Monte Carlo Simulation Studies

Ian Spence
University of Toronto

This paper reviews the use of the monte carlo method to help illuminate various issues in the area of multidimensional scaling. Both two-way and three-way multidimensional scaling models and procedures are considered. Sampling distribution studies, studies comparing different procedures, and studies that have examined the basic capabilities of the methods under a variety of conditions are reviewed. Based upon the simulations, recommendations are given regarding several problems that face the user of multidimensional scaling techniques, for example, choosing a computer program, deciding upon the appropriate dimensionality or whether useful structure exists in the data, and dealing with large stimulus sets. Practical advice is given regarding the use of several computer programs, including M-D-SCAL, TORSCA, SSA-I, KYST, MINISSA-I, INDSCAL, ALSCAL, and MULTISCALE, as well as traditional Young-Householder-Torgerson scaling.

The invention of both the method and the name *monte carlo* is popularly credited to S. Ulam, a physicist who worked on the Manhattan Project during World War II. Certain calculations of a statistical nature concerning the amount of nuclear material necessary to achieve a critical mass were too complex to be solved analytically and had to be finessed by simulation of the physical process. During the simulation, random numbers were employed; therein lies the allusion to games of chance and, hence, the name *monte carlo*. Surprisingly, in statistics the method has a history almost as long as the subject itself. Apart from formal work on probability, which dates from the 16th century at the earliest (Cardano, 1501–1576; *Liber de ludo aleae*, published 1663) and the life tables of John Graunt (1662/1939), there is really little work of a truly statistical nature until the late 19th and early 20th century. Then, the demographic, psychometric, and biometric traditions flowered with the efforts of such men as Quetelet, Galton, and Pearson.

During this great intellectual upheaval, a paper was published by an obscure chemist who was employed as Brewer to Messrs. Guinness of Dublin. William Sealy Gosset was, by necessity, forced to consider the sampling properties of statistics computed on the basis of small samples—the assessment of quality in the brewing of beer can be a chancy business due to variations in lot characteristics, ambient temperature, humidity, and so forth, and consequently much has to be inferred on the basis of small samples. In his paper ''The Probable Error of a Mean'' (Student, 1908) Gosset discovered the distribution of the standardized sample mean. The exact statistic employed differs by a constant from the one used today and the proof is incomplete, since some steps are conjectured rather than proven; but this does not diminish the importance of a paper that has subsequently assumed the status of a classic in statistics. The result is, of course, familiar to all beginning students of statistics and is the basis of the t distribution.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 4, Fall 1983, pp. 405–425
© Copyright 1983 Applied Psychological Measurement Inc.
0146-6216/83/040405-21$2.30

Possibly as a consequence of his self-confessed lack of mathematical ability, Gosset felt that he had to conduct an experiment to bolster his slightly shaky formal argument. Using a known empirical distribution (actually the heights of 3,000 criminals!), he drew 750 random samples of size 4, calculated 750 values of what is now called *t*, and then compared the empirical frequency distribution with the distribution that his theoretical work had suggested. As reported in *Biometrika*, the two distributions were found to be in good agreement. Thus was seen, perhaps for the first time, the use of what is essentially the monte carlo method to help illuminate a problem in theoretical statistics. In this case the problem may be tackled and solved with full mathematical rigor and generality, as Fisher did in the 1920s; but the important thing is that Gosset's experimental demonstration helped bolster his tentative mathematics, and possibly was the crucial piece of support he needed to convince him to publish and, hence, allow others to benefit from his great discovery.

Other Applications of the Monte Carlo Method

The use of the monte carlo method is not confined to the examination of sampling distributions, although this has been the most common area of application in statistics. The method may be useful in any situation where a complete mathematical analysis of a problem is difficult or intractable. In recent years the method has frequently been applied to situations involving comparisons among estimators (cf. Andrews, Bickel, Hampel, Huber, Rogers, & Tukey, 1972) and to the problem of comparing methods or algorithms where mathematical analysis is problematic (e.g., Dempster, Schatzoff, & Wermuth, 1977). Although the monte carlo method can never be a totally satisfactory substitute for a thorough mathematical analysis, it is a simple fact of life that, for a variety of reasons, it may be difficult or impractical to obtain the desired analytic results. Even if an appropriate formal development is available, empirical demonstrations often provide insights that may lead to further development or, at the very least, provide some help in making the esoteric more intuitively accessible to the nonmathematical experimenter.

However, use of the method can have its drawbacks, and if unskillfully applied, it may sometimes do more harm than good. The monte carlo experiment is a designed experiment and, as such, is capable of displaying the same virtues and vices to be found in designed experiments in more familiar settings. Bad design, sloppy execution, and inept analysis lead to the same kinds of difficulties in the world of computer simulation as they do in the laboratory. On the other hand, if the computer experiment is carefully tailored to fit the problem at hand, if the work is executed with scrupulous attention to detail, and if the examination of the results is conducted with care and sensitivity, the outcome may have great utility, either as an adjunct to, or substitute for, a formal mathematical analysis.

Purpose of the Present Paper

This paper is intended to address two audiences. The first, and larger of the two, is composed of typical users of multidimensional scaling procedures; it is to be hoped that they may benefit from some knowledge of what has been done in the area of evaluation. The second audience, though small, is equally important. These are the people who will be conducting monte carlo experiments in the future. Those who have performed monte carlo studies during the last few years were, to some extent, learning the ropes as they went along. Much of what has been done could, on occasion, have been done better or more efficiently. Perhaps the first category of readers will be patient during a discussion of some of the issues that seem important regarding the conduct of simulation experiments. In the future this may help make for better monte carlo work and may make it easier for the consumer of multidimensional scaling procedures to evaluate the evaluators.

Limitations of the Monte Carlo Technique

Suppose that, as in Gosset's case, it is desired to examine the sampling distribution of the random variable $[\overline{X} - E(X)]/[S_X^2/n]^{1/2}$. This is the standardized sample mean, based on random samples of size n. Assume X to be Gaussian with known mean $E(X)$, but do not presume knowledge of the variance. Consequently, the random variable S_X^2, the unbiased estimator of the variance of $\overline{X}$, is employed in the standardization of X. If it was not known that the standardized mean was distributed as Student's t with $n - 1$ degrees of freedom, there might be some temptation to perform a monte carlo experiment.

It might seem that such an experiment would be child's play. A single replication would merely consist in computing the value of the standardized mean based on n values from a random Gaussian generator. The whole process would then be repeated as many times as thought necessary. There are many traps for the unwary in all of this. The first lies in the area of random number generation. Virtually all Gaussian generators depend upon having a good source of rectangularly distributed numbers, and although there are many publicly available programs, it cannot be said that all are good—indeed, some routines, such as the widely used RANDU from the IBM Scientific Subroutine Package, have atrocious properties. Given a source of respectable uniform random variates, there are several ways to produce Gaussian variates. Again, some are not very good. For sampling distribution work, approximate methods (such as Teichrow's) are not good enough. Excellent sources of advice regarding random number generators are Knuth (1969) and Kennedy and Gentle (1980).

The next difficulty will be in deciding on the values of n to be included in the experiment. This is a problem in selecting the levels of a factor similar to that encountered in any experiment. In this case, the unsophisticated approach of taking $n = 1, 2, 3, \ldots$ might be considered. This is unlikely to cause any problems in the simple situation considered here; but above $n = 35$, or thereabouts, essentially the same results will be obtained, and therefore the experiment will be terminated without excessive waste of computer time. A more intelligent approach, which will certainly serve the experimenter better in more complicated situations, should have been used: A small amount of pilot work would easily show the changes with n to be nonlinear and to reach asymptote before n becomes very large. Then, the advice of Cox (1958, chap. 7) regarding the choice of number and spacing of the levels of a factor could be heeded. One reason for preferring the more thoughtful approach is that the detail provided by the brute force approach is not needed—in most cases, sufficient accuracy may be obtained for intermediate values of n by interpolation. Another, better reason is that the experiment is unlikely to be as simple as first outlined. As described, the design is a simple one-way layout; but if it were considered desirable to extend the study to include parent populations other than the Gaussian, it can be seen that the amount of computing necessary could rise rather rapidly and, if the factor levels were not chosen intelligently, might turn out to be very expensive.

Most monte carlo studies have to be conducted within fixed budgets. Therefore, exhaustive evaluation is rarely a sensible, or feasible, option. Careful planning is important. The choice of factor levels and the inclusion of appropriate blocking variables will pay off here just as in conventional experimental settings. For example, in many experiments, the same random number seeds should be used across various factor levels in order to eliminate a source of variance that will otherwise help reduce the precision of the estimates. Often, it may be appropriate to run a sequence of small, related, carefully planned factorials rather than a gigantic multiple factor design that will use excessive amounts of computer time, and probably make it necessary to settle for an inadequate number of replications in each condition. Remember that in such studies the precision of the estimated treatment combination means will usually be required to be high, forcing a reasonable number of replications. In this context, the alert experimenter will entertain the possibility of using fractional factorial or other incomplete layouts if the likelihood of appreciable nonzero high order interactions is low. Not to be forgotten, also, is the possibility of using variance reduction techniques, monte carlo "swindles," in certain situations (Andrews et al., 1972). In general, however, there will be the usual tradeoff between precision of estimation and the number of replications. The only good general advice that can be given here is that there is little point in doing most monte carlo work if the level of precision is not adequate; a priori calculation of the standard errors of the estimates should be routine, even if some assumptions and approximations have to be made.

Monte carlo experiments typically result in the production of much more data than in the average psychological laboratory experiment. Thus, the data analysis and presentation of the ensuing summaries can present considerable difficulty. Although compact tables and appropriate graphics will constitute the usual mode of presentation, it is frequently possible to summarize by the use of analytic approximating functions; these may be theoretically motivated or merely empirical curve fitting. It is difficult to discuss these issues much further without exploring specific examples (e.g., Hoaglin, 1977; Spence, 1979), but the important point is that the presentation of a table of treatment combination means is unlikely to do full justice to the data.

Multidimensional Scaling and the Monte Carlo Method

Methods of two-way multidimensional scaling (cf. Torgerson, 1958), based on the theorems of Young and Householder (1938), have not stimulated any significant amount of monte carlo work. This is largely because the mathematics of the situation are very well understood. Nonmetric scaling procedures (cf. Guttman, 1968; Kruskal, 1964b; Shepard, 1962a), on the other hand, require the use of iterative processes whose properties are less easy to evaluate. For example, although a solution is guaranteed, it may not be the best possible one: The iterative procedure may converge to a position where no further improvement is possible by making small adjustments, and yet a better solution may exist somewhere else. This is the local optimum problem plaguing many algorithms that use successive approximations. Certain strategies may render a particular procedure more or less susceptible to this problem than others, and it is important to have some way of assessing the value of such refinements. (For a general review and a discussion of the operation of both two- and three-way algorithms, see Spence, 1978, or Davison, 1983).

In the three-way multidimensional situation, even in the metric case, virtually all commonly used algorithms are iterative in their operation and thus potentially suffer the same local optimum problem. Also, it is of interest to have some idea of the general capabilities of three-way scaling programs: how performance is affected by increasing the number of stimuli or subjects, and what effect error has on recovery. Additionally, other problems that are difficult to treat mathematically arise, for instance, the effects of missing data or, in the case of programs that can accommodate different assumptions regarding the strength of the data, how performance is affected.

In multidimensional scaling, therefore, it might be desirable to know something about the sampling behavior of goodness-of-fit statistics or to have some objective basis for choosing one algorithm as opposed to another. It is also important to know how missing data affect the performance of the procedures, and whether there is some way of making better decisions with regard to the choice of appropriate dimensionality. These, and other, concerns may be approached by use of the monte carlo method.

Two-Way Scaling: A Review of Monte Carlo Studies

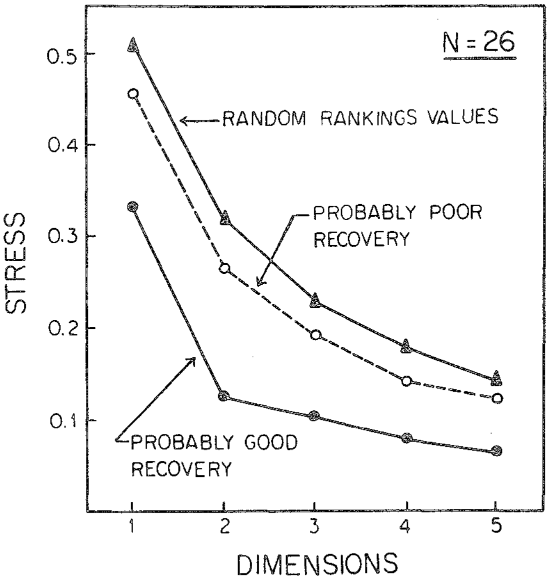
Sampling Distribution Studies

Apart from some small demonstrations, the first monte carlo experiments in the area of multidimensional scaling were concerned with an examination of the behavior of the goodness-of-fit statistic STRESS Formula One (Kruskal, 1964a) under the null hypothesis that the data input to a nonmetric scaling program are essentially random numbers (Klahr, 1969; Spence & Ogilvie, 1973; Stenson & Knoll, 1969; Tschudi, 1972; Wagenaar & Padmos, 1971). These experiments display a variety of good and bad practice in the design, execution, and analysis of monte carlo experiments, and none may be said to be completely satisfactory. However, the experimental situation is a relatively simple one, and consequently the results obtained are useful despite the problems. For a discussion of some of the difficulties, and a comparison of the studies, see Spence and Young (1978; also reprinted in Young & Lewyckyj, 1981). Partly because of the advantage of hindsight, perhaps the most useful summary of the results of this kind of experimentation is to be found in Spence (1979; reprinted in Davies & Coxon, 1982).

The null hypothesis to be tested is rather a weak one: It asserts that the input data are random. This will rarely be the case in practice; consequently, the null hypothesis will probably be rejected in most situations in which the results of the above papers are applied. It is possible to reject the hypothesis of randomness and yet the data may contain very little useful structure. The situation is akin to rejecting the hypothesis of a null population coefficient of correlation: It may be comforting to do so but does not guarantee that the fit of the data to the model is either good, or useful in practice. Perhaps the most valuable gain for the experimenter from an examination of the results of scaling random data is an intuitive appreciation of the worth of the data (see Figure 1 for an illustration). If the obtained stress values are close to the random values, then even though the formal hypothesis of randomness may be rejected, the multidimensional scaling representation may not be very useful. On the other hand, if the obtained stress values are only a third or a half as large as the results obtained from scaling random data, then the experimenter may have much more confidence in the solution.

The advantage of using the Spence (1979) results in this context is that an analytic approximation has been provided. This approximation is valid over a wide range of values of both the number of points and dimensions. Rather than the user having to look up a table, or interpolate in a graph, the appropriate values may be entered in a simple equation to produce the corresponding random rankings STRESS value. A simple hand calculator (with a logarithmic function key) may be used. The error of the approximation is of the same order as the standard error of the estimates in any of the above-mentioned random rankings studies.

Figure 1
Obtained STRESS Values When the Solutions Represent (1) Essentially Random Rankings, (2) Probably Good Recovery of Useful Structure, and (3) Probably Poor Recovery or Non-useful Structure



Basic Capabilities

Although several small demonstrations (e.g., Kruskal, 1964b; Shepard, 1962a, 1966) had suggested that nonmetric procedures could do very well in recovering known multidimensional structure, not much was known about how performance might vary as a function of the number of points, dimensions, or the error level in the data. Extensive studies by Young (1970), Spence (1970), and Sherman (1972) provided the answers. These studies, as well as most of the other studies discussed in this paper, employed the same general method. First, a configuration of points in a space of fixed dimensionality was generated; this was done either randomly or systematically. Then, error was introduced in some fashion to distort the true distances of this known configuration, and sometimes a nonlinear transformation of the distances was also employed. Various magnitudes of error variance were used, ranging from zero up to very high values.

It should be noted that the specification of an error model can often be a difficult choice for the experimenter. An attempt is being made to simulate an aspect of the real world, but error in the real world can come in many varieties, depending upon the judgment task or other aspects of the experimental situation; it is difficult to cover all possibilities in the simulation. Another consideration, especially when comparisons of algorithms are being made, is that different procedures make, either explicitly or implicitly, different assumptions about the distribution of errors; consequently, the performance of a given procedure may or may not be significantly altered by the kind of error distribution employed. Cohen and Jones (1974) have given a good discussion of these and various other issues relating to the choice of error models in simulations.

The error-perturbed distance matrix, however obtained, is then considered to be a matrix of dissimilarities and is input to a multidimensional scaling program with the objective of seeing how well the recovered configuration will match the known true configuration. The process may then be repeated a number of times with different samplings from the error distribution and the population of configurations. Finally, depending on the purpose of the study, this procedure is repeated with different values of the parameters of interest, such as the number of points, or dimensions, or fraction of missing data, or number of subjects in the three-way case.

Broadly speaking, nonmetric procedures have been shown to perform best when the true dimensionality is low, the number of points is large, and the error level is low. One very important finding from these studies is that the minimum number of stimuli (points) to be scaled should be no less than about six times the expected number of dimensions. Thus, in one dimension 6 points is the minimum, in two dimensions the minimum is about 12, and so forth. This is to ensure that the problem of degenerate solutions is minimized. Degenerate solutions are solutions where the goodness-of-fit statistics can look quite reasonable but where the solution is actually very poor. For example, in one dimension, it is usually possible to arrange 3 or 4 points on a line in many ways such that the rank order of the distances is the same as, or close to, the rank order of the dissimilarities. Since only one solution can be ‘‘correct,’’ however, it follows that all others are, in some sense, degenerate. Indeed, many of these other solutions may bear almost no resemblance to the correct solution. With more points, the relationship between the solution and the data is much more tightly constrained and the result is much less likely to be degenerate. (Note that metric procedures are not susceptible to this difficulty, since they make use of more than just rank-order information about the distances.) The reader is referred to Young (1970), Sherman (1972), Spence (1972), and Shepard (1974) for more detail.

Incomplete Design Studies

When the number of points is large, the labor of collecting all possible pairwise judgments is great. Several methods of reducing this work to manageable proportions have been proposed. Two main lines of research may be traced. First, the data may be collected interactively under the control of a computer (Young & Cliff, 1972). Second, the choice of pairs to be judged may be decided a priori by experimental design. Such a design may ignore the possible characteristics of the data (Spence & Domoney, 1974) or may make some use of a knowledge of the likely interpoint distances (Graef & Spence, 1979). Alternatively, an approach that combines features of both may be considered (Isaac, 1982). Monte carlo studies have been conducted to help decide which of the above approaches will work best. This is a difficult area in which to do good work, since the number of factors of interest, and the possible number of levels, is large. Consequently, the conclusions drawn are sometimes tentative. Perhaps the best recent source of collected wisdom on this topic is to be found in Golledge and Rayner (1982). Several chapters in this book, by diverse authors, deal with the topic of incomplete designs and interactive data collection. Some of the findings are summarized here, but without going into much detail concerning the simulation studies.

Spence and Domoney (1974) were the first to conduct a systematic monte carlo study examining the effects of having missing paired comparisons, by design, in a scaling experiment. They found that several methods of reducing the number of pairs could provide satisfactory results, providing that a sufficient fraction of the possible set of pairs was obtained. Although this minimum fraction depends upon the level of error in the data, it seems that if the fraction collected exceeds the ratio (6 × number of dimensions/number of points), then this will be adequate in practice. If the fraction of data collected is much less, then there may not be enough information in the available data to support the construction of a good multidimensional representation. The issue is no different here than with any other kind of parameter estimation: Too little data means that good estimates cannot be expected, and if the data are truly sparse it may be that no estimation is possible.

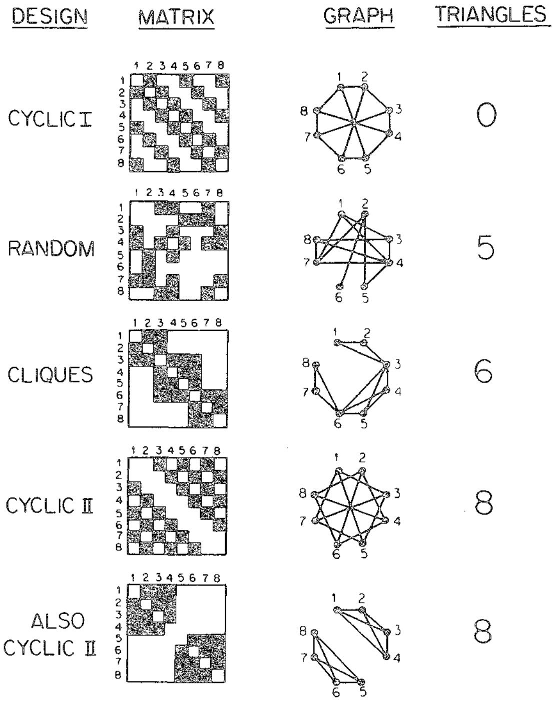
However, Spence and Domoney (1974) have qualified this advice by pointing out that, even with sufficient fractions, only designs that have high "global connectedness" should be used (see below for illustrations of this concept). This is further supported in Spence (1982a), where it is shown that as the degree of global connectedness decreases, the quality of a multidimensional scaling solution deteriorates. Some small illustrations of the graphs of a few designs are shown in Figure 2. In each of the graphs, if two vertices are connected by an edge, this signifies that the dissimilarity between that pair of stimuli was observed, whereas if two vertices are not connected, the dissimilarity corresponding to that pair is missing. The reader is cautioned that the points in such graphs are positioned arbitrarily and, of course, bear no relationship to the positions of the points in a multidimensional scaling solution. Also, these are merely illustrations and, in practice, the order of the data matrix will be much larger; more than 30 points would typically be involved, rather than 8.

Global connectedness may be assessed in a variety of virtually equivalent ways, one of which is by counting the number of triangles in the graph of the design.

The number of triangles in the graph of a design is an index of how well the design may be expected to perform; fewer triangles are associated with better performance. Further details are to be found in Spence (1982a), but for the moment, the reader may gain an intuitive appreciation of the concept of global connectedness versus "local connectedness" by studying the examples in Figure 2. Note that a large number of triangles are associated with local connectedness, whereas few triangles indicate global connectedness.

Even randomly generated designs will usually work well, but perhaps the simplest way to obtain a good design is to construct a cyclic design (so-called because of the way in which the pairs to be observed are written down, see Spence, 1982a). Most cyclic designs will perform well, but it is wise to choose one with high global connectedness.

Figure 2
Some Possible Incomplete Experimental Designs with the CYCLIC II Designs Equivalent (Obtained by Relabeling the Points)



ALL DESIGNS
REPRESENT A 12/28 = 43% FRACTION

Note, for example, that Cyclic II in Figure 2 has a large number of triangles in its graph and is not even connected. This means that the positions of points (1, 3, 5, 7) and (2, 4, 6, 8) relative to each other could *never* be determined. Spence (1982a) provides an algorithm for finding an optimal cyclic design as well as a simple paper-and-pencil technique that will yield an optimal, or close to optimal, design with much less effort.

The kinds of a priori designs that should be avoided are those that contain strongly connected blocks with very few linking comparisons (see Figure 2). For example, a design that collects all paired comparisons for subsets of the stimuli and only a few comparisons involving stimuli from *different* subsets will probably not yield good results. However, Spector and Rivizzigno (1982) have shown that if the number of linking comparisons is large enough, reasonable recovery may be obtained, providing that the error level is low. Their results for high levels of error are somewhat difficult to interpret, since they used a fairly small, though seemingly adequate, fraction of the possible data (37%). Perhaps their results suggest that when the error level in the data is expected to be high, then either multidimensional scaling should not be employed or the user should attempt to collect a fraction well in excess of the recommended minimum, whatever the design. (Of course, nobody believes that his or her data will turn out to be excessively noisy, so this is something of a counsel of perfection.) Unfortunately, none of the existing monto carlo studies is adequate to the task of providing more explicit guidelines for situations where the error level is quite high, and a complicating factor is that the results might well vary as a function of the distribution of true distances (Isaac, 1982).

Interactive Data Collection

An alternative to the use of a priori fixed incomplete designs is the interactive collection of data. This requires the use of a dedicated or timesharing computer for the acquisition and possible analysis of the data and has the advantage that the collection of particular data may be guided and modified by the results of examining the subject's previous responses. At least two different ways of going about this are available. The first allows the simultaneous acquisition, and scaling analysis of the data and is typified by a program like ISIS (Young & Cliff, 1972). In the second approach, there is no need to have the scaling done at the same time as the data are being collected. This may be an advantage in terms of computer storage and execution time: The data might be collected using a small laboratory computer or a microcomputer and then subsequently analyzed on a large mainframe computer. Young, Null, Sarle, and Hoffman (1982) have described and evaluated a program called ISO, which is designed to operate in this way.

ISIS (Young & Cliff, 1972) requires that the distance judgments of the subject be considered to be Euclidean. This might seem to be quite restrictive; but, in practice, many data sets will satisfy this assumption. The ISIS program first establishes a "frame," which consists of a small subset of points that are relatively far apart from each other; this is done by collecting all pairwise judgments for a larger, but still small, subset and then eliminating some points after solving for the subset configuration using traditional methods. Subsequently, additional points may be added using only a knowledge of their distances from the frame points. Distances among the added points do not need to be determined. Various refinements, such as periodic updating of the frame, have been developed by Young and Cliff (1972). Monte carlo evaluation (Baker & Young, 1978; Girard & Cliff, 1976; Hamer, 1978) has shown that when the Euclidean assumption is satisfied, ISIS performs well using only 25% to 45% of the possible pairs and is generally superior to fixed designs with the same fraction of data. On the basis of monte carlo, and other work, an improved program called INTERSCAL has been developed (Cliff, Girard, Green, Kehoe, & Doherty, 1977; Green & Bentler, 1979) and should probably be preferred to the earlier versions of ISIS. Of course, when the data do not satisfy the Euclidean assumption, it may be desirable to consider fixed designs or a procedure like ISO, which is discussed below.

A nonmetric multidimensional scaling algorithm, such as KYST (Kruskal, Young & Seery, 1978; see next section), basically requires the rank order of all possible *pairs* of stimuli. For example, if there are 20 stimuli, the program needs to know the rank order of the 190 possible pairs of stimuli, from smallest to largest, in order to obtain a solution. One task for a subject could be to select the most similar pair, then the next most similar pair, and so on, until the 190 pairs had been ordered. This can be a lot of work, even with only 20 stimuli. Fortunately, several good algorithms originally designed for sorting numerical data may be adapted to make the human subject's task easier. This is what has been done by Young et al. (1982) in the ISO program. They have adapted a sorting procedure that was developed for the rapid sorting of numbers in order of their magnitudes. Their evaluation of the procedure, using stimulus sets of moderate size, has shown that this represents a convenient alternative way of reducing the amount of work required of the subject.

Algorithm and Program Comparisons

This is another difficult area. Algorithms are often modified during their early lives, and commercially obtainable computer programs are even more frequently altered by their authors or distributors. Deciding which to use on the basis of monte carlo results is a little like trying to buy a 1983 car based on a report in a 1973 consumer magazine. Neither the programs nor the cars remain the same—and new models are not even considered. All is not lost, however: There are enough results that endure to help make a sensible choice.

At least two approaches to the problem of designing monte carlo studies may be distinguished. In the first, *algorithms* may be compared and contrasted, sometimes without any immediate intention of commenting on particular computer programs. Such an approach is exemplified by the extensive study of Lingoes and Roskam (1973). Alternatively, publicly available computer *programs* may be evaluated by running them, in essentially their original state, with common simulated data as input; the experiments of Spence (1972) employed this strategy. The two approaches overlap considerably, and it may be instructive to compare their findings. Often they concur, but sometimes there are disagreements on points of detail.

Spence (1972) examined the performance of three well-known programs: M-D-SCAL, SSA-I, and TORSCA-9. Each algorithm was run with simulated data under a wide range of conditions—the number of points was varied from 6 to 36, the number of dimensions from 1 to 4, and the error level from zero to a moderate amount. It was found that all three programs performed very well in recovering the known structure, with TORSCA perhaps the best, but with M-D-SCAL producing a moderately large number of suboptimal solutions, especially in one dimension. Based on an analysis of the performance of the various starting configurations employed by the three programs, Spence (1972) attributed M-D-SCAL's deficiencies to the inadequacy of the arbitrary starting configuration employed. TORSCA's starting configuration was easily the best, and it is interesting to note that subsequently Young and Kruskal, the authors of the programs TORSCA and M-D-SCAL, have joined forces to produce a program that combines the TORSCA start with the M-D-SCAL iterative process. This program is called KYST (Kruskal, Young, & Seery, 1978). Although the program has never been evaluated in an extensive monte carlo study, it seems reasonable to assume, based on Spence (1972) and on Lingoes and Roskam (1973), that it is less susceptible to the suboptimal solution problem than most other nonmetric scaling programs that are publicly available.

Possibly the only program that will match KYST's performance is MINISSA-I (Roskam & Lingoes, 1970). MINISSA-I may be regarded as an improved version of SSA-I, and since Spence (1972) found that SSA-I performed almost as well as TORSCA, it would seem safe to predict that MINISSA-I is the equal of TORSCA, and possibly of KYST. The differences between these three programs in terms of performance will be so small that a choice of one will probably be more a matter of convenience, related to the input/output and other options offered by the programs.

Many readers will be curious to know whether the performance of ALSCAL (Takane, Young, & de Leeuw, 1977) matches that of KYST when a two-way scaling solution is obtained. Unfortunately, it is not easy to answer that question, since no systematic comparisons have been made. It is not appropriate to conjecture regarding the capabilities of ALSCAL, as was done above with MINISSA-I, since ALSCAL fits *squared* distances to *squared* dissimilarities rather than distances to the raw data. This could possibly accentuate the effects of error and cause overall performance to be inferior to that of procedures fitting the simple dissimilarities directly. For further discussion of this issue, see Krane and Spence (1980).

The Lingoes and Roskam (1973) study represents the only other major empirical attempt to compare the various available algorithms. Both Spence (1972), and Lingoes and Roskam (1973), have provided references to, and discussion of, a number of smaller scale studies that have examined some aspects of particular algorithms, including, most importantly, Gleason (1967). The conclusions drawn by Lingoes and Roskam do not differ substantially from those of Spence: They agree that a good initial configuration is very important and that a single arbitrary, or random, configuration should not be used.

Lingoes and Roskam (1973) conducted a detailed analysis of the performance of the iterative portions of M-D-SCAL and SSA-I. In these iterations certain quantities are required at each step during the minimization process; these are termed "targets" by Lingoes and Roskam and "pseudo distances" by Spence. Without going into detail, there are two frequently used ways of getting these: the first is attributable to Kruskal (1964b), who called them "*d*-hats," the other is attributable to Guttman (1968), with the resulting quantities called "rank images," or "*d*-stars." The major difference between the two approaches is that the *d*-hats provide a least squares fit to the data during each step of the iterative process, whereas the *d*-stars do not. Lingoes and Roskam (1973) have shown that even though the rank images lack the least squares property possessed by the *d*-hats, they perform quite well in the iterative process; indeed, if the initial configuration is far from optimal, they may have some advantage over the *d*-hats in terms of helping the algorithm avoid locally optimum, but undesirable, solutions. This observation was also made by Spence (1972). It should not be forgotten, however (cf. Spence, 1972), that if *d*-stars are used throughout, the final solution will not possess the least squares property; but since both SSA-I and MINISSA-I allow the user to use a concluding set of iterations with *d*-hats, this objection is largely academic.

Lingoes and Roskam (1973) examined other aspects of the behavior of nonmetric algorithms, such as the effect of ties in the data and the different strategies for dealing with this circumstance. Additionally, they investigated the effects of various modifications to the basic iterative process. Their monograph is an indispensable source of information to the expert in the construction of algorithms for nonmetric scaling; the average user, if unfamiliar with previous work such as Kruskal (1964b) and Guttman (1968), may find some of the detail a little difficult to follow.

Determining Dimensionality

A number of studies, going back as far as Stenson and Knoll (1969), have attempted to provide information that bears on the problem of deciding upon the appropriate number of dimensions. Stenson and Knoll suggested that the user make a choice of dimensionality based on a criterion such as interpretability, or perhaps the existence of an "elbow" in the graph of STRESS versus dimensionality (see Figure 1). Then, the obtained STRESS value in that dimensionality should be compared with the value obtained by scaling random data: "If the empirical value of [STRESS] is too close to the null hypothesis value, the appropriateness of the chosen number of dimensions should be questioned" (Stenson & Knoll, 1969, p. 125). However, it has been shown by others (e.g., Spence & Graef, 1974; Wagenaar & Padmos, 1971) that if the data are not random, the STRESS for *any* recovered dimensionality will be smaller than the random data value in that number of dimensions—and this includes the case of an overestimated number of dimensions (see Figure 1). It is thus doubtful that Stenson and Knoll's (1969) procedure accomplishes the desired end.

Isaac and Poor (1974) have tried to refine the basic Stenson and Knoll (1969) idea by considering the magnitudes of the *differences* between random data STRESS values and obtained STRESS values in spaces of successively increasing dimensionality. They introduced a measure called Constraint, which is simply the difference between the random rankings STRESS value and the obtained STRESS value. In their simulation they showed that the Constraint value was often maximal in the space of appropriate dimensionality, especially with low to moderate error levels. Unfortunately, their simulation was quite small in scope, and it is not known whether the Constraint criterion will perform adequately in general; for example, its application to the simulated data collected by Spence (1970) frequently yielded incorrect results. However, the idea is appealing by virtue of its simplicity and perhaps deserves further consideration and evaluation.

Wagenaar and Padmos (1971) proposed an ingenious paper-and-pencil technique for determining dimensionality. Based on a small set of monte carlo data obtained by scaling data with known dimensionality and error level, they described a two-stage procedure for determining the dimensionality and error level. The details will not be repeated here, since the technique to be described below is quite similar and probably represents an improvement.

Spence and Graef (1974) have developed a procedure that is implemented in a FORTRAN IV program called M-SPACE. Based on extensive monte carlo data, this program attempts to find the dimensionality and error level that best characterize the obtained STRESS values in a least squares sense. The idea is similar, in some respects, to that of Wagenaar and Padmos (1971), but a much more extensive set of data is used, and the fitting is done objectively, rather than subjectively. M-SPACE is capable of determining dimensionality up to four dimensions, with stimulus sets of from 12 to 36 objects inclusive. An important aspect of the method is that the approximate level of error in the data is determined; consequently, the user has a firm basis for evaluating the likely worth of the solution.

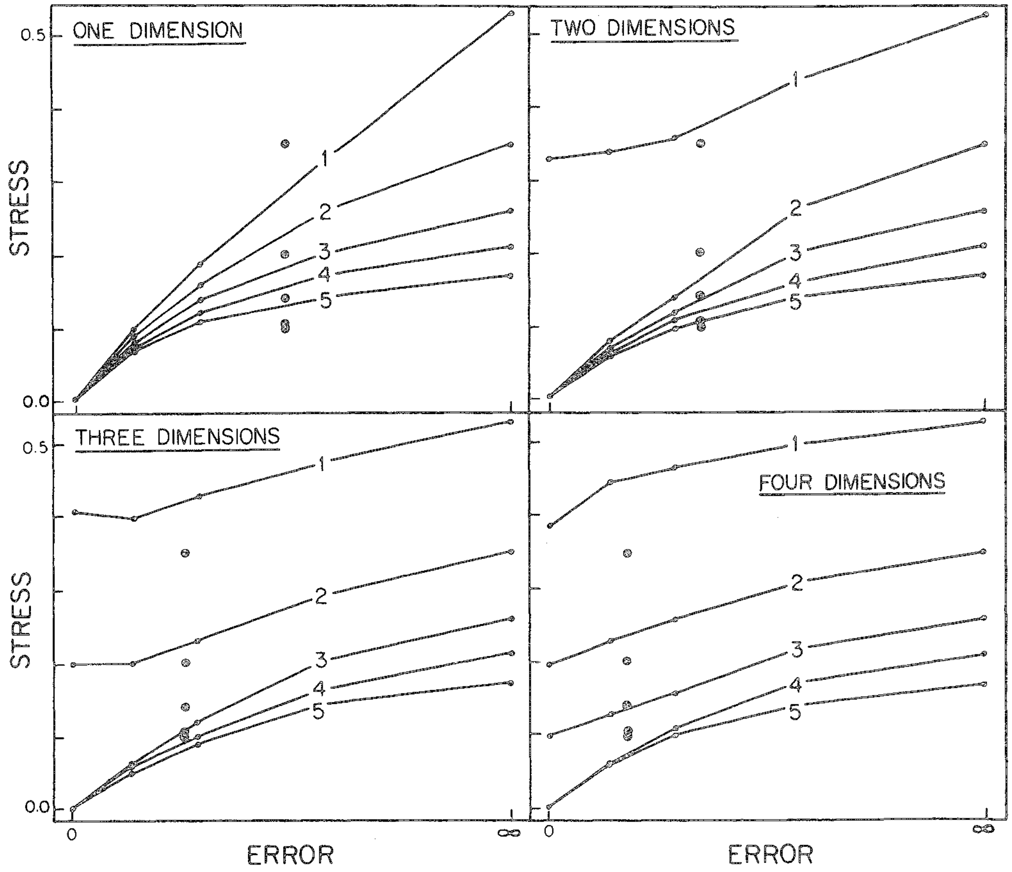
Figure 3 shows an application of M-SPACE to the Rothkopf (1957) Morse code data previously analyzed by Shepard (1963). The unbroken lines in the graphs are based on monte carlo simulation data for the appropriate number of points, which is 36 in this case; the closed circles represent the obtained STRESS values from scaling the Rothkopf data in one through five dimensions. M-SPACE has found the error level where the STRESS values most closely fit the monte carlo data. Once this has been done for each of four true dimensionalities (from the computer simulation), it can be seen that although not perfect, the best fit is to the two-dimensional monte carlo data. The best candidate for the dimensionality of the Rothkopf data is thus two, and the error level is moderate. It is interesting to note that since there is a lack of a clearly defined "elbow," application of the elbow criterion would be difficult in this case (see Figure 4). Shepard (1963) arrived at the same conclusion regarding the number of dimensions, based upon the interpretability of the solution.

M-SPACE has been used successfully in a variety of applications, for example, ethology (Miller, 1975), market research (Day, Deutscher, & Ryans, 1976), color perception (Tansley & Boynton, 1978), psychopathology (Chan & Jackson, 1978), and occupational mobility (Coxon & Jones, 1980).

Robustness

Although there has been much work on statistical methods that are resistant to the effects of outliers during the last 15 years, it is only recently that attention has turned to this topic in multidimensional scaling. Spence (1982b) has shown that traditional metric methods can be adversely affected by even a single outlying observation. Null and Sarle (1982) have also demonstrated that the results obtained by conventional methods may be improved upon by using a robust procedure. Spence's (1982b) algorithm uses a modified Newton iteration technique where the median of successive corrections is employed, whereas Null and Sarle (1982) have adopted a different approach involving the iterative minimization of other functions than the usual sum of squared errors employed in classical least squares. Both approaches appear to be capable of doing much better than conventional techniques.

Figure 3
Graphs of Monte Carlo STRESS Values Produced by M-SPACE
During an Application to the STRESS Values Obtained by Scaling
the Rothkopf (1957) Morse Code Confusions Data (Shown as Closed Circles)



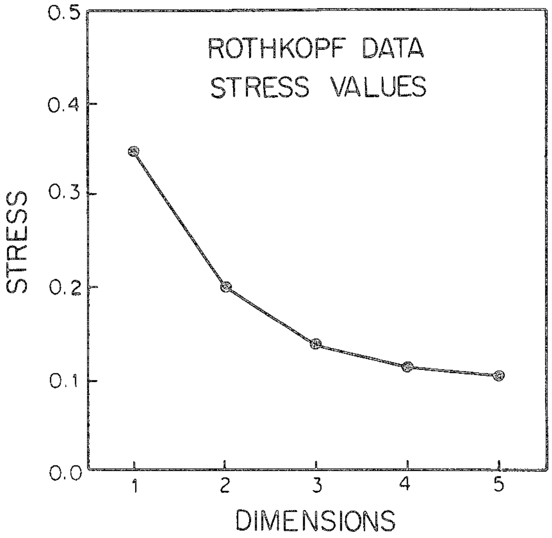
Lewandowsky and Spence (1983) have performed a fairly large monte carlo simulation showing that the program TUFSCAL is capable of tolerating a very large number of outliers. They have also shown that TUFSCAL can tolerate outliers in the data much better than can KYST, MULTISCALE, or traditional metric scaling. Also, when outliers are *not* present, it appears that TUFSCAL's performance is as good, if not better, than these other procedures.

Both Spence (1982b) and Null and Sarle (1982) have developed algorithms for three-way, as well as two-way, scaling. Practical, publicly available programs should be ready soon.

Three-Way Scaling: A Review of Monte Carlo Studies

Currently, several procedures are available for three-way multidimensional scaling. Originally, only one program was widely used (Carroll & Chang, 1970), despite the fact that the weighted Euclidean model had been independently considered by a number of other workers (e.g., Bloxom, 1968; Harshman, 1970; Horan, 1969). In addition to INDSCAL (Carroll & Chang, 1970), programs such as ALSCAL (Takane et al., 1977), MULTISCALE (Ramsay, 1977, 1978), SUMSCAL (de Leeuw & Pruzansky, 1978), and direct least squares

Figure 4
Obtained STRESS Values for the Rothkopf (1957) Morse Code Confusions Data Versus Dimensionality



metric procedures by Schonemann (1972), Bloxom (1978), and Krane (1978), plus a few others, are now available. Although they all fit some version of the weighted Euclidian model, they differ considerably in several respects, notably in the ways in which they have chosen to handle the spatial and distance components of the model as well as the method of estimation. The spatial component of the model defines the relationship between the coordinates and the distances in the geometrical representation, whereas the distance component of the model involves the specification of the function that relates the data and the distances in the representation. Although most of the programs use some variant of a least squares fitting approach, Ramsay (1977) has chosen the method of maximum likelihood. See Krane and Spence (1980) for further discussion of these points.

Monte carlo work on three-way scaling problems is much more sparse than in the two-way case, and with one or two exceptions, there is little in the literature that can be regarded as independent evaluation. Conspicuously absent are comparative studies that compare and contrast the several approaches. The authors of each of the procedures have performed some simulations, including comparative demonstrations, but ideally it would be desirable to have more third-party evaluation. The reasons for the deficit are not mysterious. It is quite expensive to run just one of these procedures with a typical empirical data set. Consider how much more it costs to obtain three-way scaling solutions with hundreds or thousands of artificially constructed data matrices, using three or four different programs! Added to this is the problem of experimental design. As noted above, the various available programs deal with the problem of fitting the weighted Euclidean model in different ways, and this makes the problem of comparative evaluation much more troublesome than in the two-way case. Furthermore, there are so many ways in which the parameters of interest might be varied that many potential monte carlo studies have probably died on the drawing boards as their authors concluded that it would be too difficult to do useful work within a limited computing budget. Nevertheless, the situation is not hopeless. Over the years, as empirical experience has been gained with these techniques, the important issues seem to have crystallized, and with clever experimental design it should be possible to examine them without spending huge amounts of money. Some of these issues have been discussed in detail by Krane and Spence (1980), but it may be useful to give a brief recapitulation.

Basic capability

Most multidimensional scaling models are composed of two component parts. First, as noted above, there is a distance component that specifies the relation between the dissimilarity and the distance; and second, a spatial component that specifies how the coordinates are related to the distances. The three-way scaling programs described by Bloxom (1968), Carroll and Chang (1970), Takane et al. (1977), Ramsay (1977), and Krane (1978) all take different approaches to fitting these two components: No two programs fit exactly the same model. Some programs—those of Bloxom (1968) and Carroll and Chang (1970), as well as the various successors of INDSCAL—adopt a sequential approach: this means that they first fit the distance component of the model to the observed dissimilarities, producing estimated scalar products, and then fit the spatial component to these estimated scalar products. Most other programs attempt to fit both distance and spatial components simultaneously to the observed dissimilarities (or to a transformation of their squares, as in the case of ALSCAL). This distinction may appear a little confusing to the casual user and may not seem to be terribly important; however, in certain situations the programs could produce rather different results as a consequence. For example, if a sequential strategy is employed, sampling errors are ignored during the fitting of the distance component of the model. The spatial component is subsequently fitted to the estimated scalar products, and it is not clear what kind of distortion is produced by having previously ignored the sampling errors. The simultaneous estimation procedures are, of course, unaffected by this problem.

Each of the three-way programs discussed above employs a different numerical procedure: Some use simple gradient-based descent methods, whereas others use more sophisticated quasi-Newton procedures, and yet others rely on what have come to be known as alternating least squares algorithms. Convergence rate considerations are important here. Since these programs can be very expensive to use, it would be useful to know something about the empirical convergence rates of alternating least squares and a quasi-Newton method. In theory, the latter might be expected to be superior, but the functions being minimized are very complicated, and it is not clear that there can be too much reliance on the conventional theory.

Another question of interest concerns ALSCAL, a very popular three-way scaling program. As noted before, ALSCAL fits squared distances to a transformation of squared dissimilarities. It seems likely, then, that ALSCAL might be adversely affected when large sampling errors are contained in large dissimilarities. A few outliers, for example, might exert tremendous leverage on the solution. A small monte carlo experiment could provide the answer—so far it has not been performed.

Finally, the effect of the starting configuration on the final solution is of interest, just as it is in the two-way case. So far nothing has been published on this topic, but there is some empirical work that suggests that the effects of different starting positions are not trivial.

Carroll and Chang (1970) may have been the first to perform a simulation with a three-way scaling program. They examined the effects of scaling random data. However, they obtained only one replication with 25 points and 12 subjects; so the results do not permit any really useful conclusions to be drawn. Takane et al., (1977) conducted a more ambitious, though still small scale, study that examined the ability of ALSCAL to recover a known configuration. They found that their program was successful in recovering a known configuration when the error level was zero, even when a nonlinear distortion of the distances had been employed. They also found that the addition of fairly large amounts of error did not degrade the recovery of the configuration too much, although recovery of individual subject weights could be quite badly affected. This is not a very surprising finding, since the configuration is estimated on the basis of the data from all of the subjects, whereas the individual weights have to be estimated from a single subject's data. Their experiment is nonetheless a useful demonstration of the phenomenon.

MacCallum and Cornelius (1977) reported the first large-scale study examining the performance of a three-way algorithm. Using ALSCAL, they found that as in two-way scaling better recovery was associated with an increase in the number of points or a decrease in the error level. Also, if the number of true dimensions was low, then, all other things being equal, the recovery was better. Also, as in studies of two-way scaling, MacCallum and Cornelius (1977) found that the stress-like goodness-of-fit statistic was not a very good indicator of whether the configuration and weights had been well recovered. Reconstruction could be highly satisfactory, and yet the goodness of fit of the representation to the data could be poor. This can happen with any scaling program when the error level is high; unfortunately, stress can be high and the recovery also poor, and there is no good way of distinguishing the two situations. However, if the number of stimuli is quite large and the data are complete, there is a better chance of success.

The major surprising finding of the MacCallum and Cornelius (1977) study was that the number of subjects used seemed to have very little effect upon recovery. They simulated the use of 15, 25, 35, and 50 subjects and found that when the other factors were held constant, the configuration was equally well recovered with 15 as with 50 subjects. This is somewhat counterintuitive, since it might be imagined that with more data, recovery would be better. It may be, however, that the number of subjects can have a pronounced effect if it is smaller than 15. The effect of adding more subjects may reach an asymptote rather quickly, but since neither MacCallum and Cornelius, nor anyone else, has explored this further, the exact nature of the phenomenon, or how it might depend on other factors, is not known. In any event, their results seem to indicate that 15 subjects may be a sufficient number in most three-way scaling situations.

Program Comparisons

Very little has been done in this area. Ramsay (1977) conducted a small experiment that seemed to show that although there was not much difference in the recovery of the configuration, MULTISCALE performed better than INDSCAL in recovering the subject weights. In the same paper he also showed that on the basis of monte carlo results the asymptotic chi-square test for dimensionality in MULTISCALE required some adjustment. Ramsay (1980) published such an adjustment based on the results of a small monte carlo experiment.

Incomplete Design Studies

The sole attempt to examine the problem of missing data is by MacCallum (1978). In this study he manipulated the fraction of missing data when the observations were missing at random. Thus, no effort was made to exploit standard experimental designs such as partially balanced incomplete block designs (cf. cyclic designs, discussed above). Using ALSCAL, MacCallum varied the number of subjects, the level of error, and the fraction of missing data, based on a known configuration of 30 points in three dimensions. He found that very good recovery was possible with up to 60% of the data missing. He further investigated whether it made any difference to have the same or different stimulus pairs missing for each subject and found, at least up to about 40% missing data, that it did not. Above 40%, it seems that it may be better to have different judgments missing if the number of subjects is small.

Demonstrations

MacCallum (1977) performed a very interesting demonstration experiment to illustrate a problem with the interpretation—and possible subsequent analysis—of the subject weights obtained by a three-way scaling program. Either implicitly, or explicitly at the user's option, the data that are input to a three-way scaling analysis are said to be ''conditional'' or ''unconditional.'' If the data are conditional, observations cannot be meaningfully compared across individuals; they may, for example, differ in scale. If the data are unconditional, then each subject's dissimilarity matrix is assumed to be directly comparable with any other. Depending on the assumption invoked, a three-way scaling performs the operation known as ''normalization'' somewhat differently; and this has a direct effect upon the weights in the solution. If the data are assumed to be conditional, and the subsequent normalization is done separately for each subject's data, then the weights for different subjects cannot be legitimately compared. No such problem arises when the data are assumed unconditional. Some programs (e.g., ALSCAL) allow the user to specify the normalization desired, whereas others (e.g., INDSCAL) perform separate normalizations for each subject, thus implicitly assuming the data to be conditional. Most of the empirical examples in the literature where weights have been compared across subjects have used INDSCAL, and so it appears that, even now, many people are not aware of the difficulty. They should consult MacCallum (1977) for a full description and some suggestions.

The ALSCAL program is capable of allowing a variety of relations between the input data and the distances in the representation. In addition to the usual linear transformations, ALSCAL permits ordinal and category-preserving transformations. A fairly large study by Young and Null (1978) showed that ALSCAL can perform well in recovering a known configuration if the level of measurement is known, and even if it is not, ALSCAL may be useful in helping to determine the level of measurement and to recover the underlying structure in the data.

Practical Advice

Choosing a Program

It is difficult to make firm recommendations with respect to the choice of programs, as it is in any area where change is rapid. Also, different users are likely to find certain features of a given program compatible, whereas others may find the same characteristics objectionable. For example, in the area of statistical packages, it is this author's opinion that SAS is the ''best buy'' from a number of points of view, including the very important ones of file-handling capability and employing numerically sound algorithms. Should a long-time SPSS user, who is happy with the system, make the switch? There may, indeed, be several advantages to using SAS, but ultimately the user must decide what the costs and benefits of conversion will be. Likewise, there would be hesitation in trying to convince the sophisticated user of GENSTAT to give up the many advantages of that package for the increased ''user friendliness'' of the SAS package and its documentation. Finally, whatever this author's opinion might be in 1983, there is no guarantee that the same one will be held in 1985.

The following comments, then, will be general and will recommend only widely available routines of 1983 vintage that perform well under most common circumstances. By and large, problems of implementation, documentation, and use will not be considered unless there is something quite unusual that requires comment.

Two-way scaling. For two-way scaling, users might first consider the possibility of using a traditional metric scaling program based on the work of Young and Householder (1938) and Torgerson (1958). Unfortunately, no single well-packaged routine is commercially available. Many institutions have such programs, but they are usually of the home-grown variety. Based mainly on the work of Torgerson (1958) or Gower (1966), such a routine is easy to program, easy to use, rapid computationally, economic of storage, and immune to the local minimum and degeneracy problems that can bedevil the nonmetric counterparts. If the user has access to such a program, its use should be given some serious consideration. In most cases it will do just as good a job as the nonmetric programs—and cheaper! However, it should be noted that if there is reason to suspect nonlinearity in the data, or excessive error, or outlying observations, it may be worth considering other options.

A sophisticated and flexible metric routine is available in the MULTISCALE package (Ramsay, 1978). This is based on the maximum likelihood method of estimation and operates quite differently from the traditional procedure discussed above. It is potentially susceptible to the problems suffered by other iterative routines, although since very little monte carlo work has been done, it is difficult to know how much of a problem this might be in practice. The procedure does possess many advantages: statistical hypotheses (e.g., regarding dimensionality) may be tested, since a specific probability model is assumed for the errors; the input and output options are extensive; and nonlinear relationships between the data and the distances are permitted so long as they come from a power family. Additionally, since the procedure operates on the data directly, it will be less sensitive than traditional metric scaling to large errors in the data.

If a nonmetric scaling routine is desired, the author's choices, in order of preference, would be KYST, MINISSA-I, and SSA-I. However, as noted before, the differences are minor: They all do the same job, and do it quite well. KYST is probably the most flexible, but this is paid for in terms of slightly increased computing costs. If TORSCA-9 is available, this will serve almost as well. M-D-

SCAL has its considerable strengths also, but the user should be aware that it is a bit more prone to the local optimum problem, especially in one dimension. Most other nonmetric routines have not been kept up to date by their authors and, in any case, are less readily available. As noted previously, ALSCAL may not be a wise choice for two-way scaling, since it uses the squares of the data and is thus more readily affected by large errors in the data.

Once a program has been chosen, the user will have to make additional decisions regarding the setting of the operating parameters of the program. In most cases this is not likely to cause much trouble, since the authors of all widely used programs have provided reasonable default values for these parameters. In the vast majority of cases, these will serve well. However, there is one choice that may perplex the user. Shepard (1974, 1980), following Arabie (1973), has advised that whenever nonmetric scaling programs are employed, the user should make *several* runs with the same data but should start the iterative process from a different random configuration each time. The purpose of this expensive exercise is to help ensure that suboptimal solutions are not obtained; the theory behind the strategy is based on the idea that if 20 or more attempts are made, one of them is bound to be better than the result of using a single systematically chosen configuration.

This point of view is not supported by empirical evidence (Lingoes & Roskam, 1973; Null & Young, 1978; Spence & Young, 1978), nor is it shared by the authors of all publicly distributed computer programs where a rational starting configuration has been provided as the *default* option. There are various ways of providing a program with a good starting position; most are based on some modification of the classical Young-Householder-Torgerson approach (Torgerson, 1958; Young & Householder, 1938). As Spence and Young (1978) have shown in a detailed review of several monte carlo studies, a rational start is much superior to a single random start; furthermore, the results of Lingoes and Roskam (1973) and of Null and Young (1978)—summarized in Spence (1979)—show that the use of multiple random starts is less effective than the use of a rational start. Thus, it seems that it is not necessary to invest several times as many computer dollars to guard against being trapped in a suboptimal position. The default rational starting configuration offered by all current programs will serve very well.

Three-way scaling. In the three-way case it is more difficult to give good advice regarding the choice of a program. INDSCAL, or its computationally more efficient relatives, SUMSCAL and SINDSCAL (de Leeuw & Pruzansky, 1978), probably continues to be most popular, although ALSCAL is now very widely used, especially since its incorporation into SAS. ALSCAL is probably more expensive to use in the majority of cases, especially by comparison to more recent versions of INDSCAL such as SINDSCAL, but is more flexible from a number of points of view. However, this very flexibility places a greater onus on the user to become familiar with the consequences of exercising the various options and to understand thoroughly the properties of the models that are implemented in ALSCAL. This is one program that should not be used casually; the user should have read the papers in the collection edited by Young and Lewyckyj (1981) at the very least. Apart from MULTISCALE (Ramsay, 1978), most of the other three-way scaling programs are not very widely used as yet. This may change—especially if monte carlo studies that reveal strengths and shortcomings are published.

Distinguishing Noise from Structure

This is easier to do in the two-way case. The use of tables, graphs, or analytic approximations (as earlier in this paper) will quickly give the user an idea of the worth of the data. It should be borne in mind that this may only be done in the case of complete data: No one has so far provided comparable monte carlo results for the situation in which some of the data are missing. Incidentally, if M-SPACE (Spence & Graef, 1974) is used to help determine the appropriate dimensionality, random rankings STRESS values are automatically exhibited as part of the graphical output, as well as an indication of the level of error in the data. Even without the use of such aids, a plot of STRESS versus dimensionality is often revealing. If there is a clear elbow in the graph, it is probable that the data contain only a low or moderate level of error. On the other hand, if no such discontinuity is present, it may be that the error variance is quite large. This comment also applies to other goodness-of-fit statistics.

In three-way scaling no comparable aids are available. However, the device of plotting the goodness-of-fit statistics against dimensionality may be employed. Whether this is a correlation-like statistic (such as in INDSCAL) or a sum of deviation squares (as in ALSCAL) makes very little difference. It should be noted that this could be done individually for each subject. If Ramsay's (1977) MULTISCALE procedure is used, it should be remembered that a maximum likelihood estimate of the error variance is obtained as part of the process of model fitting. The user's manual (Ramsay, 1978) should be consulted regarding interpretation.

Determining Dimensionality

The most important piece of advice that can be given here is that this should not be a mechanical procedure. By all means examine plots of STRESS versus dimensionality or use M-SPACE; but finally, other considerations must be given great weight. Probably interpretability is the most important one; and this does not mean staring at coordinates or dimension by dimension plots and then generating fanciful stories based upon the ordering of stimuli. Preferably some recourse should be had to external analysis. The use of regression or canonical correlation techniques to relate the dimensions of the space to the results of obtaining unidimensional scales on a number of clearly defined variables is highly recommended. Good discussions of this are to be found in Kruskal and Wish (1977) and Davison (1983). This may be done whether two- or three-way scaling has been employed. If Ramsay's (1977) procedure has been used, there is a statistical test for dimensionality, but even in this case collateral analysis and attention to interpretability is desirable.

Another possibility is to match the obtained configuration to a hypothesized target configuration by the use of an orthogonal Procrustes procedure (e.g., Davison, 1983; Schonemann & Carroll, 1970). Examination of the goodness-of-fit statistics over several different recovered dimensionalities with the same fixed dimensioned target may be revealing. Finally, Shepard (1974) has given some excellent general advice on the question of determining dimensionality.

Large Experiments

At least three approaches may be taken. The best will depend on individual circumstances. The first involves attempts to avoid the problem of having to collect pairwise judgments. One possible technique is the method of sorting, of which there exist several variants. A typical task consists in having the subjects sort the stimuli into a number of homogeneous categories and then computing one of several indices of similarity prior to entry to a multidimensional scaling program. Very little systematic monte carlo experimentation has been done to assess the strengths and limitations of this method. Drasgow and Jones (1979) have done work suggesting that this method may not be as effective as some have claimed. Undoubtedly, in a few instances, it has been applied successfully, but there is enough in the way of anecdotal evidence to convince most people that use of this method can lead to problems, especially if the error level is high or the similarity matrix is computed on the basis of only a few subjects.

The interactive approach (Cliff et al., 1977; Young & Cliff, 1972) seems to have much to recommend it. Simulations have shown that interactive scaling programs can perform very well and, indeed, may be superior to the use of incomplete experimental designs that are fixed a priori. This is perhaps not surprising, since the fixed experimental designs do not take into account the nature of the data, whereas an interactive program is able to modify its choice of pairs dependent on the subject's previous responses. The major problem with the approach is probably a logistic one; access to timesharing facilities or to a dedicated computer system is essential; this may preclude testing in the field or the simultaneous acquisition of data from a large number of subjects. Also, it must be remembered that the data have to satisfy the restriction of being, at least approximately, Euclidean distances.

If an incomplete experimental design is to be used, the most important thing is to have a design with high global connectedness. Cyclic designs, or even random designs, will usually perform well. Construction of these is generally quite easy, with or without the aid of a computer; but perhaps the simplest effective approach is the paper-and-pencil method described by Spence (1982a, p. 39), which is even easier than constructing a random design. Whenever the design is incomplete, it is wise to collect as much data as possible to guard against the effects of rather high levels of error. Although the minimum fractions recommended by Spence (1982a) will usually be sufficient, collection of more data is to be strongly recommended if at all possible. If incomplete data are to be used with a program that fits the weighted Euclidean model, then it is comforting to know that it is not necessarily advantageous to use different experimental designs with each subject (MacCallum, 1978). All subjects may therefore complete the same experimental materials, thus saving a great deal in their preparation.

References

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey. J. W. *Robust estimates of location: Survey and advances*. Princeton NJ: Princeton University Press, 1972.

Arabie, P. Concerning Monte Carlo evaluations of nonmetric multidimensional scaling algorithms. *Psychometrika*, 1973, *38*, 607–608.

Baker, R. F., & Young, F. W. A note on an empirical evaluation of the ISIS procedure. *Psychometrika*, 1975, *40*, 413–415.

Bloxom, B. Individual differences in multidimensional scaling (ETS RB 68–45). Princeton NJ: Educational Testing Service, 1968.

Bloxom, B. Constrained multidimensional scaling in *N* spaces. *Psychometrika*, 1978, *43*, 397–408.

Carroll, J. D., & Chang, J. J. Analysis of individual differences in multidimensional scaling via an *N*-way generalization of ''Eckart-Young'' decomposition. *Psychometrika*, 1970, *35*, 283–319.

Chan, D. W., & Jackson, D. N. Implicit theory of psychopathology. *Multivariate Behavioral Research*, 1978, *14*, 3–19.

Cliff, N., Girard, R., Green, R. S., Kehoe, J. F., & Doherty, L. M. INTERSCAL: A TSO FORTRAN IV program for subject computer interactive multidimensional scaling. *Educational and Psychological Measurement*, 1977, *37*, 185–188.

Cohen, H. S., & Jones, L. E. The effects of random error and subsampling of dimensions on recovery of configurations by non-metric multidimensional scaling. *Psychometrika*, 1974, *39*, 69–90.

Cox, D. R. *Planning of experiments*. New York: Wiley, 1958.

Coxon, A. P. M., & Jones, C. Multidimensional scaling: Exploration to confirmation. *Quality and Quantity*, 1980, *14*, 31–73.

Davies, P., & Coxon, A. P. M. (Eds.) *Key texts in multidimensional scaling*. London: Heinemann, 1982.

Davison, M. L. *Multidimensional scaling*. New York: Wiley-Interscience, 1983.

Day, G. S., Deutscher, T., & Ryans, A. B. Quality, level of aggregation, and nonmetric multidimensional scaling solutions. *Journal of Marketing Research*, 1976, *13*, 92–97.

de Leeuw, J., & Pruzansky, S. A new computational method to fit the weighted Euclidean distance model. *Psychometrika*, 1978, *43*, 479–490.

Dempster, A. P., Schatzoff, M., & Wermuth, N. A simulation study of alternative to ordinary least squares. *Journal of the American Statistical Association*, 1977, *72*, 77–91.

Drasgow, F., & Jones, L. E. Multidimensional scaling of derived similarities. *Multivariate Behaviorial Research*, 1979, *14*, 227–244.

Girard, R., & Cliff, N. A Monte Carlo evaluation of interactive multidimensinal scaling. *Psychometrika*, 1976, *41*, 43–64.

Gleason, T. C. *A general model for nonmetric multidimensional scaling* (Michigan Mathematical Psychology Program Report MMPP 67-3). Ann Arbor MI: University of Michigan, 1967.

Golledge, R. G., & Raynor, J. N (Eds.) *Proximity and preference: Problems in the multidimensional analysis of large data sets*. Minneapolis: University of Minnesota Press, 1982.

Gower, J. C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 1966, *53*, 325–338.

Graef, J., & Spence, I. Using distance information in the design of large multidimensional scaling experiments. *Psychological Bulletin*, 1979, *86*, 60–66.

Graunt, J. *Natural and political observations mentioned in a following index, and made upon the bills of mor-*

tality. (Edited by W. F. Willcox). Baltimore MD: The Johns Hopkins Press, 1939. (Originally published, London, 1662).

Green, R. S., & Bentler, P. M. Improving the efficiency and effectiveness of interactively selected MDS data designs. *Psychometrika*, 1979, *44*, 115–119.

Guttman, L. A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, 1968, *33*, 469–506.

Hamer, R. *Nonmetric interactive scaling with multiple subjects*. Unpublished doctoral dissertation, University of North Carolina, Chapel Hill, 1978.

Harshman, R. A. Foundations of the PARAFAC procedure: Models and conditions for an ''explanatory'' multi-model factor analysis. *UCLA Working Papers in Phonetics*, 1970, *16*, 1–84.

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

Horan, C. B. Multidimensional scaling: Combining observations when individuals have different perceptual structures. *Psychometrika*, 1969, *34*, 139–165.

Isaac, P. D. Considerations in the selection of stimulus pairs for data collection in multidimensional scaling. In R. G. Golledge & J. N. Rayner (Eds.), *Proximity and preference: Problems in the multidimensional analysis of large data sets*. Minneapolis: University of Minnesota Press, 1982.

Isaac, P. D., & Poor, D. D. S. On the determination of appropriate dimensionality in data with error. *Psychometrika*, 1974, *39*, 91–109.

Kennedy, W. J., & Gentle, J. E. *Statistical computing*. New York: Marcel Dekker, 1980.

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

Knuth, D. E. *The art of computer programming*: Vol. 2. Seminumerical algorithms. Reading MA: Addison-Wesley, 1969.

Krane, W. R. Least squares estimation of individual differences in multidimensional scaling. *British Journal of Mathematical and Statistical Psychology*, 1978, *31*, 193–208.

Krane, W., & Spence, I. *An evaluation of algorithms for analyzing individual differences in multidimensional scaling*. Paper presented at the 22nd International Congress of Psychology, Leipzig, GDR, July 1980.

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to be a nonmetric hypothesis. *Psychometrika*, 1964, *29*, 1–27. (a)

Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 1964, *29*, 115–129. (b)

Kruskal, J. B., & Wish, M. *Multidimensional scaling*. Beverly Hills CA: Sage Publications, 1977.

Kruskal, J. B., Young, F. W., & Seery, J. B. *How to use KYST-2, a very flexible program to do multidimensional scaling and unfolding*. Murray Hill NJ: Bell Laboratories, 1978.

Lewandowsky, S., & Spence, I. *The robustness of two-way scaling algorithms*. Paper presented at the annual meeting of the Psychometric Society, Los Angeles CA, June 1983.

Lingoes, J. C., & Roskam, E. E. A mathematical and empirical analysis of two multidimensional scaling algorithms. *Psychometrika Monograph Supplement*, 1973, *38* (4, Pt. 2, Monograph No. 19).

MacCallum, R. C. Effects of conditionality on INDSCAL and ALSCAL weights. *Psychometrika*, 1977, *42*, 297–305.

MacCallum, R. C. Recovery of structure in incomplete data by ALSCAL. *Psychometrika*, 1978, *44*, 69–74.

MacCallum, R. C., & Cornelius, E. T. A Monte Carlo investigation of recovery of structure by ALSCAL. *Psychometrika*, 1977, *42*, 401–428.

Miller, E. H. Walrus ethology. I. The social role of tusks and applications of multidimensional scaling. *Canadian Journal of Zoology*, 1975, *53*, 590–613.

Null, C. H., & Sarle, W. *Robust multidimensional scaling*. Paper presented at a joint meeting of the Psychometric and Classification Societies, Montreal, P.Q., Canada, June 1982.

Null, C. H., & Young, F. W. *A Monte Carlo investigation of initial configuration strategies in KYST*. Paper presented at the European Meeting on Psychometrics and Mathematical Psychology, Uppsala, Sweden, June 1978.

Ramsay, J. O. Maximum likelihood estimation in multidimensional scaling. *Psychometrika*, 1977, *42*, 241–266.

Ramsay, J. O. *MULTISCALE: Four programs for multidimensional scaling by the method of maximum likelihood*. Chicago: National Educational Resources, 1978.

Ramsay, J. O. Some small sample results for maximum likelihood estimation in multidimensional scaling. *Psychometrika*, 1980, *45*, 141–146.

Roskam, E. E., & Lingoes, J. C. MINISSA-I: A FORTRAN IV (G) program for the smallest space analysis of square symmetric matrices. *Behavioral Science*, 1970, *15*, 204–205.

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

Schönemann, P. H. An algebraic solution for a class of subjective metrics models. *Psychometrika*, 1972, *37*, 441–451.

Schönemann, P. H., & Carroll, R. M. Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika*, 1970, *35*, 245–256.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with 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,* 219–245. (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. Metric structures in ordinal data. *Journal of Mathematical Psychology,* 1966, *3,* 287–315.

Shepard, R. N. Representation of structure in similarity data. *Psychometrika,* 1974, *39,* 373–421.

Shepard, R. N. Multidimensional scaling, tree-fitting, and clustering. *Science,* 1980, *210,* 390–398.

Sherman, C. R. Nonmetric multidimensional scaling: A Monte Carlo study of the basic parameters. *Psychometrika,* 1972, *37,* 323–355.

Spector, A. N., & Rivizzigno, V. L. Sampling designs and recovering cognitive representations of an urban area. In R. G. Golledge & J. N. Rayner (Eds.), *Proximity and preference: Problems in the multidimensional analysis of large data sets.* Minneapolis: University of Minnesota Press, 1982.

Spence, I. *Multidimensional scaling: An empirical and theoretical investigation.* Unpublished doctoral dissertation, University of Toronto, Toronto, 1970.

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

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

Spence, I. A simple approximation for random rankings stress values. *Multivariate Behavioral Research,* 1979, *14,* 355–365.

Spence, I. Incomplete experimental designs for multidimensional scaling. In R. G. Golledge & J. N. Rayner (Eds.), *Proximity and preference: Problems in the multidimensional analysis of large data sets.* Minneapolis: University of Minnesota Press, 1982. (a)

Spence, I. *Robust multidimensional scaling.* Paper presented at a joint meeting of the Psychometric and Classification Societies, Montreal, P.Q., Canada, June 1982. (b)

Spence, I., & Domoney, D. W. Single subject incomplete designs for nonmetric multidimensional scaling. *Psychometrika,* 1974, *39,* 469–490.

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

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

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

Student. The probable error of a mean. *Biometrika,* 1908, *6,* 1–25.

Takane, Y., Young, F. W., & de Leeuw, J. Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika,* 1977, *42,* 1–67.

Tansley, B. W., & Boynton, R. M. Chromatic border perception—role of red-sensitive and green-sensitive cones. *Vision Research,* 1978, *18,* 683–697.

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

Tschudi, F. The latent, the manifest, and the reconstructed in multivariate reduction models: A study of multidimensional scaling and similarity data *(FSBN 82-569-0046-6).* University of Oslo, Institute of Psychology, Norway, 1972.

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

Young, F. W. Nonmetric multidimensional scaling: Recovery of metric information. *Psychometrika,* 1970, *35,* 455–473.

Young, F. W., & Cliff, N. Interactive scaling with individual subjects. *Psychometrika,* 1972, *37,* 385–415.

Young, F. W., & Lewyckyj, R. *ALSCAL-4: Collected papers.* Chapel Hill NC: University of North Carolina, Psychometric Laboratory, 1981.

Young, F. W., & Null, C. H. Multidimensional scaling of nominal data: The recovery of metric information with ALSCAL. *Psychometrika,* 1978, *43,* 367–379.

Young, F. W., Null, C. H., Sarle, W. S., & Hoffman, D. L. Interactively ordering the similarities among a large set of stimuli. In R. G. Golledge & J. N. Rayner (Eds.), *Proximity and preference: Problems in the multidimensional analysis of large data sets.* Minneapolis: University of Minnesota Press, 1982.

Young, G., & Householder, A. S. Discussion of a set of points in terms of their mutual distances. *Psychometrika,* 1938, *3,* 19–22.

Acknowledgment

This work was supported by Grant No. A8351 from the Natural Sciences and Engineering Research Council of Canada.

Author's Address

Send requests for reprints or further information to Ian Spence, Department of Psychology, University of Toronto, Toronto, Ontario M5S 1A1, Canada.