landing at a local minimum other than the true overall minimum. This is a widely known difficulty—in fact, it could be called a standard difficulty of such problems.

In certain important minimization problems the only local minimum is the true overall minimum, so this difficulty does not arise. With this important exception, there are few minimization problems (in numerical analysis) in which the local minimum difficulty can truly be vanquished. At best, we can hope for reasonable confidence that we have the true minimum.

In our minimization problem the difficulty is quite mild. In most cases of interest it need not be a serious concern, for these reasons.

First, we can easily start the method of steepest descent from a variety of different initial configurations. In principle, each initial configuration could lead to a different local minimum. While only the smallest of these could possibly be the true minimum, we would wonder about other still smaller local minima. Usually in our minimization problems, most initial configurations lead to the same local minimum, and this local minimum is much smaller than the few other local minima we find. It seems needless to worry when this is so.

Second, the local minimum configuration which we suppose to be the true overall minimum is not in itself the final end product of the analysis, which must be accepted blindly. In most cases this configuration is of interest

only if it makes sense, only if it can be interpreted or is useful in giving the experimenter insight. If a configuration does this, it is unlikely to be seriously at fault.

Third, unless the stress of the supposed true minimum configuration is sufficiently small, we will not be interested anyhow. A minimum-stress configuration whose stress is above 20% is unlikely to be of interest. Above 15% we must still be cautious; from 10% to 15% we wish it were better; from 5% to 10% is satisfactory; below 5% is impressive.

Fourth, many checks are possible by detailed comparison of the configuration and the data, and by separate analysis of parts of the data.

6. *Numerical Technique*

In principle the iterative technique we use to minimize the stress is not difficult. It requires starting from an arbitrary configuration, computing the (negative) gradient, moving along it a suitable distance, and then repeating the last two steps a sufficient number of times. In this section we discuss some computational aspects which are entirely independent of which computer and which programming language are used.

Since the stress is invariant under translation and uniform stretching and shrinking, we always normalize a configuration by first placing its centroid (center of gravity) at the origin and then by stretching or shrinking so that the root-mean-square distance of the points from the origin equals one. (In our program we have arbitrarily chosen to use Euclidean distance for this purpose, regardless of which Minkowski distance is being used for the interpoint distances.) A configuration which has these properties is said to be normalized.

If ordinary Euclidean distance is used, then the stress is invariant under all rotations, so it becomes possible, and for some purposes desirable, to normalize the angular attitude of the configuration. A natural way to do this is to rotate the configuration so that its so-called principal axes coincide with the coordinate axes (in the natural order). On the other hand, using Minkowski r-metric distance for $r \neq 2$, the only rotations which leave stress invariant are those which transform coordinate axes into coordinate axes. In this case it is possible, and perhaps desirable, to normalize the angular attitude by rotating so that the so-called one-dimensional variance decreases from one coordinate axis to the next. While these normalizations are not difficult, they can easily be left as a separate operation. Therefore we do not discuss them further here.

If a fairly good configuration is conveniently available for use as the starting configuration, it may save quite a few iterations. If not, an arbitrary starting configuration is quite satisfactory. Only two conditions should be met: no two points in the configuration should be the same, and the configuration should not lie in a lower-dimensional subspace than has been chosen for the analysis. If no configuration is conveniently available, an arbitrary configuration must be generated. One satisfactory way to do this is to use the first n points from the list

$$(1, 0, 0, \cdots, 0, 0),$$
$$(0, 1, 0, \cdots, 0, 0),$$
$$\cdots$$
$$(0, 0, 0, \cdots, 0, 1),$$
$$(2, 0, 0, \cdots, 0, 0),$$
$$(0, 2, 0, \cdots, 0, 0), \text{ etc.}$$

Another way would be to generate the points by use of a pseudorandom number generator. In either case the resulting configuration should be normalized.

Suppose we have arrived at the configuration x, consisting of the n points $x_1, \cdots, x_n$ in t dimensions. Let the coordinates of x_i be $x_{i1}, \cdots, x_{it}$. We shall call all the numbers x_{is}, with $i = 1, \cdots, n$ and $s = 1, \cdots, t$, the coordinates of the configuration x. Suppose the (negative) gradient of stress at x is given by g, whose coordinates are g_{is}. Then we form the next configuration by starting from x and moving along g a distance which we call the *step-size* α. In symbols, the new configuration x' is given by

$$x'_{is} = x_{is} + \frac{\alpha}{\text{mag } (g)} g_{is}$$

for all i and s. Here mag (g) means the relative magnitude of g and is given by

$$\text{mag } (g) = \sqrt{\sum_{i,s} g_{is}^2} \Big/ \sqrt{\sum_{i,s} x_{is}^2}.$$

If we assume that x is normalized, then a simpler formula is valid:

$$\text{mag } (g) = \sqrt{\frac{1}{n} \sum_{i,s} g_{is}^2}.$$

We give the formulas for g in another section. Of course, x' should be normalized before further use.

The step-size α is varied from one iteration to the next. The step sizes used do not affect the solution ultimately obtained. However, they profoundly affect the number of iterations required to reach the solution, and are an important computational consideration.

The step-size procedure given here is the result of considerable numerical experimentation. No claim is made that it is optimal in any sense. However, it seems to be reasonably fast, it is robust, and it avoids many pitfalls which we discovered in earlier procedures. It provides large steps during the early stages of calculation and small steps at the end. It is capable of providing a very exact solution when desired.

The initial value of α with an arbitrary starting configuration should be about 0.2. For a configuration that already has low stress, a smaller value should be used. (A poorly chosen value results only in extra iterations.) Thereafter the step size is determined by the following formula.

$$\alpha_{\text{present}} = \alpha_{\text{previous}} \cdot (\text{angle factor}) \cdot (\text{relaxation factor}) \cdot (\text{good luck factor}),$$

where

$$\text{angle factor} = 4.0^{(\cos\theta)^3 \cdot 0},$$

θ = angle between the present gradient and the previous gradient,

$$\text{relaxation factor} = \frac{1.3}{1 + (\text{5-step-ratio})^{5.0}},$$

$$\text{5-step-ratio} = \min\left[1, \left(\frac{\text{present stress}}{\text{stress 5 iterations ago}}\right)\right],$$

$$\text{good luck factor} = \min\left[1, \left(\frac{\text{present stress}}{\text{previous stress}}\right)\right].$$

If five iterations have not yet been calculated, the "stress five iterations ago" may be taken as the first stress computed. Similar artifices may be used in connection with "previous stress" and θ on the very first iteration. If g is the present gradient and g'' the previous gradient, $\cos\theta$ may be calculated by

$$\cos\theta = \frac{\sum_{i,s} g_{is} g''_{is}}{\sqrt{\sum_{i,s} g_{is}^2} \sqrt{\sum_{i,s} g''^2_{is}}}.$$

As the computation proceeds, successively smaller values of stress are achieved. (Occasionally, an iteration may increase the stress rather than decrease it.) Eventually the stress "levels off," and further iterations cause little or no improvement. When is it time to stop and consider the configuration then obtained as sufficiently accurate? There is no really good answer to this question. However, the following rough guide, sensibly used, provides a practical answer.

As the computation proceeds, the relative magnitude mag (g) of the successive gradients decreases. At a configuration which is precisely a minimum, the gradient and its magnitude are zero. Subject to the following qualifications, we suggest that when the magnitude mag (g) reaches a value of approximately 2 per cent of its value for a typical arbitrary configuration, then the iteration may be terminated. This value of mag (g) at which we stop will be called the *local minimum criterion*. For data with large statistical variations, larger values are appropriate, and conversely. For large values of n and t (say $n = 40$, $t = 2$ or $n = 15$, $t = 5$), a larger local minimum criterion is appropriate, and for small values of n and t (say $n = 9$, $t = 2$ or $n = 15$, $t = 1$), a smaller value is appropriate. A larger value of the criterion could mean 5 per cent or conceivably as high as 10 per cent. A smaller value could mean 0.5 per cent or anything down to 0 per cent. If it appears possible and desirable to achieve a stress of zero, then of course the computation should continue until the gradient is zero.

Any iterative minimization procedure is in danger of converging to a local minimum which is not the overall minimum, that is, a solution from

which no *small* change is an improvement and yet which is not the best solution. In our situation, experience shows that this is not a serious difficulty because a solution which appears satisfactory is unlikely to be merely a local minimum. The following simple technique may be used to investigate the situation.

After reaching a solution which we fear is only a local minimum, we apply a violent motion to create a new arbitrary configuration, and we start all over again. We do this repeatedly, and obtain several solutions. We may take the best of these to be the true best solution. A handy way to apply a violent motion is simply to use a very large value of α. It is also reasonable to use the present g (after normalization) as the new x. (In effect this makes α infinite.) If we intend to compute several solutions as described, it is appropriate to relax the local minimum criterion to a fairly large value, and later to continue iterative convergence with a more stringent criterion starting from the best solution.

7. *Programming Technique*

Since the procedure described in this paper is entirely impractical without the aid of an automatic computer, it seems desirable to describe the procedure in sufficient detail that an experienced programmer can easily program it.

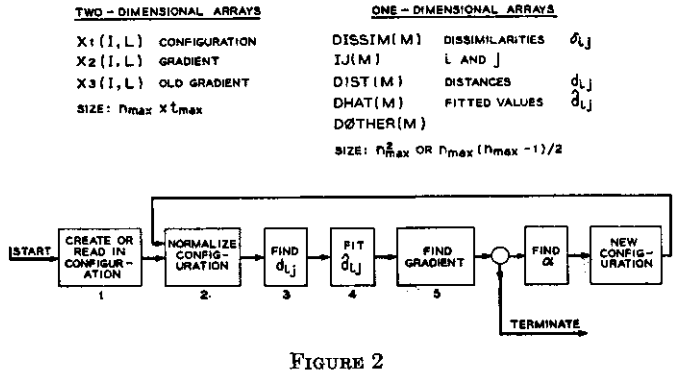

FIGURE 2

A block diagram of the procedure appears in Fig. 2. We start by creating or reading in a configuration. After normalizing the configuration, we calculate the distances d_{ij} . Then we fit the numbers $\hat{d}_{ij}$. (The rank order of the dissimilarities is used only at this stage of the iteration.) From d_{ij} and $\hat{d}_{ij}$, we calculate the stress and the gradient of the present configuration. Then we decide whether we have found a (local) minimum of the stress yet, or whether the normal iterative process should be continued. (It is also desirable to have other termination rules, notably a limit on the number of iterations.) If a local minimum has been reached, then the configuration, the stress, and other useful information should be printed. The configuration should be punched out or saved in some way. Also, printing a history of

the more important variables is desirable. If the calculation is to continue, then the step size is determined, and the new configuration is calculated. This starts a new step of the iteration.

Let n_{max} be the greatest number of objects and t_{max} the greatest number of dimensions which the program is meant to handle. The major blocks of storage needed are two-dimensional arrays X1, X2, X3 and one-dimensional arrays DISSIM, IJ, DIST, DHAT (d-hat), and DØTHER (d-other) as shown in Fig. 2. At the start of an iteration, X1 holds the configuration and X3 holds the old gradient which was used to find it. (Thus X1 (I, L) holds the value x_{il} , and X3 (I, L) holds the previous value of g_{il}.) The gradient at the present configuration is put into X2. Next cos θ is calculated, where θ is the angle between the two gradients. The new configuration is calculated and placed in X1. Then everything is ready for the next step of the iteration.

DISSIM contains the original dissimilarities (or similarities) δ_{ij} , as shown in Fig. 3. Each cell of IJ contains (packed together in one cell) the

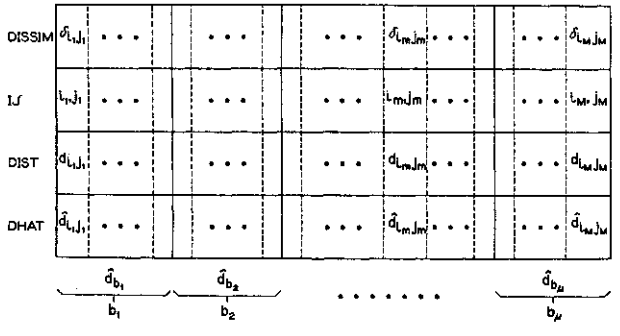

FIGURE 3

values of i and j for the dissimilarity in the corresponding cell of DISSIM. When the dissimilarities are first read, they are placed in DISSIM without any gaps (that is, the cells of DISSIM are filled one by one in order, with no intermediate cells remaining empty). If the dissimilarities are inherently symmetric, or have previously been made symmetric by a separate calculation, then only the entries from one-half the matrix are put into DISSIM. If a dissimilarity is missing (presumably this fact is signalled by a very special artificial value of δ_{ij}), then no entry is made in DISSIM nor is any space reserved.

At the same time that each entry is placed in DISSIM, the corresponding values of i and j are packed together in the corresponding cell of IJ. Thus, although the dissimilarities are put into DISSIM in a manner which ignores their subscripts, this essential information is still present in IJ. Let the number of entries actually placed in DISSIM be M.

After DISSIM and IJ have been filled, then the M entries in DISSIM are sorted in order of increasing algebraic value. (An efficient sorting pro-

cedure should be used, such as the radix-exchange method of Hildebrandt and Isbitz [5].) However, if the measurements in DISSIM are similarities instead of dissimilarities, decreasing order is used. (This is the *only* way in which similarities and dissimilarities are treated differently.) During the sorting procedure the M entries in IJ are simultaneously rearranged so as to preserve the correspondence between cells in corresponding positions of DISSIM and IJ. After the sorting is complete, the values in DISSIM are no longer strictly needed, as their rank order is now available in IJ. However, it is convenient for several reasons to retain the original data.

In cases of ties (equal dissimilarities), the order in which they occur does not matter. However, the IJ cell corresponding to the first dissimilarity of a tie-block (that is, a block of equal dissimilarities) should contain the number of dissimilarities in that tie-block. (This number must be packed in together with i and j.) Also, these cells must be distinguished from other cells, so that the presence of the tie-block can be noted later.

The first stage of each iteration is to find the distances d_{ij} . For Minkowski r-metric distance, use the formula

$$d_{ij} = \left[\sum_{s=1}^{t} |x_{is} - x_{js}|^r \right]^{1/r}.$$

In the case of ordinary Euclidean distances, $r = 2$ and

$$d_{ij} = \sqrt{\sum_{s=1}^{t} (x_{is} - x_{js})^2}.$$

In the case of city-block metric, $r = 1$ and

$$d_{ij} = \sum_{s=1}^{t} |x_{is} - x_{js}|.$$

One very important point concerns the order in which the distances are computed. They should *not* be computed using a double loop on i and j. Instead they should be computed using a single loop in which m runs from 1 to M, corresponding to the entries in IJ. Thus at the mth pass through the loop, the mth entry in IJ is consulted, its values of i and j are used, and the resulting value of d_{ij} is placed in the mth position of DIST.

The next stage of each iteration is to fit the numbers $\hat{d}_{ij}$. We describe how to do this in another section. After fitting, we calculate the stress. It is convenient to set

$$S^* = \sum (d_{ij} - \hat{d}_{ij})^2,$$
$$T^* = \sum d_{ij}^2,$$
$$S = \sqrt{S^*/T^*}.$$

It is best to calculate S^* and T^* by a single loop on m from 1 to M. On the mth pass through the loop we use i and j from the mth entry of IJ. We add to the partially accumulated values of S^* and T^* the quantities $(d_{ij} - \hat{d}_{ij})^2$ and d_{ij}^2 . At the end of the loop, S is calculated from S^* and T^*.

To calculate the (negative) gradient we use the following formulas. For Minkowski r-metric, component g_{kl} , which is to be placed in X2 (K, L),

is given by

$$g_{kl} = S \sum_{i,j} (\delta^{ki} - \delta^{kj}) \left[\frac{d_{ij} - \hat{d}_{ij}}{S^*} - \frac{d_{ij}}{T^*} \right] \frac{|x_{il} - x_{jl}|^{r-1}}{d_{ij}^{r-1}} \text{ signum } (x_{il} - x_{jl}).$$

Here δ^{ki} and δ^{kj} denote the Kronecker symbols ($\delta^{ki} = 1$ if $k = i$, $\delta^{ki} = 0$ if $k \neq i$) and must not be confused with dissimilarities δ_{ij}. Signum is $+1$ for a positive number, -1 for a negative number, and 0 for 0. In case of Euclidean distance $r = 2$ and this becomes

$$g_{kl} = S \sum_{i,j} (\delta^{ki} - \delta^{kj}) \left[\frac{d_{ij} - \hat{d}_{ij}}{S^*} - \frac{d_{ij}}{T^*} \right] \frac{(x_{il} - x_{jl})}{d_{ij}}.$$

In the case of city-block distance, $r = 1$ and it becomes

$$g_{kl} = S \sum_{i,j} (\delta^{ki} - \delta^{kj}) \left[\frac{d_{ij} - \hat{d}_{ij}}{S^*} - \frac{d_{ij}}{T^*} \right] \text{signum } (x_{il} - x_{jl}).$$

To calculate the gradient use a single iterative loop on m from 1 to M. On the mth pass through this loop, use i and j from the mth cell of IJ. If $i = j$, then $\delta^{ki} - \delta^{kj} = 0$ for all k, so the corresponding term in the formula vanishes, and we may skip to the next value of m. If $i \neq j$, then for $l = 1$ to t, add the following term into g_{il} (that is, X2 (I, L)) and subtract it from g_{jl} (that is, X2 (J, L)):

$$\left[\frac{S}{S^*} (d_{ij} - \hat{d}_{ij}) - \frac{S}{T^*} d_{ij} \right] \frac{1}{d_{ij}^{r-1}} |x_{il} - x_{jl}|^{r-1} \text{ signum } (x_{il} - x_{jl}).$$

At the end of the loop, the gradient g has been accumulated in X2.

Once the gradient has been calculated, it is time to decide whether or not a local minimum has been reached. If it has, suitable output is created, and either the calculation terminates, or else it continues after applying a violent motion to create a new arbitrary configuration. If a local minimum has not been reached, the new step size is calculated, the new configuration is calculated and normalized, and the iteration is ready to start over again.

8. *Algorithm for Fitting*

We describe our algorithm for calculating the numbers $\hat{d}_{ij}$. We first describe it supposing that there are no ties (equal dissimilarities). Afterwards we describe the simple modification needed in case ties are present.

Algorithms for essentially the same purpose, though more general because weights are permitted, may be found in Miles ([8], pp. 319–320), Barton and Mallows ([1], pp. 426–427), and Bartholomew ([2], pp. 37–38) and ([3], pp. 242–244). Algorithms and useful facts for the very much more general situation in which the dissimilarities are only partially ordered, not linearly ordered, and for which the function being minimized is much more general then a sum of squares may be found in van Eeden ([11], pp. 134–136) and ([12], pp. 508–512). Our algorithm is essentially the same as algorithm $\mathfrak{A}_1$ of Miles ([8], p. 539). However, we feel for several reasons that it is worthwhile to describe our algorithm. First, Miles' algorithm involves many arbitrary choices, and how these are made affects the efficiency of the com-

putation. Our algorithm is fully explicit; in effect, we make these arbitrary choices in an intelligent manner. Second, none of these algorithms are described in sufficient detail or in simple enough notation to make easy their use on an automatic computer. Third, the efficiency of these algorithms varies very widely. We believe ours to be as efficient as any of those mentioned.

Imagine the dissimilarities $\delta_{i_m j_m}$ arranged from smallest to largest in DISSIM, as in Fig. 3. The subscript pairs (i_m , j_m) are arranged in the same order in IJ. The correct values $\hat{d}_{ij}$ can be described in this way. There is a partition of the dissimilarities into consecutive blocks $b_1 , \cdots , b_u$ such that within each block b the value of $\hat{d}_{ij}$ is constant, and this common value $\hat{d}_b$ is the average of the d_{ij} values in the block. As this is true, it is only necessary to find the correct partition in order to calculate the numbers $\hat{d}_{ij}$.

Our algorithm starts with the finest possible partitions into blocks, and joins the blocks together step by step until the correct partition is found. The finest possible partition consists naturally of M blocks, each containing only a single dissimilarity.

Suppose we have any partition into consecutive blocks. We shall use $\hat{d}_b$ to denote the average of the d_{ij} in block b. If b_- , b, b_+ are three adjacent blocks in ascending order, then we call b *up-satisfied* if $\hat{d}_b < \hat{d}_{b_+}$ and *down-satisfied* if $\hat{d}_{b_-} < \hat{d}_b$. We also call b up-satisfied if it is the highest block, and down-satisfied if it is the lowest block.

At each stage of the algorithm we have a partition into blocks. Furthermore, one of these blocks is *active*. The active block may be *up-active* or *down-active*. At the beginning, the lowest block, consisting of $d_{i_1 j_1}$, is up-active. The algorithm proceeds as follows. If the active block is up-active, check to see whether it is up-satisfied. If it is, the partition remains unchanged but the active block becomes down-active; if not, the active block is joined with the next higher block, thus changing the partition, and the new larger block becomes down-active. On the other hand, if the active block is down-active, do the same thing but upside-down. In other words, check to see whether the down-active block is down-satisfied. If it is, the partition remains unchanged but the active block becomes up-active; if not, the active block is joined with the next lower block into a new block which becomes up-active. Eventually this alternation between up-active and down-active results in an active block which is simultaneously up-satisfied and down-satisfied. When this happens, no further joinings can occur by this procedure, and we transfer activity up to the next higher block, which becomes up-active. The alternation is again performed until a block results which is simultaneously up-satisfied and down-satisfied. Activity is then again transferred to the next higher block, and so forth until the highest block is up-satisfied and down-satisfied. Then the algorithm is finished and the correct partition has been obtained.

After the final partition has been found, then for every block b the value $\hat{d}_b$ is placed in every DHAT cell of b. This completes the fitting computation.

In case there are ties among the dissimilarities, the algorithm for fitting

only needs to be modified by preprocessing. If we adopt the primary approach to ties described in [7], that is, if the only constraints on the $\hat{d}_{ij}$ are those in Section 1, then this preprocessing simply consists of arranging the dissimilarities within each tie-block in such a way that the distances d_{ij} within that block form an increasing sequence. After this preprocessing, the algorithm is carried out as before.

In case we adopt the secondary approach to ties, that is, if we further constrain the $\hat{d}_{ij}$ to be equal when the corresponding δ_{ij} are equal, the preprocessing is still simpler. Instead of starting the algorithm with the finest possible partition, we start it with the partition into tie-blocks. More specifically, the block containing δ_{ij} consists of all dissimilarities which equal δ_{ij}. (In case no other dissimilarities happen to be tied with δ_{ij}, the block contains only one dissimilarity.)

In programming the above algorithm, it is convenient to keep track of the blocks of the partition in the following way. If a block b starts with the mth dissimilarity and contains ν dissimilarities, with $\nu \geqq 2$, then the first DØTHER cell should contain ν and the last DØTHER cell should contain m; also, the first DHAT cell should contain $\hat{d}_b$ and the second DHAT cell should contain $\sum d_{ij}$, where the sum is over all d_{ij} in the block. (Of course,

$$\hat{d}_b = \frac{1}{\nu} \sum d_{ij} ,$$

so we are storing redundant information.) If a block contains only one dissimilarity, then the DØTHER cell should be recognizably blank, and the DHAT cell should contain $\hat{d}_b = d_{ij} = \sum d_{ij}$.

This structure makes it easy to check whether b is up-satisfied or down-satisfied and makes it easy to join two adjacent blocks together. When joining takes place, the joined $\sum d_{ij}$ should be formed by adding the two separate sums, and the new $\hat{d}_b$ formed by dividing by ν. This minimizes round-off error.

If the primary approach to ties is adopted, the sorting of the d_{ij} in each tie-block (during preprocessing) must of course be accompanied by a simultaneous identical rearrangement of the corresponding cells in IJ. If large tie-blocks are anticipated, then the sorting should be done by an efficient procedure such as the radix-exchange technique of Hildebrandt and Isbitz [5].

9. *Summary*

We have described the numerical methods necessary to use our approach to multidimensional scaling. We have included sufficient detail so that an experienced programmer should not have difficulty in creating a program to perform these computations.

REFERENCES

[1] Barton, D. E. and Mallows, C. L. The randomization bases of the amalgamation of weighted means. *J. roy. statist. Soc.*, Series B, 1961, **23**, 423–433.

[2] Bartholomew, D. J. A test of homogeneity for ordered alternatives. *Biometrika*, 1959, **46**, 36–48.
[3] Bartholomew, D. J. A test of homogeneity of means under restricted alternatives (with discussion). *J. roy. statist. Soc.*, Series B, 1961, **23**, 239–281.
[4] Hardy, G. H., Littlewood, J. E., and Polya, G. *Inequalities*. (2nd ed.) Cambridge, Eng.: Cambridge Univ. Press, 1952.
[5] Hildebrandt, P. and Isbitz H. Radix-exchange—An internal sorting method for digital computers. *J. Assoc. computing Machinery*, 1959, **6**, 156–163.
[6] Kolmogorov, A. N. and Fomin, S. V. *Elements of the theory of functions and functional analysis*. Vol. 1. *Metric and normed spaces*. Translated from the first (1954) Russian Edition by Leo F. Boron, Rochester, N. Y., Graylock Press, 1957.
[7] Kruskal, J. Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. *Psychometrika*, 1964, **29**, 1–28.
[8] Miles, R. E. The complete amalgamation into blocks, by weighted means, of a finite set of real numbers. *Biometrika*, 1959, **46**, 317–327.
[9] Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika* (I and II), 1962, **27**, 125–139, 219–246.
[10] Spang, H. A. III. A review of minimization techniques for nonlinear functions. *SIAM Rev.*, 1962, **4**, 343–365.
[11] van Eeden, C. Maximum likelihood estimation of partially or completely ordered parameters, I. *Proc. Akademie van Wetenschappen*, Series A, 1957, **60**, 128–136.
[12] van Eeden, C. Note on two methods for estimating ordered parameters of probability distributions. *Proc. Akademie van Wetenschappen*, Series A, 1957, **60**, 506–512.

Manuscript received 4/11/63
Revised manuscript received 8/22/63

5 Metric Structures in Ordinal Data*

Roger N. Shepard

Prior to the development of procedures of measurement specifically for the behavioral sciences, discussions of scientific measurement (e.g., Cambell, 1920, 1928) tended to focus on the physical sciences and, hence, on scales in which the distance of separation (or spacing) between points on the scale is uniquely determined except for an arbitrary multiplicative constant (viz., the "scale factor" that fixes the size of the unit). Such discussions, that is, were almost exclusively concerned with what Stevens (1951) has classified as "interval" and "ratio" scales.

Subsequently, however, the rather different character of data collected in the behavioral sciences has led psychologists to construct weaker types of scales such as the "ordered metric" type proposed by Coombs (1950). With this type of scale the separations between the points are determined only to within the enormously wider class of order-preserving transformations. Recent investigations have indicated, though, that the difference between an interval and an ordered metric scale, say, may not be as great as at first appears. Accordingly the corresponding distinction between metric and nonmetric information may be susceptible to further clarification.

I. BACKGROUND: ONE-DIMENSIONAL SCALES

Those scales are usually considered nonmetric in which the relations among the interpoint distances are specified by inequalities only (rather than by the strict equalities that are required to fix an interval scale). Simple ordinal scales (Stevens, 1951) as well as the ordered metric and "higher ordered metric" scales (Coombs, 1950, 1964; Fagot, 1959; Siegel, 1956) are all essentially nonmetric according to this definition, since the strongest information that is given about any two values or interpoint separations concerns only which one is the larger. Nothing is explicitly stated about how much larger.

Approximation of Nonmetric Scales to Metric Scales

Actually though, if nonmetric constraints are imposed in sufficient number, they

[1] I am greatly indebted to my colleague J. B. Kruskal for enabling me to carry out extensive Monte Carlo explorations using his improved computer programs for nonmetric scaling. I have also profited from discussions of various aspects of this investigation with him and with J. D. Carroll. The preparation of artificial data and the running and analysis of the computer calculations were carried out by Miss Maureen Sheenan. The paper itself has benefited from the critical comments of several readers including, particularly, C. H. Coombs and R. D. Luce.

*reprinted from the *Journal of Mathematical Psychology*, **3**, 1966, pp. 287–300, 308–311 and 313–315

begin to act like metric constraints. In the case of a purely ordinal scale the nonmetric constraints are relatively few and, consequently, the points on the scale can be moved about quite extensively without violating the inequalities (i.e., without interchanging any two points). As these same points are forced to satisfy more and more inequalities on the interpoint distances as well, however, the spacing tightens up until any but very small perturbations of the points will usually violate one or more of the inequalities.

Abelson and Tukey (1959, 1963) have established this intuitively plausible fact on a more quantitative and rigorous footing. Given any nonmetric scale (i.e., set of inequalities on the coordinates or coordinate differences for the points on a scale), they propose an associated interval scale (i.e., set of explicit metric coordinates for those points) that best represents the nonmetric scale in this "maximin" sense: The smallest "formal" product-moment correlation between the proposed coordinates and any other coordinates that satisfy the same inequalities shall be as large as possible.

They have examined, particularly, the case of four points on a one-dimensional scale. In this case they found that, if only the rank order of the points themselves is known (the ordinal scale), the squared maximin correlation, r^2, is already .65. If, in addition, the rank order of the distances between adjacent points is known (an ordered metric scale), r^2 increases to between .67 and .94 (depending upon the particular ordering given). Finally, if the complete ordering of *all* interpoint distances is known (a higher ordered metric scale), r^2 increases still further to between .91 and .97 (depending, again, on the particular ordering). For many practical purposes, then, a knowledge of the rank order of the interpoint distances may become almost as good as a knowledge of the actual distances themselves.

Frank Goode (1962; Coombs, 1964, pp. 96-102) has investigated a different ("worksheet") method for the conversion from higher ordered metric scales to interval scales. Apparently owing to the strong constraints in this type of nonmetric scale, though, the canonical representation proposed by Goode (his "equal-delta" solution) typically agrees very closely with the maximin r^2 solution of Abelson and Tukey (e.g., see Coombs, 1964, p. 102).

Effect of Number of Scale Points on the Approximation

The results of Abelson and Tukey, particularly, demonstrate that different types of nonmetric scales vary in the extent to which they approximate (or behave like) metric scales. But the question remains, for any one type, as to whether this degree of approximation increases or decreases with the number, n, of points on the scale. Neither Abelson and Tukey nor Goode have reported any systematic examination of higher ordered metric scales consisting of more than four or five points, but Abelson (1959) and Abelson and Tukey (1959) have reported investigations of the dependence of maximin r^2 on n for the ordinal case and for certain simple ordered metric cases.

Whether the over-all effect of the nonmetric constraints should become looser or tighter with an increase in n is not immediately apparent for these particular cases, since in both the ordinal and simple ordered metric cases the number of constraints increases linearly with the number of points.

In fact, however, Abelson and Tukey found that maximin r^2 generally decreased with increasing n. For example, as n increased from 3 to 20, r^2 declined from .75

to .41, for the purely ordinal scale, and from .93 to .69, for the "diminishing returns" ordered metric scale (in which the separation between each successive pair of adjacent points is less than the preceding separation).

These systematic declines may in part have been a consequence peculiar to the maximin criterion adopted by Abelson and Tukey, though. For, each increase in n introduces opportunities for satisfying the given nonmetric constraints by means of more and more bizarre spacings (Abelson and Tukey's "corner sequences"). However, owing to the improbability of such irregular extremes of spacing in nature, their practical significance may be small. For some purposes, then, the maximin r^2 may not be the most useful criterion. Whether a criterion based on some expected deviation (rather than the most extreme possible deviation) would also show this general decline with increasing n has not, apparently, been determined.

There is reason to suppose that an exploration of the higher ordered metric case would lead to conclusions that diverge sharply from those just reviewed for the ordinal and simple ordered metric cases. In particular, since the number of inequalities that must be satisfied increases like n^2 (rather than like n) in the "higher" case, increases in n should lead to a tighter and tighter over-all constraint on a higher ordered metric scale. That is, such a scale should become more and more like a fully metric scale (Shepard, 1962b, p. 239).

Suppes and Winet (1955) and Aumann and Kruskal (1958, p. 118) have reported results, for a particular limiting case of infinite n, that provides further support for this intuitive argument. Specifically, they establish that a complete ordering of the distances between *all* points on a closed interval determines those distances except for a multiplicative scalar. Thus, in the limit of a linear continuum at least, the information in the higher ordered metric scale becomes precisely equivalent to the information in the corresponding interval scale.

Extensions to be Explored

The need is rapidly developing for further explorations of the conditions under which essentially metric information can be extracted from seemingly nonmetric data. In addition to the potential relevance of such explorations for theories of measurement in general, there is also the clarification that they should supply for certain scaling procedures recently developed at the Bell Telephone Laboratories in particular. These latter include a variety of multidimensional scaling which has sometimes been called "analysis of proximities" (Klemmer and Shrimpton, 1963; Kruskal, 1964a, b; Shepard, 1962a and b, 1963, 1964, 1965) and a variety of "nonmetric factor analysis" (Shepard and Kruskal, 1964). Both of these two types of procedures have been found in practice to yield tight, metric solutions even when the input data consisted of nothing but rank orders. As in the case of one-dimensional higher ordered metric scales, moreover, the extent to which this conversion from qualitative to quantitative information could be fully realized seemed to increase with the number of points, n. Accordingly, particular attention will be given, here, to the systematic effects of this variable.

II. THE CASE OF A SINGLE ORDERING OF PAIRS OF OBJECTS: ANALYSIS OF PROXIMITIES

Among the types of nonmetric scales already mentioned, the higher ordered metric type is of particular interest here because it evidently represents the closest approximation to a truly metric scale. There is even reason to suspect that the approximation becomes asymptotically perfect as n is indefinitely increased in an appropriate manner. Moreover, psychological data are readily collected in just the form required for this type of scale. For example, subjects may be presented with all pairs from a set of n stimuli and asked to rank order these $n(n-1)\,2$ pairs with respect to the subjective similarity (or dissimilarity) of the two stimuli in each pair.

Or, even if a numerical rating of similarity is obtained for each pair, these ratings may be meaningful only to within a monotonic transformation (i.e., on an ordinal scale). Thus, an analysis of Ekman's (1954) data on rated similarity of pairs of colors showed that the original ratings were curvilinearly related to interpoint distance in the underlying psychological scale (Shepard, 1962b, p. 237). Since the form of the curvilinear relation was not known in advance of the analysis, only the rank order of the original ratings could be assumed of significance for the scale.

Data of this general form are not of course confined to ratings of subjective similarity. There are also many other types of measures of the psychological closeness of pairs of objects, stimuli, or people—such as frequency of confusion, disjunctive reaction time, strength of association or mutual choice. All of these seem subsumable under the more generic term "proximity relation" (Coombs, 1964) or "proximity measure" (which will be used here).

Analysis of Proximities as a Multidimensional Problem

Clearly, any scale used to represent a set of n objects should have the property that psychological proximity is preserved in the geometry of the scale. That is, if objects A and B are more closely related psychologically than objects C and D, then the corresponding scale-points A and B should be closer together than the points C and D. The central question with which we are here concerned then takes this form: Given a complete set of $n(n-1)/2$ proximity measures for n objects, to what extent is a metric arrangement of the corresponding n points rigidly determined by the requirement that the rank order of the interpoint distances be just the reverse of the rank order of the given proximity measures?

We must notice, however, that some sets of proximity measures cannot be accommodated in this sense by *any* one-dimensional scale. For example, four points A, B, C, D cannot be arranged on a line in such a way that their six interpoint distances have the order $BC < CD < AB < AD < BD < AC$. In order to satisfy these constraints, the points must be arranged in a two-dimensional space. In short, although all ordinal scales and simple ordered metric scales can be one-dimensional, many higher ordered metric scales are necessarily multidimensional.[2] The present investigation is primarily concerned with such multidimensional scales for two reasons. First, they have not previously been studied as thoroughly as unidimensional scales and, second, they are more general in the sense that unidimensional scales can always be subsumed as a special case.

[2] In general, any ordering of all $n(n-1)/2$ distances among n points (including ties) can be realized in a Euclidean space of $n-1$ dimensions (Bennett and Hays, 1960, pp. 37-38; Shepard, 1962a, pp. 129-130).

Limitations of Purely Nonmetric Representations

Before proceeding further, mention should be made of a technique for the multidimensional analysis of proximity measures that has already been developed by W. L. Hays and refined by Coombs and his other associates (see Coombs, 1964, Chs. 21-22). This technique differs from others to be considered here in that it does not attempt to recover a metric configuration. It does start with a (partial) ordering of the interpoint distances, but the result consists of *only* an ordering of the points on the "axes" of the solution.

The question, here, is whether such a purely nonmetric solution preserves all of the significant information in the original nonmetric data. Now if it did, then any particular metric spacing along the axes that might be adopted for the purposes of constructing a spatial picture of the solution should be equally consistent with the original data. However this is not at all the case; for, clearly, the rank order of the interpoint distances changes drastically with alterations in this spacing. Indeed solutions that have been presented spatially in this way (Coombs, 1964, pp. 476, 493) can be shown to violate the original data for just this reason.[3] Coombs himself acknowledges that "the figure does not adequately reflect the metric relations nor the relative lengths of the two dimensions" (p. 476). Still, the fact that inconsistencies arise from an arbitrary "metrization" of nonmetric solution conclusively shows that some of the information in the initial data has been ignored in the process of constructing the solution.

Metric Solution by Iteration on a Computer

As early as 1954 there were indications that the kinds of constraints considered by Abelson, Tukey, and Goode, in one dimension, (but ignored by Hays and Coombs, in more than one dimension) might permit the recovery of essentially metric solutions even in the multidimensional case. In some preliminary explorations, which later led to an application of multidimensional scaling to the study of stimulus generalization (Shepard, 1955, 1958), it became apparent that a small number of points on a plane typically could be moved only slightly without disrupting the rank order of the interpoint distances. Some crude two-dimensional solutions were even obtained by a trial-and-error process of alternately adjusting the positions of small movable markers on a flat surface and measuring the distances between the markers until a satisfactory approximation to the desired ranking was achieved. Entirely independently, Frank Goode used a somewhat similar trial-and-error process (using pencil and worksheet) to show that a particular configuration of nine points in two dimensions could be closely reproduced on the basis, solely, of the rank order of the 36 distances among those points (Goode, 1957).

Nevertheless, metric representations of higher ordered metric scales in more than one dimension did not become practicable on a large scale until the crude method of successively adjusting the positions of markers on a plane was explicitly formalized to the point where it could be implemented on a digital computer. The great speed

[3] In the first of these two spatial solutions presented by Coombs, for example, the distance between the points B and E is considerably greater than the distance between the points A and D. But, in the partial ordering of distances from which this solution was obtained, the distance BE is actually eight levels below (i.e., smaller) than the distance AD (see Coombs, 1964, p. 469).

and storage capacity of the computer permitted solutions with much larger numbers of points than could ever have been attempted by hand. Moreover, by recasting the method into the more abstract language of a computer program, the inherent restriction of physical adjustments to two or, at most, three dimensions could be removed without difficulty. Finally, in the analysis of real data, the errors of measurement (always present in empirical estimates of proximity) could be balanced against each other in a more objective manner.

In the first program of this kind (Shepard, 1962a, b), the coordinates of a trial configuration were iteratively adjusted in a manner designed to bring the rank order of the interpoint distances into closer and closer coincidence with the inverse of the rank order of the given proximity measures. A considerable number of solutions have now been obtained by applying this program to real psychological data as well as to artificial data for which the "true" configuration was known in advance. The configurations to which this program has converged have generally supported the notion that the over-all constraint does indeed increase with the number of points, n. With as many as 15 points, the solutions possessed a degree of metric precision that was, at the time, quite surprising (e.g., see Shepard, 1962b, Fig. 3).

More recently, Joseph Kruskal (1964a, b) has introduced some significant refinements into this general type of method. In place of the originally adopted measure of departure from a perfect inverse ranking (Shepard, 1962a, p. 136), Kruskal proposed the measure

$$\sqrt{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2 \Big/ \sum_{i<j} d_{ij}^2}\,, \tag{1}$$

where d_{ij} is the distance between the trial points i and j during the current iteration and where $\hat{d}_{ij}$ is the corresponding value in that sequence which (a) is monotonically related to the given proximity measures and (b) minimizes the expression (1) for the current d_{ij} values. In words, the measure is simply the square root of a suitably normalized sum of squared deviations from the best-fitting monotonic sequence.

A desirable feature of this measure is that, while it is strictly invariant under any monotonic transformation of the given proximity measures, it does vary continuously with changes in the coordinates of the configuration. Kruskal was thus able to use the negative gradient of this measure to construct an iterative process that seeks its minimum by the method of steepest descent. The advantage of Kruskal's steepest-descent algorithm appears to be primarily theoretical or conceptual, however; the stationary configurations to which his process converges have been found in practice to be essentially indistinguishable from the solutions obtained by the earlier iterative process. Again, this is to be expected for any reasonable method owing to the very tight constraints imposed by the input data as soon as n becomes at all large (cf., Kruskal, 1964a, Fig. 11).[4]

[4] One advantage of the method of steepest descent is that the computational algorithm is directly dictated by the definition of the criterion to be optimized. Creative understanding is therefore needed only for the construction of the criterion—not for the construction of the algorithm. This fact has undoubtedly facilitated the extension of the general type of iterative approach considered here to other cases discussed in the ensuing sections; viz., cases of nonmetric factor analysis (by Kruskal and Shepard), analysis of pair-comparison data (Carroll and Chang, 1964b), and nonmetric analysis of factorial experiments (Kruskal, 1965).

Still, no systematic exploration of the effect of n on the accuracy of solutions obtained by these methods has previously been reported. For this reason a rather extensive exploration of this kind was recently undertaken (using Kruskal's refined algorithm), and the results will be described in a following section. First, however, it is instructive to consider some mathematical arguments concerning what happens in the theoretical limit as the number of points becomes infinite.

Mathematical Consideration of Some Limiting Cases of Large n.

If more and more points are randomly imbedded in some k-dimensional convex region of Euclidean space, this region will gradually become uniformly filled with points. As a helpful heuristic, therefore, let us pass immediately to the limiting case in which all points in this region have been filled in. The question, analogous to the one answered by Aumann and Kruskal (1958) and Suppes and Winet (1955) for one dimension, then becomes this: Does a knowledge of only the rank order of the distances between the points in this region determine the distances themselves (except for an arbitrary scale factor)? Fortunately the transition to more than one dimension entails no qualitative changes. An affirmative answer to the question follows rather directly from already established geometrical results.

Notice, first, that every possible distance that does not exceed the diameter of the region under consideration will occur for some pair of points in that region. Hence any transformation of the set of points that preserves the rank order of the interpoint distances necessarily has the property that, if two distances were initially equal to each other, they must remain equal to each other (although they may of course change *together* during the transformation). Clearly then, since all points on the $(k-1)$-dimensional surface of any sphere are equally distant from the center point, spheres must be preserved under the transformation. Now it is already known that every sphere-preserving transformation is either a similarity transformation or the product of an inversion (in a sphere) and an isometry (Coxeter, 1961, p. 104). The possibility of an inversive transformation can immediately be ruled out, however. It preserves neither the rank order of concentric spheres nor the equality of nonconcentric spheres, whereas both of these invariances are required by the given rank order of the interpoint distances. Hence we can conclude that the given rank order determines the set of points to within a similarity transformation and the distances themselves to within multiplication by an arbitrary scale factor.[5]

The same conclusion can also be reached by considering the obvious fact that the distribution of interpoint distances for any particular convex region approaches a definite limiting form as $n \to \infty$. Hammersley (1950), for example, has shown that the frequency function of distance, d, between points uniformly distributed in a k-dimensional hypersphere is given by

$$f_k(d) = 2^k k d^{k-1} I_{1-d^2}(\tfrac{1}{2}k + \tfrac{1}{2}, \tfrac{1}{2}), \tag{2}$$

[5] In the computer solutions described earlier this scale factor is usually fixed by some convention such as that the mean distance among the points (or, alternatively, the mean square distance of the points from the centroid) be unity. Further conventions can also be imposed, if desired, to fix the position and orientation of the configuration. Sometimes, for example, it is convenient to insure that the centroid is at the origin and that the principal axes of the configuration coincide with the orthogonal axes of the coordinate system.

where the scale factor is chosen so that the largest distance is unity, and where $I_x(p, q)$ is the incomplete beta-function ratio tabulated by Pearson (1934). Provided that the number, n, of points is sufficiently large, then, the inverse of the distribution function obtained by integrating $f_k(d)$ should permit a conversion from the rank of a distance to a number essentially proportional to that distance. Whether such a procedure would be useful in practice, of course, depends upon how insensitive the function $f_k(d)$ is to departures from the assumed uniform spherical distribution of points.

In a recent investigation Benzécri (1964, 1965) examines the case in which the points are assumed to be drawn from a distribution that is spherical but normal (or Gaussian) rather than uniform (as assumed just above). In addition to its naturalness, this assumption has the advantage of permitting the coordinates to be treated as independent random variables. For this case Benzécri establishes rather stronger results than those given above. First, he defines the dimension, k, of an ordering on all $n(n-1)/2$ distances among n points to be the minimum number of dimensions (in Euclidean space) in which that ordering can be realized. He also says of such an ordering that it determines a configuration to within ϵ if the difference between corresponding distances of any two suitably normalized configurations achieving that order (in Euclidean k-space) is less than ϵ. (By a suitable normalization is meant, for example, a rescaling of the configurations so that each has a maximum interpoint distance of unity.)

If, then, the points are assumed to be sampled from a spherically normal distribution, the two theorems of Benzécri can be stated as follows: First, for any specified number of dimensions, k, and any positive quantities δ and ϵ, there exists an n such that, with probability greater than $(1-\delta)$, the order of the distances among n points will be of dimension k and will determine a configuration to within ϵ. Second, for any number of dimensions, k, and any positive δ and ϵ, there exists a number, n_0, such that all distances among n points, where $n > n_0$, are determined to within ϵ by a well-defined function of their order; viz.,

$$d_{ij} \approx F_k\left(\frac{r(d_{ij}) - \frac{1}{2}}{n(n-1)/2}\right), \tag{3}$$

where $r(d_{ij})$ is the (integer) rank of the distance d_{ij} $[1 \leq r(d_{ij}) \leq n(n-1)/2]$, and where the function F_k is the square root of the inverse χ^2 function for k degrees of freedom. This last result depends upon the fact that, when the points are distributed normally (rather than uniformly in a sphere, as considered above), the squared interpoint distances will be distributed in accordance with χ^2 (rather than as indicated in connection with expression (2) above). Proofs for these two theorems can be found in Benzécri's report.

Benzécri also presents some illustrative examples that suggest that the method (based on his second theorem) of converting ranked proximity measures into distance estimates may have some practical utility. Still, the usefulness of the theorems themselves is somewhat limited by the fact that they tell us nothing about what values of ϵ and δ can be attained for any given number of points, n. Moreover, it is known that there exist "degenerate" configurations that can be drastically distorted in certain ways without altering the rank order of the interpoint distances. Perhaps the simplest of these pathological cases is that in which the points are divided into two clusters in such a way that all between-cluster distances exceed all within-cluster distances.

In this case the ratio of the between to the within distances is quite indeterminate and, can, in fact be made arbitrarily large (Shepard, 1962b). Benzécri's results indicate that the probability of such degenerate configurations becomes vanishingly small for sufficiently large numbers of random points. However their prevalence among cases of practical interest remains to be determined.

Monte Carlo Explorations for Intermediate Values of n

In order to gain some more definite, quantitative information about how the constraints tighten up as n increases over the range of practical interest, an empirical investigation was undertaken by means of the computer method described above. Specifically, the accuracies of computer-generated metric reconstructions were evaluated in a number of cases in which these reconstructions were obtained solely on the basis of the rank orders of interpoint distances in two-dimensional configurations of from 3 to 45 random points.

The degree of constraint in each case was assessed by comparing the metric configuration reconstructed by the computer with the original "true" configuration from which the rank order was initially obtained. The particular measure of agreement chosen was the product-moment correlation of the $n(n-1)/2$ interpoint distances in the reconstructed configuration with the corresponding distances in the original "true" configuration. The solutions were secured by applying Kruskal's steepest-descent algorithm to an arbitrary starting configuration (see Kruskal, 1964a, pp. 10-11).

Since the "true" configurations were two-dimensional in every case, a two-dimensional solution could always be found (in principle at least) for which the departure from monotonicity (1) would be zero. However the reduction of this measure to zero does not entail a perfect metric reconstruction. For a finite number of points there generally will exist a roughly polyhedral region in the $n{\cdot}k$-dimensional space of the coordinates within which the desired rank order of the interpoint distances is satisfied. The iterative process, of course, stops as soon as it moves across the boundary into this region. Since the gradient vanishes there, there is no longer a basis for making any further adjustments in the configuration. What we are really interested in here is the size of this region. To the extent that the first solution found within this region is always very close to the true configuration, we can reasonably conclude that the region must be small; i.e., that the configuration is tightly constrained.[6]

Solutions were obtained for 120 different configurations, all constructed from a set of random coordinates used earlier by Coombs and Kao (1960, pp. 222-223). Two-dimensionality of the "true" configurations was insured by using only the first two of the three coordinates presented for each point by Coombs and Kao (viz., their coordinates a and b). Altogether 45 random points were thus made available by renumbering the 30 points in their Table 2 (as points 16 through 45), following the 15 points in their Table 1. Twelve subsets of these 45 points were then selected so as to include, in turn, the first 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, and 45 points. Thus, these 12 subsets were nested in the sense that each succeeding subset contained all the points in all preceding subsets. Another such sequence of 12 nested subsets was then constructed after first "circularly" renumbering the points, so that the points originally numbered 10 through 45 were now numbered 1 through 36 while the points originally

numbered 1 through 9 were now numbered 37 through 45. This process of renumbering and then selecting nested subsets was repeated until five different numberings had been completed (the sixth would have been identical to the first). This entire process was then repeated after renumbering the points in reverse order, so that the points originally numbered 45, 44, ···, 1 were now numbered 1, 2, ···, 45. The final result, then, was 10 different sequences each consisting of 12 nested subsets of points; that is, a total of 120 sets.

Of course the 10 largest configurations all comprised the same 45 points, but the fact that the numbering changed from one configuration to the next insured that the iterative process began with a quite different starting configuration in each of these 10 cases (see Kruskal, 1964a, pp. 10-11). This permits an evaluation of the extent to which the result of the iterative process is influenced by the position from which it starts. If the constraints of the data are sufficiently tight, this influence should be entirely negligible.

The results are summarized in Table 1. For each of the tested numbers of points (n) columns A and B present, respectively, the smallest and the root-mean-square of the ten correlations obtained for that number. Column C gives for each number of points similar "rms" correlations between the "true" and reconstructed distances—but for the first three points only. Thus it is possible to see how the accuracy of reconstructing the triangle formed by the first three points of each nested sequence improves as more and more additional points are included in the total configuration. The close agreement between columns B and C indicates that the first points in each nested sequence were essentially representative of the entire set of random points.

The results presented in Table 1 show that, while the reconstruction of the configuration can occasionally be quite good for a small number of points, it is apt to be rather poor (for n less than eight, say). As n increases, however, the accuracy of the reconstruction systematically improves until even the worst of the ten solutions becomes quite satisfactory with ten points and, for all practical purposes, essentially perfect with 15 or more points. For $n > 15$ the precision of the reconstruction reaches the level where further improvements are of primarily theoretical interest only. Even for these intermediate values of n, then, a purely ordinal scale of proximity entails, with overwhelming probability, an essentially ratio scale of distance.

It is also clear that the number of dimensions is determined as well. Certainly, for

[6] With the gradient method, convergence often becomes extremely slow in the close vicinity of the minimum. In order to keep computing time within reasonable bounds, therefore, a decision was made that, if the measure of departure from monotonicity (1) was not reduced to zero within 50 iterations, a configuration would be accepted as a sufficiently close approximation to the final solution if this measure was reduced at least to the value .01. (This is well below the value, .025, that Kruskal (1964a, p. 3) qualitatively rates as an "excellent" fit.) The effect of accepting the more lenient criterion (.01 rather than zero) is that the measures of agreement between the reconstructed and the true configurations (to be reported in Tables 1 and 2) are biased, if at all, in the direction of *underestimating* the metric precision of the reconstruction. Moreover, failure to reduce the departure from monotonicity strictly to zero occurred in only about 25 % of the cases.

The gradient method can also become very slow or, indeed, even trapped in a local minimum while the measure (1) remains unacceptably high. Whenever this occurred, the process was begun again from a different starting position. This proved to be necessary in only about 15 % of the cases. In the end, a satisfactory solution (with the measure (1) below .01) was achieved in every case within a total of 150 iterations.

TABLE 1

MONTE CARLO RESULTS FOR ANALYSIS OF PROXIMITIES: ACCURACY OF RECONSTRUCTION AS A FUNCTION OF THE NUMBER OF POINTS, n.

n	A min r	B rms r	C rms r (triangle)
3	.753	.935	.935
4	.896	.967	.972
5	.896	.950	.971
6	.951	.981	.986
7	.919	.980	.969
8	.985	.994	.994
9	.988	.996	997
10	.992	.998	.998
15	.999,79	.999,91	.999,96
20	.999,946	.999,979	.999,993
30	.999,950,6	.999,998,3	.999,998,9
45	.999,999,61	.999,999,75	.999,999,94

the larger values of n, the given rank orders could not have been accounted for by any solution in one dimension (cf., Shepard, 1962b, p. 227). But, since a complete account is attainable in two dimensions, there is no need to consider solutions of three or more dimensions.

Section III on nonmetric factor analysis has been omitted.

IV. DISCUSSION AND CONCLUSIONS

The geometrical arguments and Monte Carlo explorations reported here have indicated that, as the number of nonmetric constraints becomes sufficiently large with respect to the number of parameters of the spatial representation, the representation is determined essentially metrically; specifically, to within a similarity transformation (in the case of analysis of proximities) or an affine transformation (in the case of factor analysis). In the first case, an ordinal scale of proximity can be converted into a ratio scale of interpoint distance and, in the second case, several ordinal scales can simultaneously be converted into as many interval scales. The following sections take up some further questions concerning the generality and implications of these results.

INFLUENCE OF ERROR IN REAL DATA

The mathematical and Monte Carlo results have both been confined to artificial cases in which a perfect account of the given ordinal data was attainable in a space of low dimensionality. Owing to the inevitable contribution of error in real psychological data, on the other hand, one must be content with an imperfect solution. The question therefore arises as to whether the above conclusions about the recoverability of metric structures also applies when the data from which we start are not only merely ordinal but fallible as well.

The evident success of numerous applications to real psychological data indicates that this is, in fact, the case (e.g., Klemmer and Shrimpton, 1963; Kruskal, 1964a;

Shepard, 1962b, 1963, 1964, 1965). In the case of analysis of proximities, Kruskal (personal communication) has also invesitgated the robustness of the solution by imposing random deviations on artificially generated data. Generally, a moderately high level of added "noise" can be sustained before the recovered configuration suffers serious deterioration. Again, since the overdetermination of the solution apparently derives from the domination of n^2 over n (or, more strictly, of $n(n-1)$ 2 over $n \cdot k$ for a fixed number of dimensions, k), the invulnerability to error should increase with the number of points, n.

Indeed, the analysis can be regarded, in part, as a technique for purifying noisy data. Since errors often are relatively independent, their effects tend to cancel out in a solution of small dimensionality. Just as a regression line is usually a better predictor of further observations than the individual points upon which it was based, a spatial representation of the type considered here can be both more reliable and more valid than the fallible data from which *it* was derived.

Of course it is always possible to add so much noise that the underlying structure is completely obscured. This situation (which, fortunately, has seldom arisen in practice) can be recognized by the fact that quite different but locally optimum configurations can be found for which the measure of departure from monotonicity has approximately the same (nonzero) value.

Contribution to the Problems of Reduction and Interpretation of Data

An important advantage of a satisfactory solution, when it does prove attainable, is that it furnishes a more parsimonious representation of the original data. In the case of the analysis of proximities, for example, the initially given $n(n \quad 1)/2$ proximity measures can then be reconstructed from just $n \cdot k$ coordinates (where, generally the number of dimensions, k, is much smaller than n). Thus, in a reanalysis of Ekman's (1954) data on the subjective similarities among 14 colors, it was found that the 28 coordinates for the colors in a two-dimensional solution (resembling the conventional "color circle") provided a sufficient basis for the essential reconstruction of all 91 of the original similarity estimates (Shepard, 1962b). Such a complete reconstruction could not, of course, be achieved on the basis of a purely nonmetric solution (like that described by Coombs, 1964, p. 493) which, as was already noted, cannot contain all of the original information.

The reduction to a more concise representation is not, in itself, the only advantage to be gained from such an analysis however. Perhaps even more important is the finding that such a representation sometimes provides an insight into the number and nature of the psychologically important dimensions underlying the empirical data. Thus, a "proximity" analysis of Rothkopf's (1957) data on confusions among all 36 dot-and-dash signals of the Morse code, indicated that the subjects' responses were primarily governed by just two dimensions of the signals; viz., the total number of components (whether dots or dashes), and the ratio of dots to dashes among these components (Shepard, 1963, Fig. 2). Likewise, in a more recent analysis of confusions among 10 vowel phonemes, a three-dimensional solution was obtained in which the three axes were found (after rotation) to agree closely with the frequencies of the first three formants (or resonances) of these sounds as measured physically (Shepard, 1964).

A purely nonmetric solution of the type proposed by Hays can in principle provide information, too, about the underlying dimensions (cf., Coombs, 1964, p. 486).

However, although alternative nonmetric solutions can be obtained for the same data, there seems to be no basis for rotating axes so as to achieve correspondences of the kind just mentioned in connection with the metric solution for the 10 vowel sounds (see Coombs, 1964, p. 494). This, again, is a curious aspect of the nonmetric solutions, since the data themselves are strictly invariant under orthogonal rotations of the underlying configuration (from which those data are presumed to derive). The practical consequence is that, unless the axes of the solution happen to come out with a lucky orientation, nonmetric solutions in three or more dimensions may be relatively difficult to interpret. In the case of metric solutions, on the other hand, objective procedures are already available for rotating axes so as to facilitate interpretation (Carroll and Chang, 1964a; Miller, Shepard, and Chang, 1964).

Whether the method of nonmetric factor analysis considered here will also prove useful in the reduction and interpretation of real psychological data remains to be seen as this new method is applied, too, to empirical data of substantive interest.

Extension to Other Cases

Analysis of proximities and nonmetric factor analysis are not the only cases in which the number of given nonmetric constraints can exceed the number of parameters of the solution. Possibly, therefore, other classes of ordinal data may also be found capable of yielding essentially metric solutions.

Pair-comparison data represent one class that has already been examined from this viewpoint. Klemmer and Shrimpton (1963), for example, modified the method originally proposed for the analysis of proximity measures (Shepard, 1962a) in such a way that it will converge, when possible, upon a one-dimensional preference scale in which the distance between the positions of two objects, A and B, on the scale is monotonically related to the absolute difference between the number of subjects who choose A over B and the number of subjects who choose B over A. This method avoids the strong parametric assumptions of Thurstone's Case V. Yet, when it was applied to Thurstone's own data (Thurstone, 1959), it yielded a scale that correlated .999 with the scale that he had originally obtained by means of a Case V analysis. Moreover, the minute changes that did appear in the new scale evidently were in the direction of permitting a slightly better reconstruction of the original data (Klemmer and Shrimpton, 1963, p. 167).

Carroll and Chang (1964b) have developed a computer program for the analysis of pair-comparison choices that differs from that of Klemmer and Shrimpton in that it is specifically designed to provide solutions in two or more dimensions and, thereby, to account for reliable differences in the preferences of different subjects. The basic model closely resembles that for the variety of nonmetric factor analysis already considered above, except that the initial data consist (for each subject) of the individual pair-comparison choices (e.g., $a = 1$ or -1 for each ordered pair of stimuli) rather than a single rank order of all n stimuli. Correspondingly, the algorithm, instead of minimizing the measure of discrepancy adopted for nonmetric factor analysis (4), minimizes the sum of squared distances between those projections (of individual pairs of points on individual axes) that do not fall in the *pair-wise* order prescribed by the data. Under appropriate conditions such a minimization has been found capable of recovering metric configurations (again, to within an affine transformation). As in the case of nonmetric factor analysis, however, the method seems to be somewhat

susceptible to degeneracy (as, for example, when one stimulus is preferred to all others by all subjects).

A final example concerns a somewhat different type of application. Iterative algorithms resembling some of those mentioned here have recently been employed to transform data matrices so as to achieve an "additive structure" (Goode, 1964; Kruskal, 1965; Morey and Yntema, 1965, Appendix). The objective is to find a monotonic transformation on the entries of the matrix such that each transformed entry can be expressed, as nearly as possible, as a sum of two numbers; one characterizing its row and one characterizing its column. Tversky (1964) has established a necessary and sufficient condition for the achievement of additivity in this sense, and Kruskal (1965) has actually obtained solutions that proved to be essentially unique (to within a linear transformation). Again, however, there are degenerate cases that fail to yield a properly determinate solution (as when all entries in one row or column of the matrix exceed all other entries).

When a determinate solution is achievable, the transformation to additivity provides a very general method for removing undesired interaction effects in the analysis of variance. For "nondegenerate" matrices of sufficient size it can also yield, at the same time, a conversion from a purely ordinal scale of the cell entries to an essentially unique interval scale. Thus another type of application of algorithms of this kind is to the construction of metric scales in accordance with the general scheme of "simultaneous conjoint measurement" proposed by Luce and Tukey (1964).

Incidentally, an important aspect of the scheme proposed by Luce and Tukey, as they point out, is that it replaces "derived measurement," which has been characteristic of the behavioral sciences, with "fundamental measurement," which was previously restricted to the physical sciences (see Suppes and Zinnes, 1963). However all of the methods considered here for extracting metric representations from ordinal data would seem to qualify as fundamental measurement in this sense.

REFERENCES

ABELSON, R. P. Efficient conversion of nonmetric information into metric information. *American Psychologist*, 1959, **14**, 406. (Abstract)

ABELSON, R. P., AND TUKEY, J. W. Efficient conversion of nonmetric information into metric information. *Proceedings American Statistical Association Meetings*, Social Statistics Section, 1959, 226-230.

ABELSON, R. P., AND TUKEY, J. W. Efficient utilization of nonnumerical information in quantitative analysis: General theory and the case of simple order. *Annals Mathematical Statistics*, 1963, **34**, 1347-1369.

AUMANN, R. J., AND KRUSKAL, J. B. The coefficients in an allocation problem. *Naval Research Logistics Quarterly*, 1958, **5**, 111-123.

BENNETT, J. F. Determination of the number of independent parameters of a score matrix from the examination of rank orders. *Psychometrika*, 1956, **21**, 383-393.

BENNETT, J. F., AND HAYS, W. L. Multidimensional unfolding: Determining the dimensionality of ranked preference data. *Psychometrika*, 1960, **25**, 27-43.

BENZÉCRI, J. P. Analyse factorielle des proximités. Publications de l'Institute de Statistique de l'Université de Paris, I. 1964, 13; II. 1965, 14.

BIRKHOFF, G., AND MACLANE, S. *A survey of modern algebra*. (First edition.) New York: MacMillan, 1941.

BRADLEY, R. A., AND TERRY, M. E. Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 1952, **39**, 324-345.

CAMPBELL, N. R. *Physics: The elements*. Vol. 1. Cambridge: Cambridge Univer. Press, 1920.

(Reprinted as *Foundations of Science: The philosophy of theory aad experiment*. New York: Dover, 1957.)

CAMPBELL, N. R. *An account of the principles of measurement and calculation*. London: Longmans, Green, 1928.

CARROLL, J. D., AND CHANG, J.-J. A general index of nonlinear correlation and its application to the interpretation of multidimensional scaling solutions. *American Psychologist*, 1964, **19**, 540. (Abstract) (a)

CARROLL, J. D., AND CHANG, J.-J. Non-parametric multidimensional analysis of paired-comparisons data. Paper presented at the joint meeting of the Psychometric and Psychonomic Societies in Niagara Falls, October 9, 1964. (b)

COOMBS, C. H. Psychological scaling without a unit of measurement. *Psychological Review*, 1950, **57**, 145-158.

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

COOMBS, C. H., AND KAO, R. C. On a connection between factor analysis and multidimensional unfolding. *Psychometrika*, 1960, **25**, 219-231.

COXETER, H. S. M. *Introduction to geometry*. New York: Wiley, 1961.

EKMAN, G. Dimensions of color vision. *Journal of Psychology*, 1954, **38**, 467-474.

FAGOT, R. F. A method for ordered metric scaling by comparison of intervals. *Psychometrika*, 1959, **24**, 157-168.

GOODE, F. M. Numerical analysis methods of scaling. Dittoed notes, University of Michigan, May, 1957.

GOODE, F. M. Interval scale representation of ordered metric scales. Dittoed manuscript, University of Michigan, May, 1962.

GOODE, F. M. An algorithm for the additive conjoint measurement of finite data matrices. *American Psychologist*, 1964, **19**, 579. (Abstract)

GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.

GUTTMAN, L. A basis for scaling qualitative data. *American Sociological Review*, 1944, **9**, 139-150.

HAMMERSLEY, J. M. The distribution of distance in a hypersphere. *Annals Mathematical Statistics*, 1950, **21**, 447-452.

HARMAN, H. H. *Modern factor analysis*. Chicago: Univer. Chicago Press, 1960.

INDOW, T., AND KANAZAWA, K. Multidimensional mapping of Munsell colors varying in hue, chroma, and value. *Journal of Experimental Psychology*, 1960, **59**, 330-336.

KLEIN, F. *Elementary mathematics from the advanced standpoint: Geometry*. New York: Dover, 1939 (paperback edition).

KLEMMER, E. T., AND SHRIMPTON, N. W. Preference scaling via a modification of Shepard's proximity analysis method. *Human Factors*, 1963, **5**, 163-168.

KRUSKAL, J. B. Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika*, 1964, **29**, 1-27. (a)

KRUSKAL, J. B. Non-metric multidimensional scaling: A numerical method. *Psychometrika*, 1964, **29**, 115-129. (b)

KRUSKAL, J. B. Analysis of factorial experiments by estimating monotone transformations. *Journal of the Royal Statistical Society (Series B)*, 1965, **27**, 251-263.

LUCE, R. D. *Individual choice behavior*. New York: Wiley, 1959.

LUCE, R. D., AND TUKEY, J. W. Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1964, **1**, 1-27.

MILLER, J. E., SHEPARD, R. N., AND CHANG, J.-J. An analytical approach to the interpretation of multidimensional scaling solutions. *American Psychologist*, 1964, 579-580. (Abstract)

MOREY, J. L., AND YNTEMA, D. B. Experiments on systems. In J. Spiegel, and D. E. Walker (Eds.) *Second congress on the information system sciences*, Washington, D. C. : Spartan, 1965.

PEARSON, K. *Tables of the incomplete beta-function*. Cambridge: Cambridge Univer. Press, 1934.

ROTHKOPF, E. Z. A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology*, 1957, **53**, 94-101.

SHEPARD, R. N. Stimulus and response generalization during paired-associates learning. Unpublished doctoral dissertation, Yale University, 1955.

SHEPARD, R. N. Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, 1958, **55**, 509-523.

SHEPARD, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 1962, **27**, 125-140. (a)

SHEPARD, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 1962, **27**, 219-246. (b)

SHEPARD, R. N. Analysis of proximities as a technique for the study of information processing in man. *Human Factors*, 1963, **5**, 33-48.

SHEPARD, R. N. Extracting latent structure from behavioral data. In *Proceedings of the 1964 symposium on digital computing*, Bell Telephone Laboratories, May, 1964.

SHEPARD, R. N. Approximations to uniform gradients of generalization by monotone transformations of scale. In D. Mostofsky (Ed.), *Stimulus generalization*, Stanford, Calif.: Stanford Univer. Press, 1965. Pp. 94-110.

SHEPARD, R. N., AND KRUSKAL, J. B. Nonmetric methods for scaling and for factor analysis. *American Psychologist*, 1964, **19**, 557-558. (Abstract)

SIEGEL, S. A method for obtaining an ordered metric scale. *Psychometrica*, 1956, **21**, 207-216.

STEVENS, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951. Pp. 1-49.

STEVENS, S. S. On the psychophysical law. *Psychological Review*, 1957, **64**, 153-181.

SUPPES, P., AND WINET, M. An axiomatization of utility based on the notion of utility differences. *Management Science*, 1955, **1**, 259-270.

SUPPES, P., AND ZINNES, J. L. Basic measurement theory. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of mathematical psychology*. Vol. 1. New York: Wiley, 1963. Pp. 1-76.

THRALL, R. M., AND TORNHEIM, L. *Vector spaces and matrices*. New York: Wiley, 1957.

THURSTONE, L. L. A law of comparative judgment. *Psychological Review*, 1927, **34**, 273-286.

THURSTONE, L. L. *Multiple factor analysis*. Chicago: Univer. Chicago Press, 1947.

THURSTONE, L. L. *The measurement of values*. Chicago: Univer. Chicago Press, 1959.

TORGERSON, W. S. Multidimensional scaling: I. Theory and method. *Psychometrika*, 1952, **17**, 401-419.

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

TVERSKY, A. Finite additive structures. Michigan Mathematical Psychology Program Technical Report MMPP 64-6, August, 1964.

RECEIVED: March 1, 1965

6 A Simple Approximation for Random Ranking STRESS Values*

Ian Spence

Most users of nonmetric multidimensional scaling programs compare the obtained stress values to the results obtained from Monte Carlo studies in which random data have been scaled. This is done in order to be sure that the empirical data set being scaled is not essentially random, but contains some genuine underlying structure. This comparison is extremely helpful in determining whether a scaling solution is meaningful since scaling programs will provide a solution just as happily with noise as with good data. Consequently, experimenters consult the results of Klahr (1969), Wagenaar and Padmos (1971), Stenson and Knoll (1969), or Spence and Ogilvie (1973). The first two sources are rather limited in their practical usefulness since Klahr (1969) contains comparison data for $N \leq 16$, while Wagenaar and Padmos (1971) investigated a maximum of N of only 12, where N is the order of the dissimilarity matrix and is equal to the number of objects being scaled. Therefore, the most useful data are those of Stenson and Knoll (1969) and Spence and Ogilvie (1973); in the former case N ranges from 10 to 60 by steps of 10, and in the latter case ranges from 12 to 48. Stenson and Knoll (1969) only provide a graph of their results, thus necessitating visual interpolation, which can be inaccurate; Spence and Ogilvie (1973) used a combination of regression and graphical interpolation to provide a table which is probably more convenient and accurate from the user's point of view. This table ranges from 12 to 48 by steps of one.

More convenient than either a graph or a table would be a direct and simple analytic expression which would take as arguments N, the number of points, and D, the number of dimensions, and would yield the stress value expected if random data had

*reprinted from *Multivariate Behavioral Research*, 14, 1979, pp. 355–365

***In memoriam*, John C. Ogilvie 1924-1978.**

This research was supported by Grant A8351 from the National Research Council of Canada.

been scaled. Such a function could be evaluated by one or two FORTRAN statements inserted by the authors (or users) of multidimensional scaling programs and output as a matter of course. Then even the inexperienced user would immediately be aware when his data were close to being random. Alternatively, it would be a simple matter to use an electronic minicalculator to compute the function value. In their 1973 paper, Spence and Ogilvie were unable to find a satisfactory analytic approximation; in this paper, by using exploratory data analysis techniques in the spirit of Hoaglin (1977), such a function is provided.

METHOD

The data were obtained from a study by Spence and Ogilvie (1973) and an extended description may be found there. (As will be shown later this set of data is *not* subject to inflation by local minimum problems, and thus represents a suitable basis for the analytic approximation.) Briefly, a nonmetric scaling program was used to scale pseudo random data for $N = 12$, 18, 26, 36, and 48 points, obtaining solutions in $D = 1$, 2, 3, 4, and 5 dimensions. Fifteen replications were obtained and the resulting mean stress values are shown in Table 1. The next step was to determine

Table 1
Monte Carlo Mean Stress Values

Points	Dimensionality 1	2	3	4	5
12	406	222	131	082	051
18	469	291	197	144	107
26	507	327	234	178	144
36	527	352	260	204	168
48	543	370	279	222	187

Note.—Unit of measurement is 0.001

whether an additive structure would adequately describe the entries in this table. This was done by "polishing" the table, after the fashion of Tukey (1977); this merely involves double centering the table by subtracting out the row and column effects (in the analysis of variance sense), leaving the residuals in the body of the table. If the entries in Table 1 can be adequately reconstructed by an expression of the form:

Entry = Grand mean + row effect + column effect

the residuals should be close to zero. As can be seen from Table 2,

Table 2
Residuals and Effects after Polishing

Points	Dimensionality 1	2	3	4	5	Row Effects
12	001	–005	–004	002	005	–086
18	001	001	–000	000	–002	–022
26	002	001	000	–002	–002	014
36	–001	002	001	–000	–001	038
48	–004	001	003	000	–001	056
Column Effects	226	048	–044	–098	–133	Grand Mean = 264

Note.—Unit of measurement is 0.001

this is indeed the case and suggests that stress is an additive function of N and D and can, in principle, be predicted using an equation of the form:

$$[1] \qquad \text{Stress} = a_0 + a_1 f(D) + a_2 g(N)$$

where $f(D)$ is some monotone function relating the number of dimensions to the column effects, and $g(N)$ relates the number of points to the row effects. The a_i constitute a set of appropriate constants.

Figure 1 adds further support to this notion; in it the data are displayed in a two-way plot ("forget-it-plot"). In such a plot (Tukey, 1977) the vertices and line intersections of the lattice coincide with the stress values predicted by the additive relation: Stress = grand mean + row effect + column effect. The residuals are plotted vertically and show the deviation of the actual means from the means predicted by the additive model. In this case, the residuals have been *magnified by a factor of ten* so that they can be seen more clearly. It is obvious that an additive structure describes the data very well.
Also, it can be seen that stress changes less rapidly when N and D become large, suggesting the need for negatively accelerated transformations.

In order to find a suitable $f(D)$ it seems natural to plot the column effects against D. This has been done in Figure 2. An obvious candidate is the logarithmic function, and indeed when a

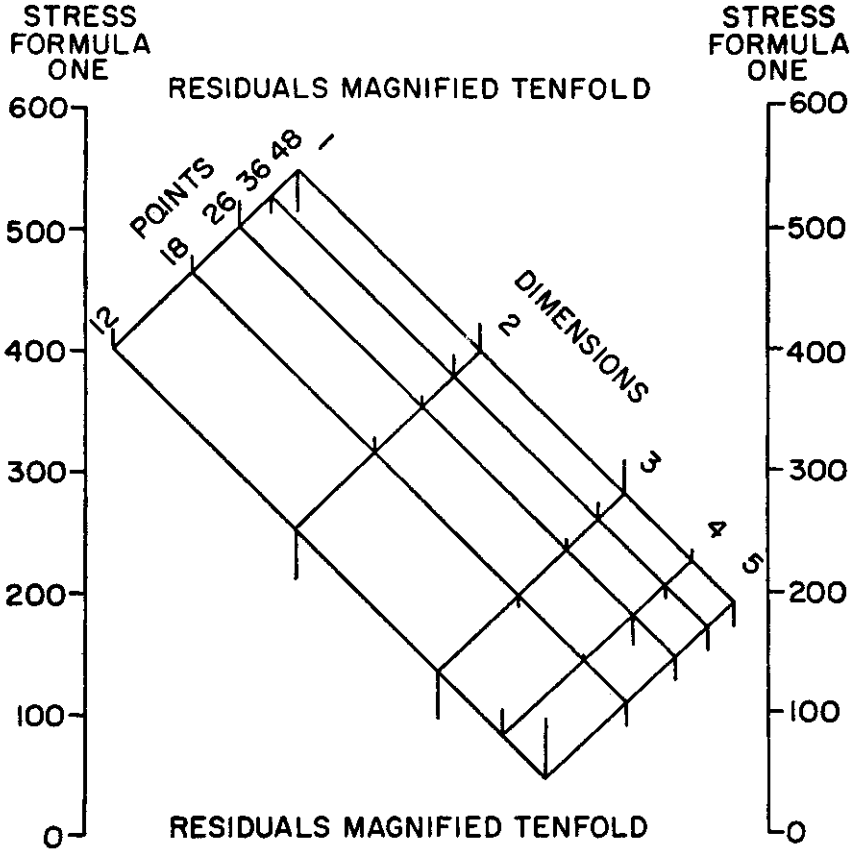

Fig. 1. Tukey two-way plot

log function is fitted by least squares the approximation is very good. A similar plot and regression has been performed to find a suitable approximation for $g(N)$, and the results are shown in Figure 3. Here a log function is not sufficiently extreme and a better fit is obtained by using $(\ln N)^{\frac{1}{2}}$, the square root of the natural logarithm.

Consequently, it seems that the data of Table 1 might be well described by an approximation like:

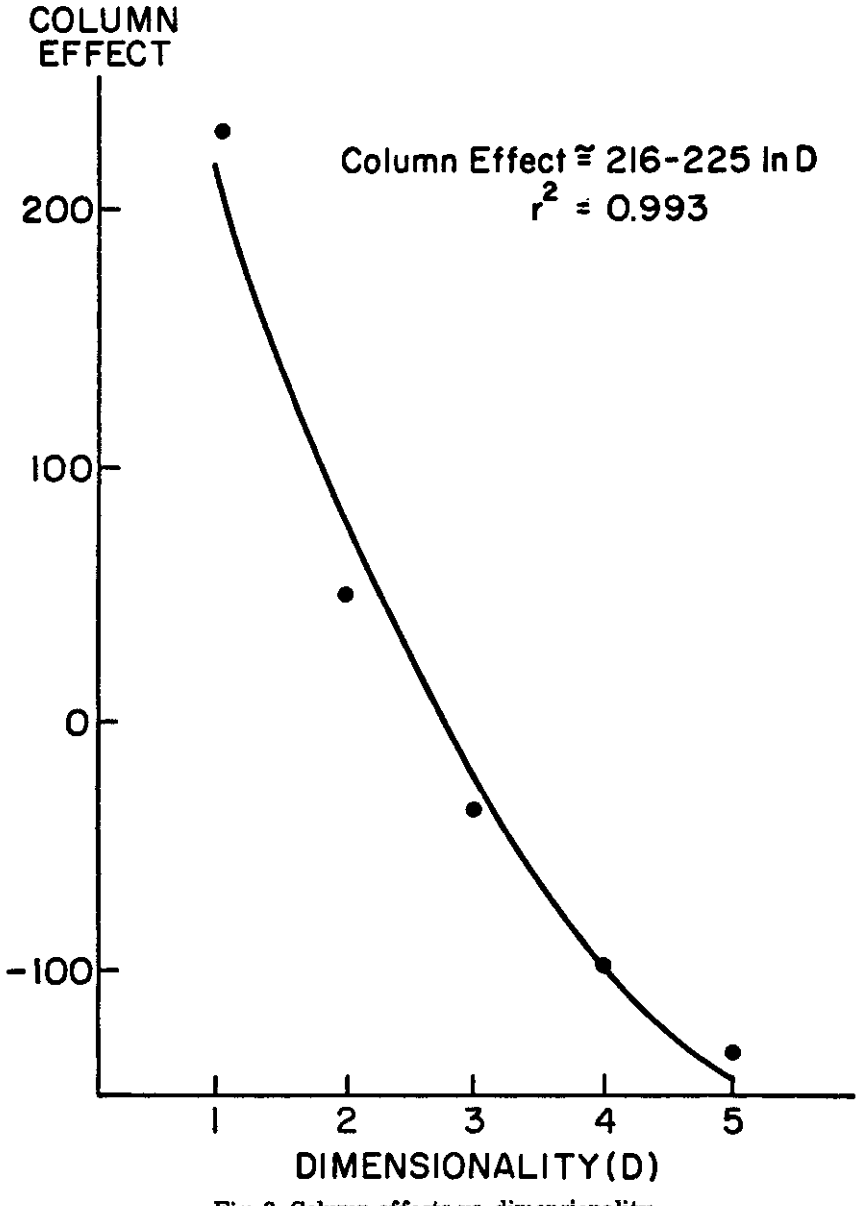

Fig. 2. Column effects vs. dimensionality

[2] $$\text{Stress} = a_0 + a_1 \ln D + a_2 (\ln N)^{1/2}$$

Although this works fairly well, the following equation has been found to provide a little more accuracy:

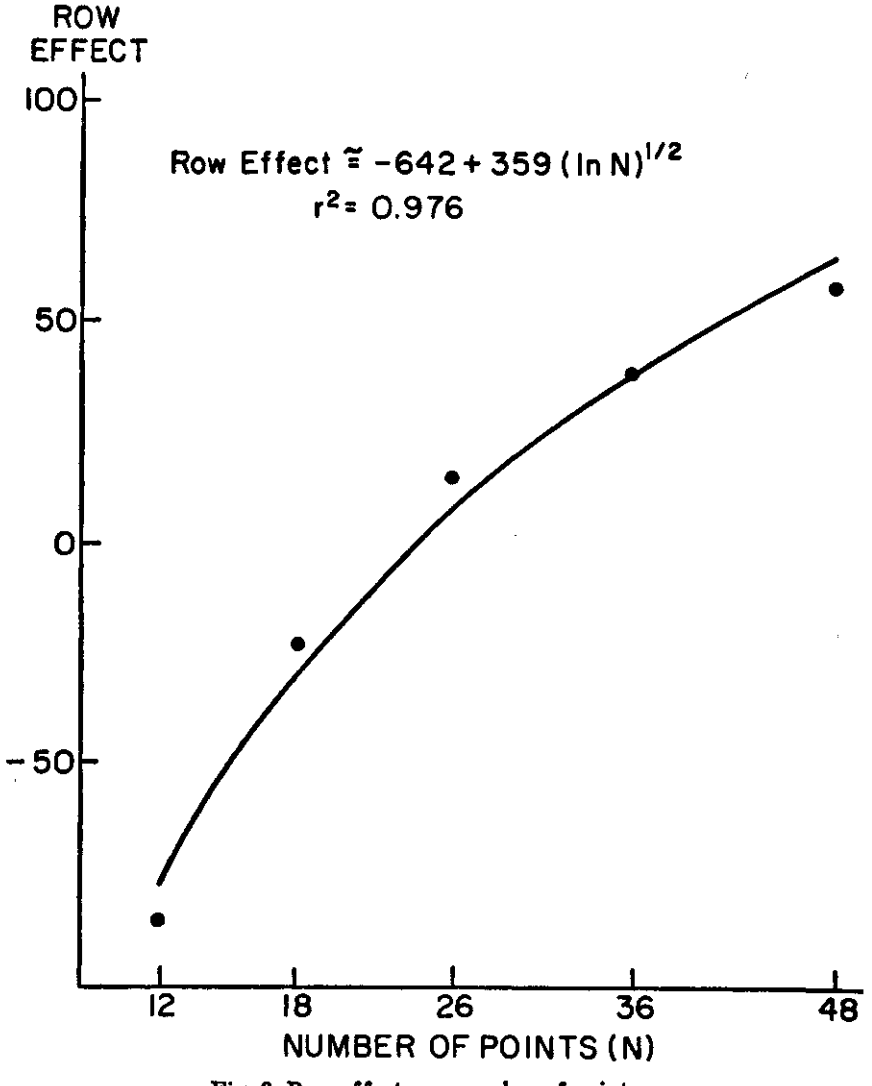

Fig. 3. Row effects vs. number of points

[3] $$\text{Stress} = a_0 + a_1 D + a_2 N + a_3 \ln D + a_4 (\ln N)^{\frac{1}{2}}$$

Consequently, it was decided to include the linear terms. Using ordinary least squares to fit this equation to the original data (5 Levels of Points x 5 Dimensions x 15 Replications = 375 Data values), the following estimates for the regression coefficients were obtained:

$$a_0 = -524.25$$
$$a_1 = 33.80$$
$$a_2 = -2.54$$
$$a_3 = -307.26$$
$$a_4 = 588.35$$

The coefficient of determination R^2 was 0.9963, and the associated regression ANOVA (not shown) yielded a lack of fit Mean Square which was one fifth the size of the replication variance. The approximation clearly fits the original data exceedingly well.

A more compelling way of demonstrating the adequacy of the approximation is to compute expected values for N = 12, 18, 26, 36, and 48 for D = 1, 2, 3, 4, and 5, and then compare the resulting numbers with the mean stress values in Table 1. Table 3 presents the values computed using the approximation, and the discrepancies between the Monte Carlo means and the approximations are listed in brackets. The largest absolute discrepancy is .006 and the mean discrepancy is .003. This compares very favor-

Table 3
Stress Values from the Approximation and Errors of the Approximation

Points	Dimensionality 1	2	3	4	5
12	407	227	137	082	047
	(−001)	(−005)	(−006)	(+000)	(+004)
18	464	285	194	140	105
	(+005)	(+006)	(+003)	(+004)	(+002)
26	505	326	236	181	146
	(+002)	(+001)	(−002)	(−003)	(−002)
36	532	353	262	207	173
	(−005)	(−001)	(−002)	(−003)	(−005)
48	545	366	275	221	186
	(−002)	(+004)	(+004)	(+001)	(+001)

Note.—Unit of measurement is 0.001. Entries in brackets are the differences between the original Monte Carlo means and the approximation values.

ably with the estimated standard error of the means in Table 1, which turns out to be about .002 ($\sqrt{\text{pooled within cell variance}/15}$).

Comparison With Other Results

Some readers may harbor a lingering suspicion that the Monte Carlo data of Spence and Ogilvie (1973) are subject to an unknown degree of distortion caused by local minimum problems and that

Section III
INTERPRETATION

7 Precipitin tests as a basis for a comparative phylogeny*

A. Boyden

In an earlier report[1] the results of a series of precipitin tests on the sera of certain common Mammalia were given. The degree of reaction as indicated by the titer of the ring tests was expressed as percent of the homologous titers. A new method of using these percent values as a basis for a quantitative phylogeny is here proposed and illustrated.

The method involves the calculation of the average values of the reciprocal relationships between pairs of species. These average values (M) constitute the primary data to be used. The values of M together with their probable errors are given in Table I.

TABLE I.

Average reciprocal values of mammalian sera (M) together with their probable errors, and the values of 100-M for all the species tested reciprocally.

Species	M (%)	$P.E._M$	100-M
Dog vs. Horse	4.9	±1.49	95.1
" " Sheep	5.5	±1.37	94.5
" " Pig	6.2	±0.87	93.8
" " Beef	10.5	±1.5	89.5
Beef vs. Horse	9.4	±1.08	90.6
" " Pig	13.2	±0.78	86.8
" " Sheep	69.3	±4.7	30.7
Sheep vs. Horse	3.7	±0.79	96.3
" " Pig	7.7	±0.92	92.3
Pig vs. Horse	5.5	±0.98	94.5

The least reliable value (dog vs. horse) is still 3.3 times its P.E., and hence the whole series is probably significant. To express these quantitative measures of relationship graphically, it is proposed to use the corresponding 100-M values for the actual distances be-

*reprinted from the *Proceedings of the Society for Experimental Biology and Medicine*, **29**, 1931, pp. 995–957

perhaps the alternative device of using the minimum stress solution from a large number of random starts should have been used (Arabie, 1978). A recent study by Null and Young (1978) shows clearly that this is not the case; some of their results are summarized in Table 4. Using identical random data sets, Null and Young used four alternative starting strategies with the same nonmetric algorithm (KYST): the Kruskal L-shaped configuration, a single random start, the *best* of 22 random starts (in terms of stress), and the TORSCA initial configuration routine. Five replications were obtained. Not surprisingly, the Kruskal L-start and the single random start performed poorly (*c.f.* Spence, 1972). The best of 22 random starts did much better, but it is important to note that this strategy did not systematically outperform the *single* TORSCA start (The TORSCA start led to a lower final stress on 13 out of 24 occasions, although the absolute numerical differences are very small). Thus, for all practical purposes, there is no real difference between using the best of 22 random starts, and a single TORSCA start except, of course, that it will cost you about 20 times as much in computer time. Using a single random start or a Kruskal L-start, is not a sensible strategy.

It is clear that the present approximation yields values which are very close to the independently obtained *minimum* values of Null and Young (1968). Indeed, even including the 10 point results, where there may easily be underdetermined solutions, the average discrepancy between the approximation values and the *smallest* of the values from 24 different starts is only .006. This is all the more remarkable when it is realized that the Spence and Ogilvie study did not collect *any* Monte Carlo data for 10, 15, 20, 25, or 30 points!

The discrepancies between the present approximation and Table 2 of Spence and Ogilvie (1973), which it is intended to replace, are also very small. This may be inferred from the values shown in the final column of Table 4.

Although it is not possible to make precise numerical comparisons with the data of Stenson and Knoll (1969), since their results were presented in graphical rather than tabular form, careful graphical interpolation shows that there is a very close correspondence between their results and the present approximation. This suggests that the approximation may be used with confidence in the range $10 \leq N \leq 60$ and $1 \leq D \leq 5$, even although it was developed for $12 \leq N \leq 48$.

Table 4
A Comparison of the Approximation with Various Other Results

No. of Points	No. of Dimensions	Kruskal L-Start[a]	Single Random Start[a]	Min. of 22 Random Starts[a]	TORSCA Start[a]	Present Approximation	S & O (1973)[b]
10	1	498	432	355	362	377	—
	2	226	214	195	206	198	—
	3	114	138	083	094	107	—
	4	066	066	043	054	052	—
15	1	524	511	456	447	440	445
	2	280	296	264	260	260	265
	3	187	194	162	164	170	175
	4	127	115	107	110	115	121
	5	085	088	080	083	080	084
20	1	529	517	503	478	477	480
	2	301	314	299	297	298	300
	3	224	225	201	209	207	210
	4	154	158	148	147	153	156
	5	119	123	112	115	118	120
25	1	539	545	523	500	502	501
	2	341	349	324	320	322	321
	3	243	240	228	227	232	231
	4	178	194	172	173	177	177
	5	147	151	137	139	142	141
30	1	556	550	542	517	518	516
	2	360	358	349	341	339	336
	3	260	270	248	246	248	246
	4	195	203	189	188	194	192
	5	157	169	155	151	159	156

[a]From Table 1, Null & Young (1978): KYST program used.
[b]From Table 2, Spence & Ogilvie (1973): TORSCA-9 program used.

Discussion

Since the approximation yields values which are indistinguishable (within the limits of sampling error) from the actual values obtained from using a TORSCA start or a *multiple* random start, it may safely be used with the TORSCA-9 program, or the KYST program using either a TORSCA start, or the best result from about 20 random starts. Further, since Spence (1972) has shown that there is little to choose between the SSA-I and TORSCA-9 algorithms, the present approximation can probably be used with SSA-I, providing the option to minimize stress in the final stage is selected by the user. Similarly, given the strong resemblance of MINISSA-I to SSA-I, the approximation can probably be used with confidence with that program.

Regarding the use of random rankings stress values, it is the opinion of the present author that they should not be used in a rigid classical hypothesis testing fashion (*c.f.* Spence & Ogilvie, 1973, p. 516; Spence, 1978, p. 214). The greatest benefit to be gained from a comparison of empirical stress values with random rankings stress values is that the investigator obtains a good intuitive feeling for the worth of the data. If the obtained values are well below the random values, say only a third or a half as large, then one can be fairly sure that the data are good. On the other hand, if the obtained values are rather close to the random values, then one should be very careful even though the hypothesis of randomness may be rejected in some technical sense. (For those who may wish to construct formal tests, however, the standard deviation associated with values produced by the approximation can be taken as .01—the square root of the replication variance.)

Finally, it is noted that the approximation has been incorporated into the M-SPACE program (Spence & Graef, 1974), replacing the stored table of random rankings stress values which was previously employed.

Reference Note

1. Null, C. H. and Young, F. W. A Monte Carlo investigation of initial configurations strategies in KYST. Presented at the European Meeting on Psychometrics and Mathematical Psychology, University of Uppsala, Uppsala, Sweden, June 15-17, 1978.

REFERENCES

Arabie, P. Random versus rational strategies for initial configurations in nonmetric multidimensional scaling. *Psychometrika*, 1978, *43*, 111-113.

Hoaglin, D. C. Direct approximations for chi-squared percentage points. *Journal of the American Statistical Association*, 1977, *72*, 508-515.

Klahr, D. A Monte Carlo investigation of the statistical significance of Kruskal's nonmetric scaling procedure. *Psychometrika*, 1969, *34*, 319-330.

Spence, I. A Monte Carlo evaluation of three nonmetric multidimensional scaling algorithms. *Psychometrika*, 1972, *37*, 461-486.

Spence, I. On random rankings studies in nonmetric scaling. *Psychometrika*, 1974, *39*, 267-268.

Spence, I. Multidimensional scaling. In P. W. Colgan (Ed.), *Quantitative Ethology*. New York: Wiley, 1978.

Spence, I. and Graef, J. The determination of the underlying dimensionality of an empirically obtained matrix of proximities. *Multivariate Behavioral Research*, 1974, *9*, 331-341.

Spence, I. and Ogilvie, J. C. A table of expected stress values for random rankings in nonmetric multidimensional scaling. *Multivariate Behavioral Research*, 1973, *8*, 511-517.

Spence, I. and Young, F. W. Monte Carlo studies in nonmetric scaling. *Psychometrika*, 1978, *43*, 115-117.

Stenson, H. H. and Knoll, R. L. Goodness of fit for random rankings in Kruskal's nonmetric scaling procedure. *Psychological Bulletin*, 1969, *72*, 122-126.

Tukey, J. W. *Exploratory data analysis*. Reading, Massachusetts: Addison-Wesley, 1977.

Wagenaar, W. A. and Padmos, P. Quantitative interpretation of stress in Kruskal's multidimensional scaling technique. *British Journal of Mathematical and Statistical Psychology*, 1971, *24*, 101-110.