PSYCHOMETRIKA—VOL. 29, NO. 1
MARCH, 1964

MULTIDIMENSIONAL SCALING BY OPTIMIZING GOODNESS OF FIT TO A NONMETRIC HYPOTHESIS

J. B. KRUSKAL

BELL TELEPHONE LABORATORIES
MURRAY HILL, N. J.

Multidimensional scaling is the problem of representing n objects geometrically by n points, so that the interpoint distances correspond in some sense to experimental dissimilarities between objects. In just what sense distances and dissimilarities should correspond has been left rather vague in most approaches, thus leaving these approaches logically incomplete. Our fundamental hypothesis is that dissimilarities and distances are monotonically related. We define a quantitative, intuitively satisfying measure of goodness of fit to this hypothesis. Our technique of multidimensional scaling is to compute that configuration of points which optimizes the goodness of fit. A practical computer program for doing the calculations is described in a companion paper.

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 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 computation.) 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
2½%	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}.

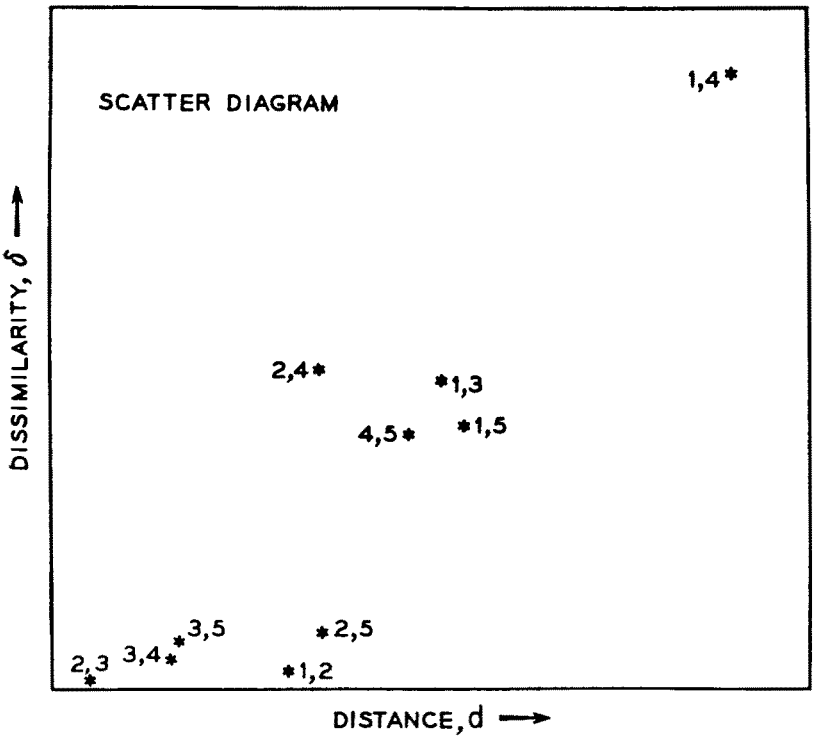

FIGURE 1

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.)

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

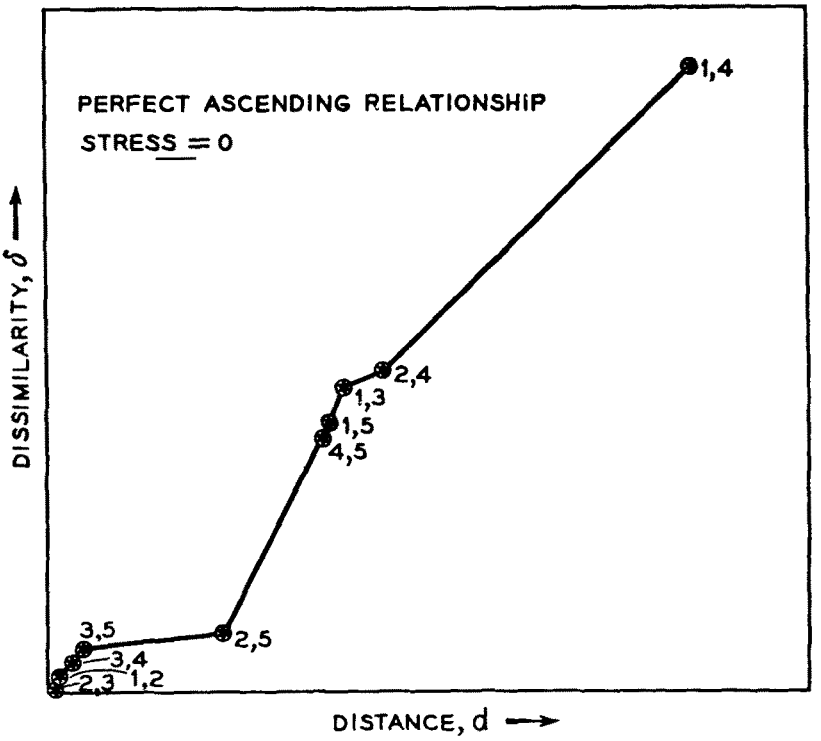

FIGURE 2

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

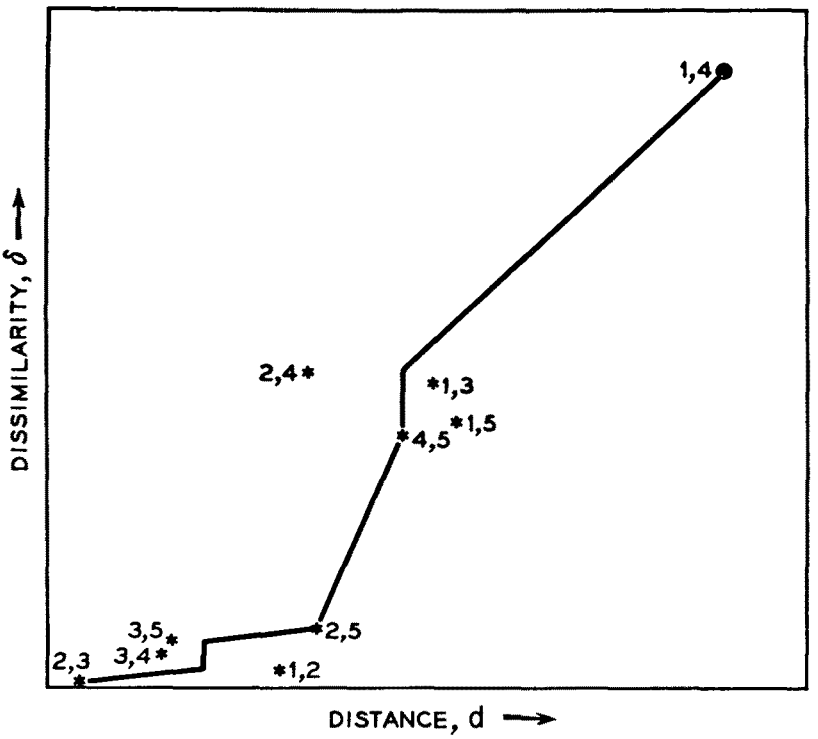

FIGURE 3

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

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

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

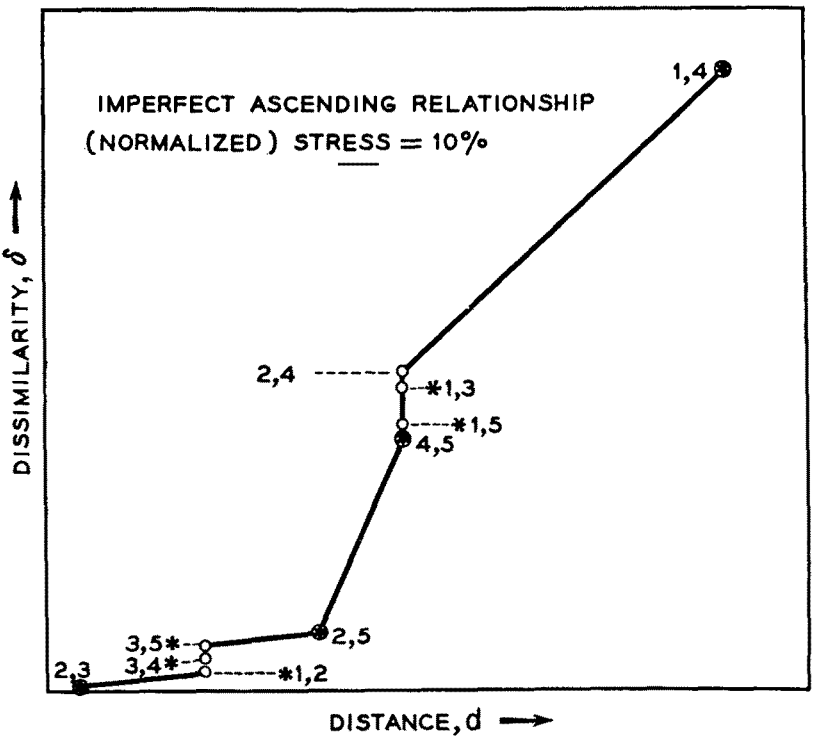

FIGURE 4

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.

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 $(d_{ii}\ ,\ \delta_{ii})$, then we denote the corresponding point by $(\hat{d}_{ii}\ ,\ \delta_{ii})$. Thus fitting the curve means no more than fitting the values of $\hat{d}_{ii}$.

We realize of course that the numbers $\hat{d}_{ii}$ are not distances. There is no configuration whose interpoint distances are $\hat{d}_{ii}$. The $\hat{d}_{ii}$ are merely a monotone sequence of numbers, chosen as "nearly equal" to the d_{ii} 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

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

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-

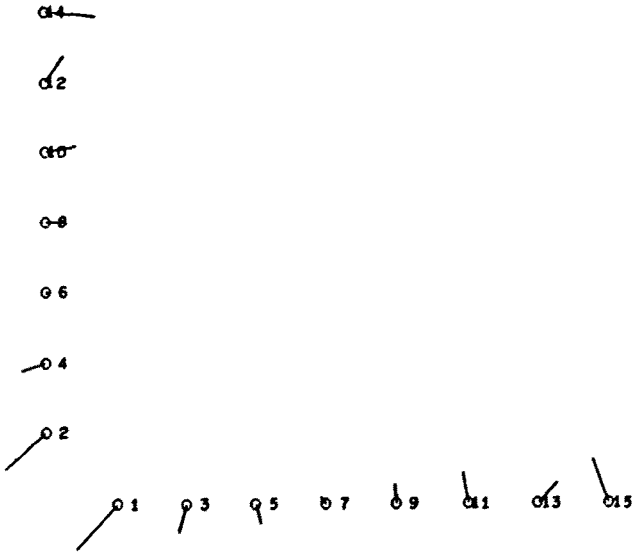

FIGURE 5
Initial Configuration (Coombs and Kao Data)

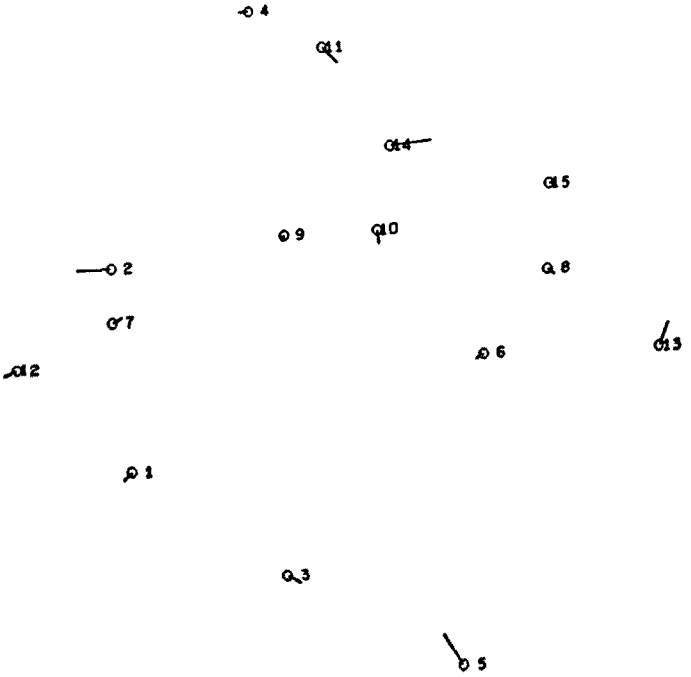

FIGURE 6
Configuration After 10 Iterations (Coombs and Kao Data)

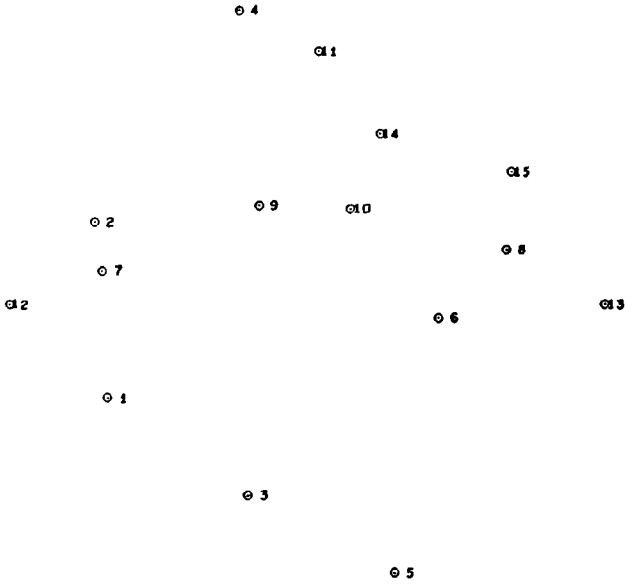

FIGURE 7
Configuration After 50 Iterations (Coombs and Kao Data)

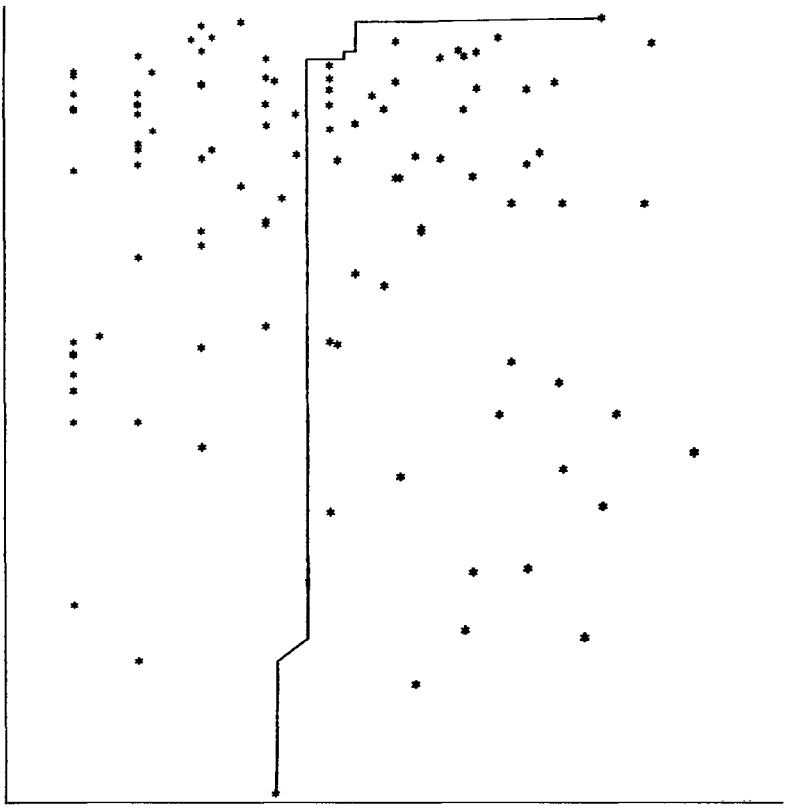

FIGURE 8
Initial Scatter Diagram (Coombs and Kao Data)

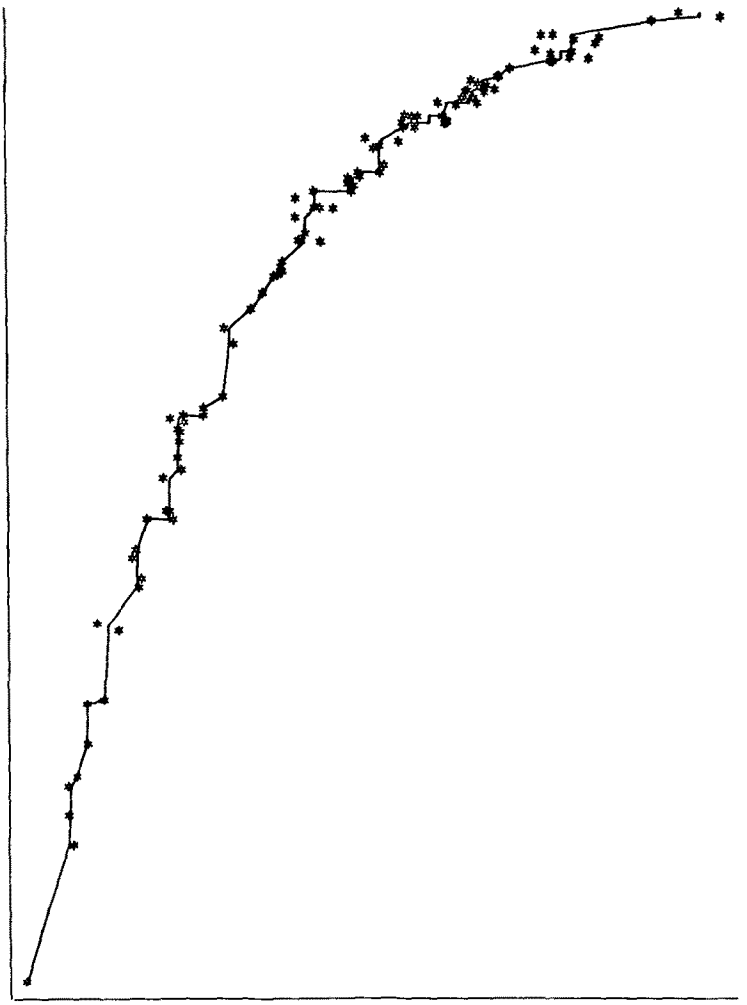

FIGURE 9

Scatter Diagram After 10 Iterations (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.

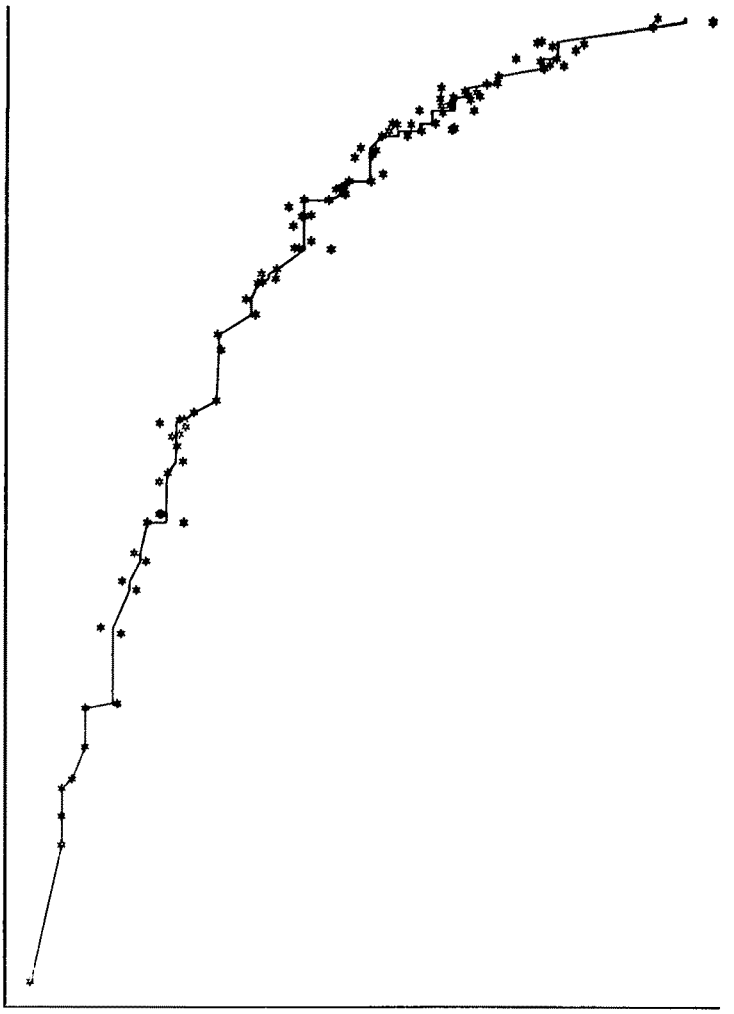

FIGURE 10

Scatter Diagram After 50 Iterations (Coombs and Kao Data)

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

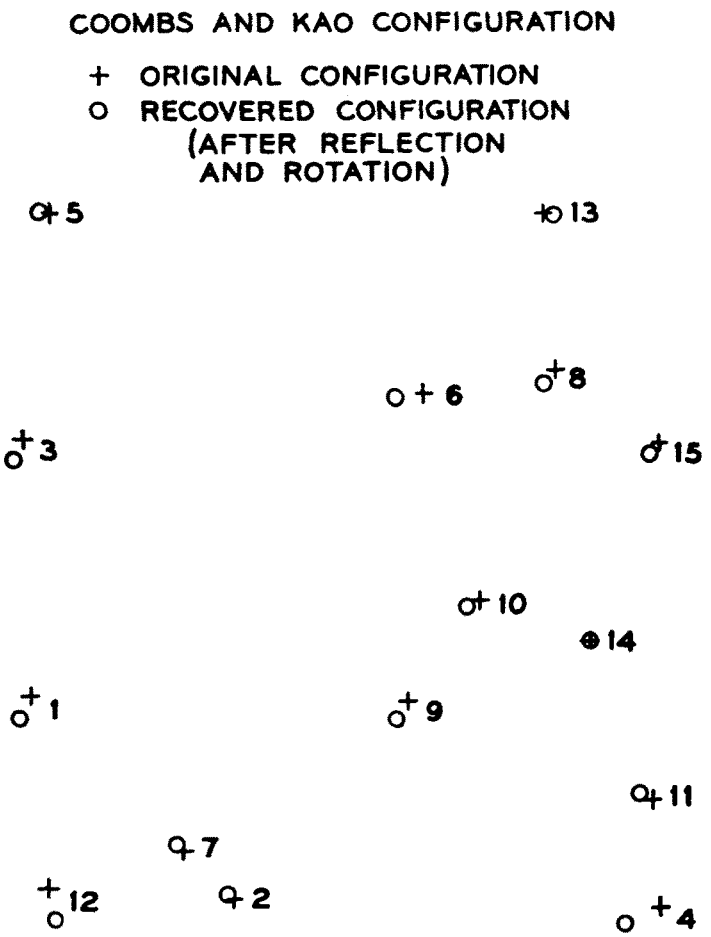

FIGURE 11

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

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.

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 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 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 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

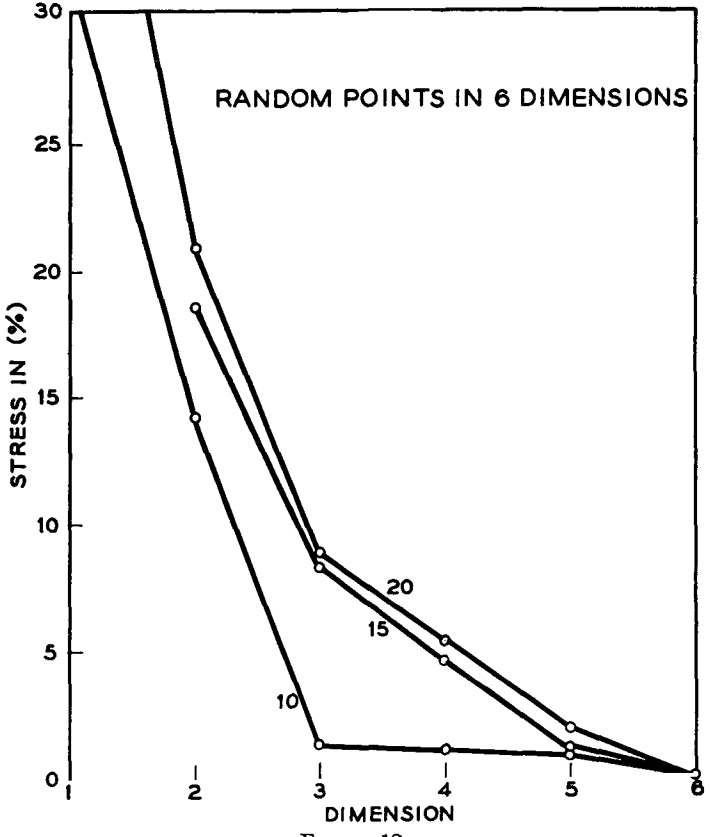

FIGURE 12

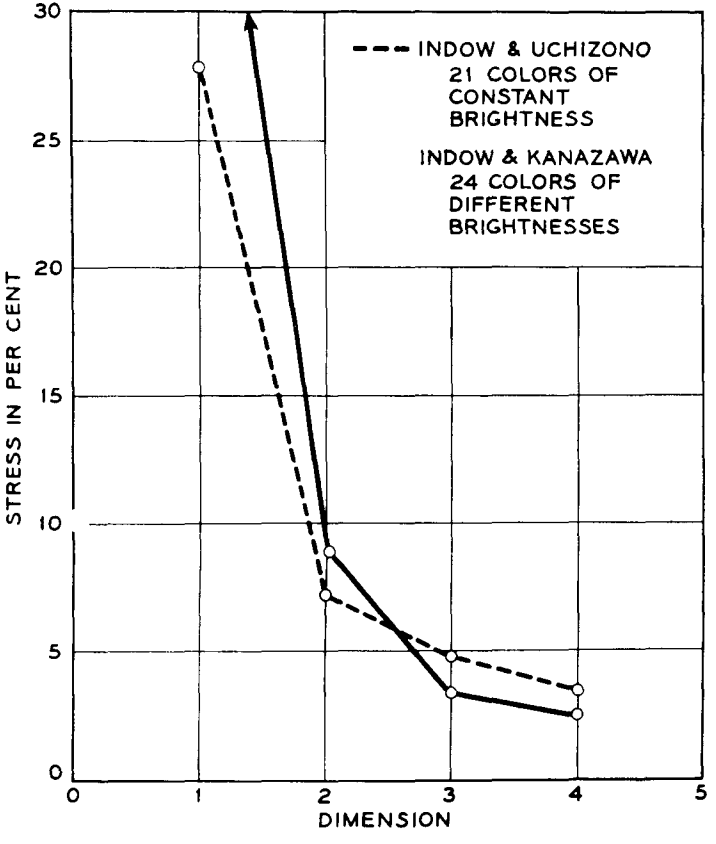

FIGURE 13

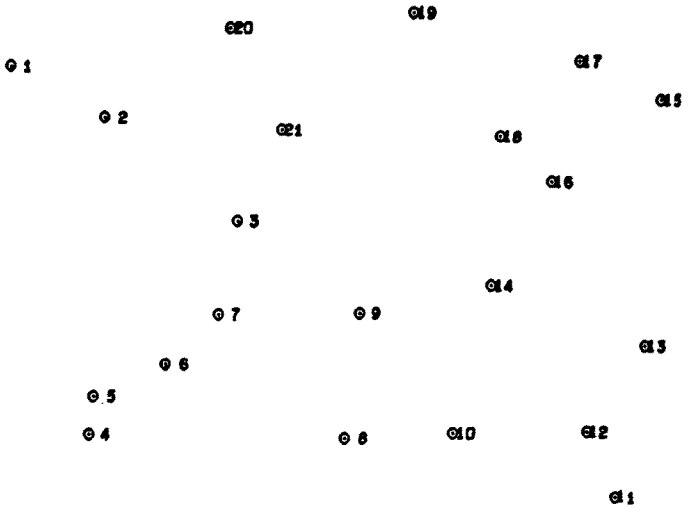

FIGURE 14
Configuration for 21 Colors of Constant Brightness
(Indow and Uchizono)

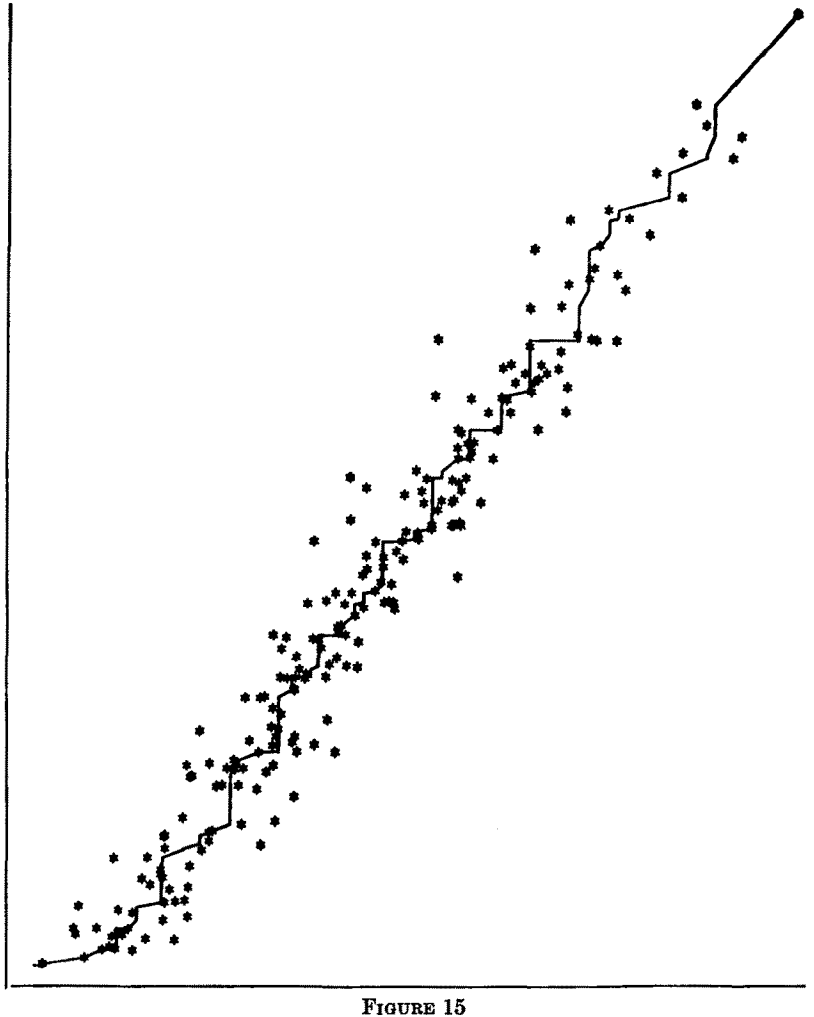

FIGURE 15
Scatter Diagram for 21 Colors of Constant Brightness
(Indow and Uchizono Data)

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

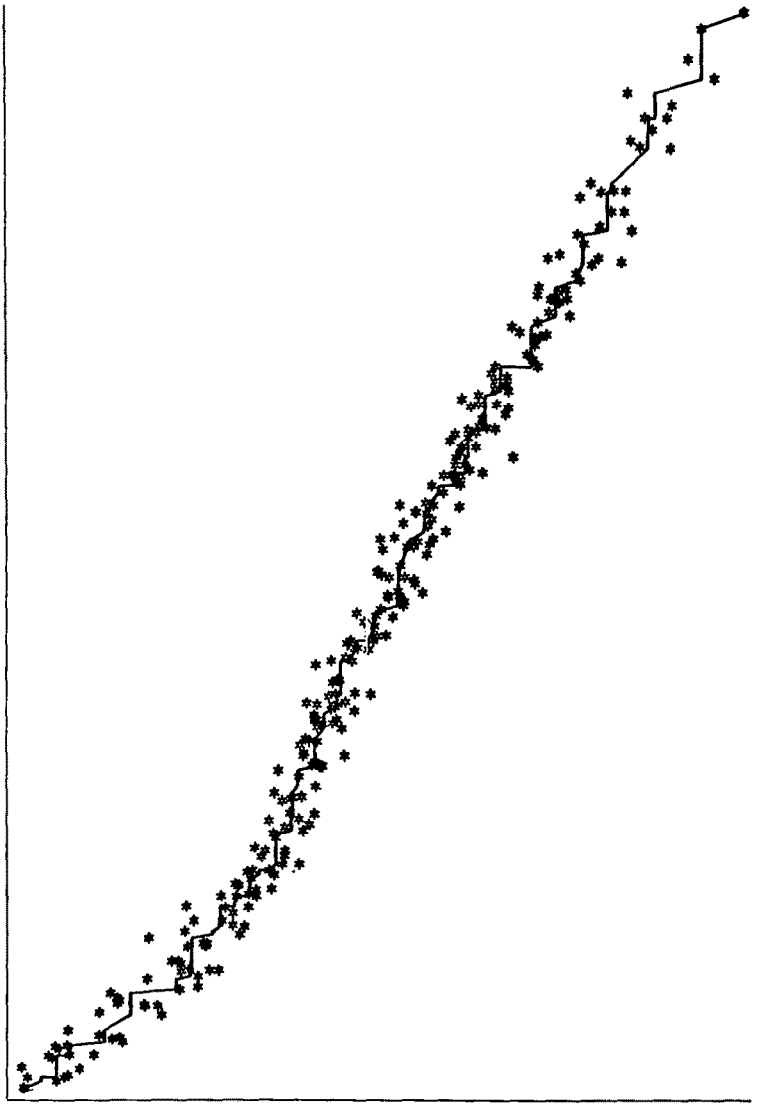

FIGURE 16
Scatter Diagram for 24 Colors of Varying Brightness (Indow and Kanazawa Data)

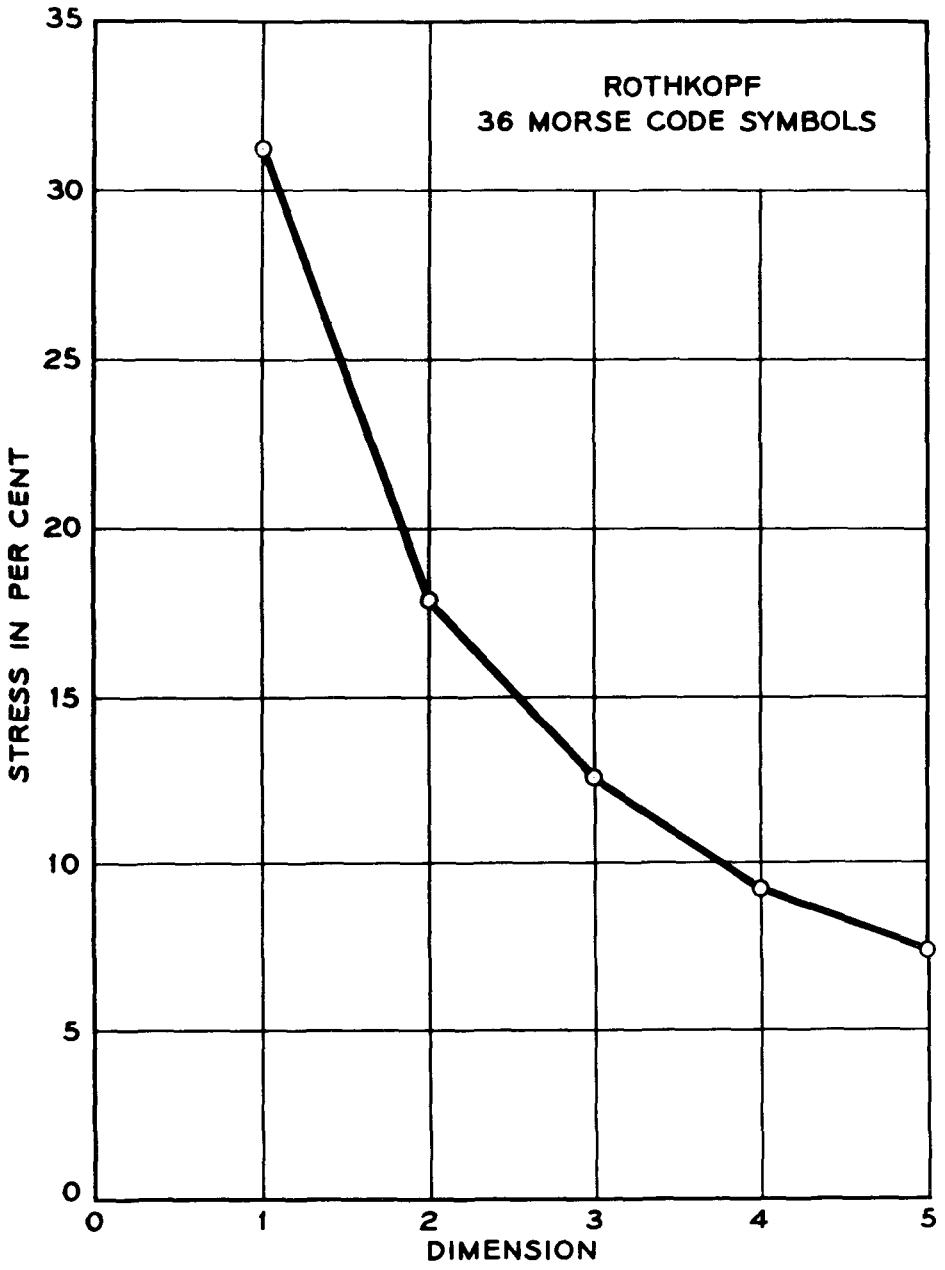
ROTHKOPF
36 MORSE CODE SYMBOLS
STRESS IN PER CENT
35
30
25
20
15
10
5
0
0
1
2
3
4
5
DIMENSION

FIGURE 17

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.

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

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

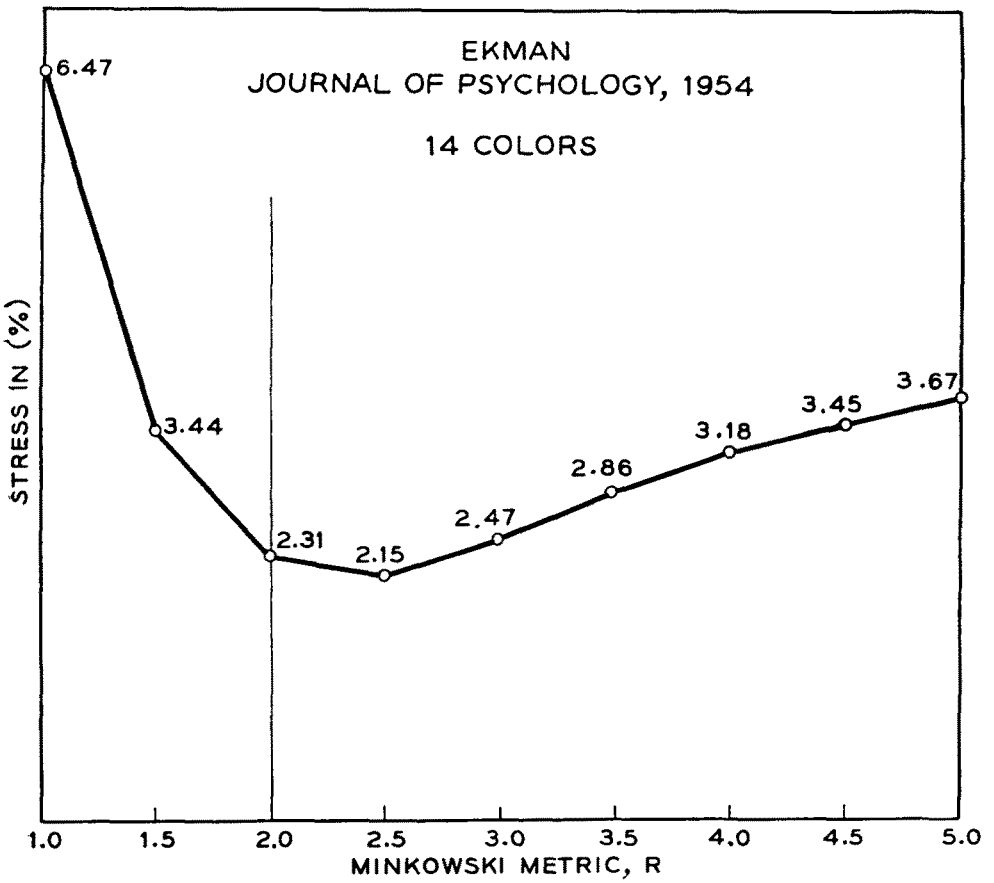

FIGURE 18

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 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