PSYCHOMETRIKA—VOL. 34, NO. 3
SEPTEMBER, 1969

A MONTE CARLO INVESTIGATION OF THE STATISTICAL SIGNIFICANCE OF KRUSKAL'S NONMETRIC SCALING PROCEDURE*

DAVID KLAHR

UNIVERSITY OF CHICAGO

Recent advances in computer based psychometric techniques have yielded a collection of powerful tools for analyzing nonmetric data. These tools, although particularly well suited to the behavioral sciences, have several potential pitfalls. Among other things, there is no statistical test for evaluating the significance of the results. This paper provides estimates of the statistical significance of results yielded by Kruskal's nonmetric multidimensional scaling. The estimates, obtained from attempts to scale many randomly generated sets of data, reveal the relative frequency with which apparent structure is erroneously found in unstructured data. For a small number of points (i.e., six or seven) it is very likely that a good fit will be obtained in two or more dimensions when in fact the data are generated by a random process. The estimates presented here can be used as a bench mark against which to evaluate the significance of the results obtained from empirically based nonmetric multidimensional scaling.

1. Introduction

Recent developments in psychological scaling [Shepard, 1962a, 1962b; Kruskal, 1964a, 1964b; Lingoes, 1965] have yielded a collection of computer based procedures for extracting metric results from nonmetric data. Since much of the data obtainable in the behavioral sciences is obtained under circumstances where there is no appropriate metric, these new procedures provide powerful tools for the analysis of such data. They are being used with increasing frequency in a diverse range of applications, e.g., color vision, Morse Code perception [Shepard, 1963], taste [Russell & Gregson, 1966], college admissions [Klahr, 1968], marketing [Green, Carmone & Robinson, 1968].

However, the great strength of these new procedures must be tempered with careful application. Torgerson [1965, p. 381], in reviewing some of the advances in multidimensional scaling, says:

> The new procedures would . . . seem to offer nothing but advantages over the old; to require very little and to yield very much. Yet there are many problems connected with their use which . . . have not been at all obvious . . . . It is like doing a factor analysis. And, like factor analysis, the methods always yield an answer. But it can be even more difficult to fully comprehend the meaning of that answer.

*A preliminary version of this paper was presented at the International Federation for Information Processing Congress 68 in Edinburgh, Scotland, August 5–10, 1968.

Torgerson goes on to discuss several of the difficulties. One which he treats most extensively in his paper is the problem of the very nature of similarity itself. He shows that under certain circumstances the implicit models underlying the scaling technique are quite inappropriate. In particular, the similarity of one item to another often depends upon the context of the judgment, i.e., upon the other relevant items in the collection of things to be judged. The scaling models do not allow for this kind of variability in similarity judgments.

A second difficulty, the problem of multiple solutions (i.e., non-uniqueness of the configuration that provides the best fit to the data), has since been treated by Shepard [1966]. Shepard has demonstrated that for more than eight points we can be very sure that the solution is unique, and for 15 or more points we can be virtually certain that we have found *the* configuration which best fits the data.

Green [1966] suggests that "these . . . procedures need Monte Carlo and other computer runs to determine their properties. . . ." This paper deals with one important aspect of nonmetric multidimensional scaling: the statistical significance of the results. We investigate, through Monte Carlo simulation, the following question: How likely are we to falsely reject the Null Hypothesis that the data to be scaled come from a random generator? There are at present no statistical methods for testing the significance of the results generated by the scaling procedures. As a first step towards filling this void, this paper presents estimates of the relative frequency with which apparent structure (i.e., a good scaling solution) is found in randomly generated data. The results to be presented are based upon a series of attempts to scale randomly generated data. They can be used as a bench mark against which to assess the significance of empirical results.

2. Nonmetric Multidimensional Scaling

All multidimensional scaling procedures assume that there is an underlying structure—a spatial configuration of items—in which interpoint distances are inversely related to empirically determined proximity measures on those items. The goal of the procedures is to construct such a configuration from the inter-item proximity measures. For example, if the proximity measures are judgments of the relative similarity of pairs of stimuli, the scaling procedure would attempt to construct a spatial configuration in which relatively similar stimuli correspond to points in the space that are relatively close together.

The three most widely used procedures—Shepard's [1962], Kruskal's [1964a, 1964b], and Lingoes' [1966]—are designed for the same purposes. That is, they yield essentially similar outputs, given the same input. Although the procedures achieve their desired result in different ways, for the purposes of our analysis it will suffice to discuss only one of them. Therefore, the remainder of this paper will be based upon Kruskal's nonmetric multidimensional scaling procedure [1964a, 1964b].

The goal of the procedure is to find the spatial configuration of a set of points in which the rank order of the interpoint distances is maximally inversely correlated with the rank order of the corresponding inter-item similarity measures. It starts with an arbitrary configuration of points and iteratively attempts to find some arrangement of the points such that the rank order of the interpoint distances is exactly the opposite of the rank order of the similarity measures. When this occurs, we have a configuration that fits the data perfectly. As the dimensionality of the space is reduced and the solution becomes more highly constrained, we are apt to get some departures from perfect fit. Some of the distances may be "out of order." A measure of departure from perfect fit, called the "stress" of the configuration, has been developed by Kruskal [1964a]; it is quite similar to a residual sum of squares. From an extensive series of empirical investigations on a variety of data, Kruskal suggests that departure from perfect fit (stress $= 0$) be interpreted as follows: .025—excellent; .05—good; .10—fair. The usual procedure is to find the best fit—the minimum stress—in spaces of decreasing dimensionality. We expect minimum stress to increase as the dimensionality decreases, starting in $n - 1$ space with zero stress.

The decision as to which configuration is the most appropriate representation of items rests upon scientific judgments and is not a direct output of the scaling technique. In most applications, the decision is based upon the stress, the dimensionality of the space, and the meaningfulness of the final configuration. Most users have used Kruskal's suggestions as to what constitutes an acceptable result in lieu of a rigorous statistical test for the significance of the stress. Although no appropriate statistical test exists at present, it is possible to obtain estimates of significance through the use of the techniques described in the next section.†

3. *Procedure*

In all applications of multidimensional scaling techniques the input consists of some measure of the proximity of each stimulus to all other stimuli. Whatever the metric for the raw data, the scaling procedure makes use of only the *rank order* of proximity measures. For this reason it is possible to simulate a wide range of empirical proximity measures on n stimuli simply by using a set of distinct values, one for each of the $n(n - 1)/2$ pairs of n stimuli.

To the set of $n(n - 1)/2$ distinct pairs of n hypothetical items we ran-

†Since this work was done, a similar study by Stenson and Knoll [1969] has appeared covering a different range of parameters: $m = 1$ to 10 in steps of 1, and $n = 10$ to 60. For large n, there is much less variance in the final stress, and the averages based on 3 data sets are sufficient. For the small n studied in our paper, there is a need for more samples. The two papers cover a wide range of conditions for MDSCAL.

domly assigned values from a uniform distribution on the open interval from 0 to 1. One hundred such sets of random proximities were generated for each value of $n = 6, 7, 8$ and 10; fifty sets each were generated for $n = 12$ and 16. Every one of the 500 sets was scaled by Kruskal's nonmetric multidimensional scaling program in spaces having from 5 to 1 dimensions. The parameter values that control Kruskal's program were: minimum stress sought—zero; maximum number of iterations—75; type of proximity measure—dissimilarities; spatial metric—Euclidean.

4. Results

Usually, the primary intent of scaling is to obtain a "picture" of the final configuration. However in this study, since the configurations are constructed from random data, they are of no intrinsic interest. Instead, we focus attention on the stress of each final configuration, for it is the behavior of stress under conditions of pure noise that this study seeks to portray.

For each value of n we have plotted the cumulative distribution of the final stress values from one, two and three dimensional configurations in Figures 1 through 3 respectively. If an analytic expression could be developed for the statistical behavior of the stress, then it would presumably be capable of generating a family of curves quite similar to these empirical estimates. In Table 1 we present some summary statistics for each value of n. The number of solutions less than or equal to Kruskal's "good" (.05) and "excellent" (.025) stress levels are included, along with the maximum, minimum, average and standard deviation of final stress values. Table 2 contains some selected percentile points from the cumulative distributions of final stress. Figure 4 is a plot of the average final stress shown in Table 1.

5. Discussion

These data may provide some assistance in detecting and avoiding spurious scaling solutions. As Shepard [1966] found in his study of uniqueness, results are extremely sensitive to n, the number of points, when n is low. For example, "good" solutions (i.e., stress $\leq$ .05) are often attainable for 6, 7 or 8 points in 3 dimensions when in fact the proximity measures are randomly generated. (For 6 points, 96 out of 100; for 7 points, 74 out of 100; for 8 points, 33 out of 100.) A small increase in the number of points (e.g., $n \geq 10$) provides a substantial reduction in the likelihood of this kind of bogus solution: none of the cases yields good stress for $n \geq 10$.

The average final stress, shown in Figure 4, is well behaved in two senses; for a given number of points, stress increases with decreasing dimensions; for a given number of dimensions, stress increases with increasing points.

One of the criteria that Kruskal suggests for selecting the appropriate dimensionality is an "elbow" in the plot of stress vs. dimensionality, i.e., a major decrease in the marginal improvement effected by an additional dimension. The lack of distinctive elbows in the plots in Figure 4 arises from the

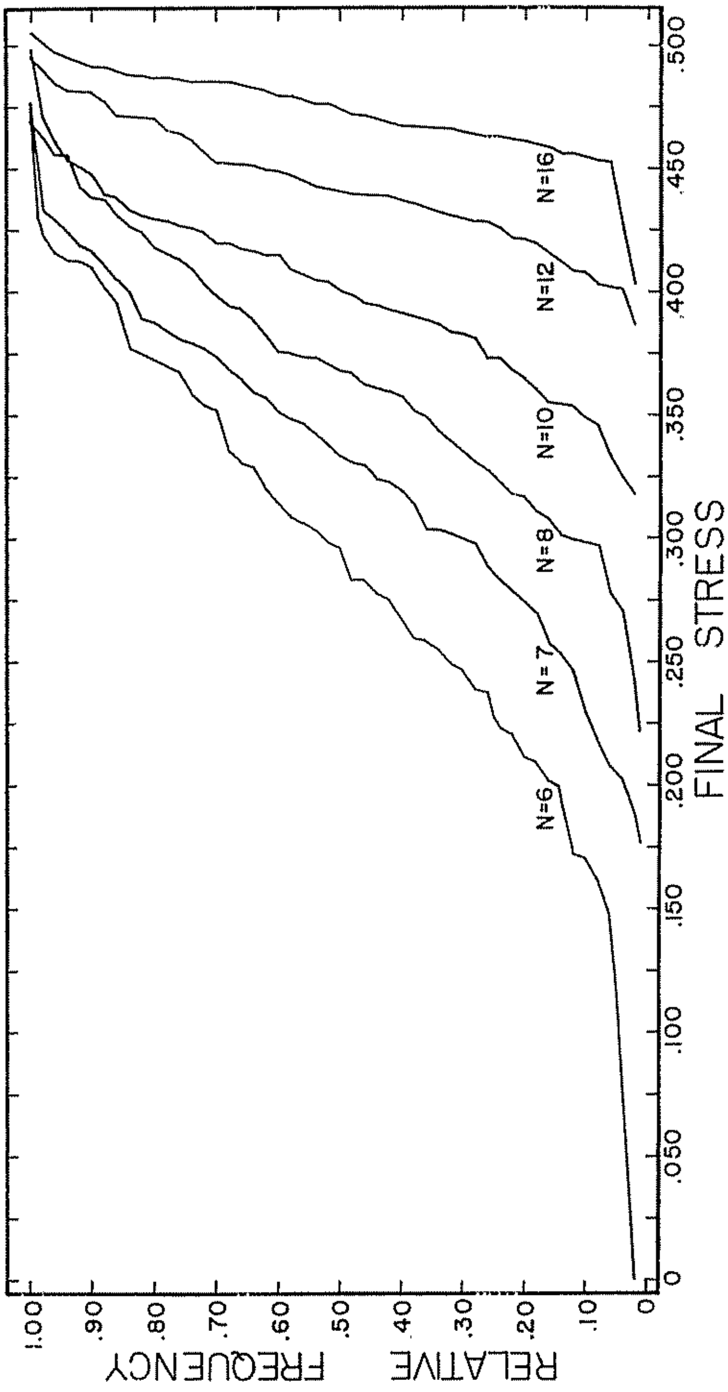

Fig. 1.--Cumulative distributions of relative frequency of final stress in a 1 dimensional configuration for 6, 7, 8, 10, 12, and 16 points.

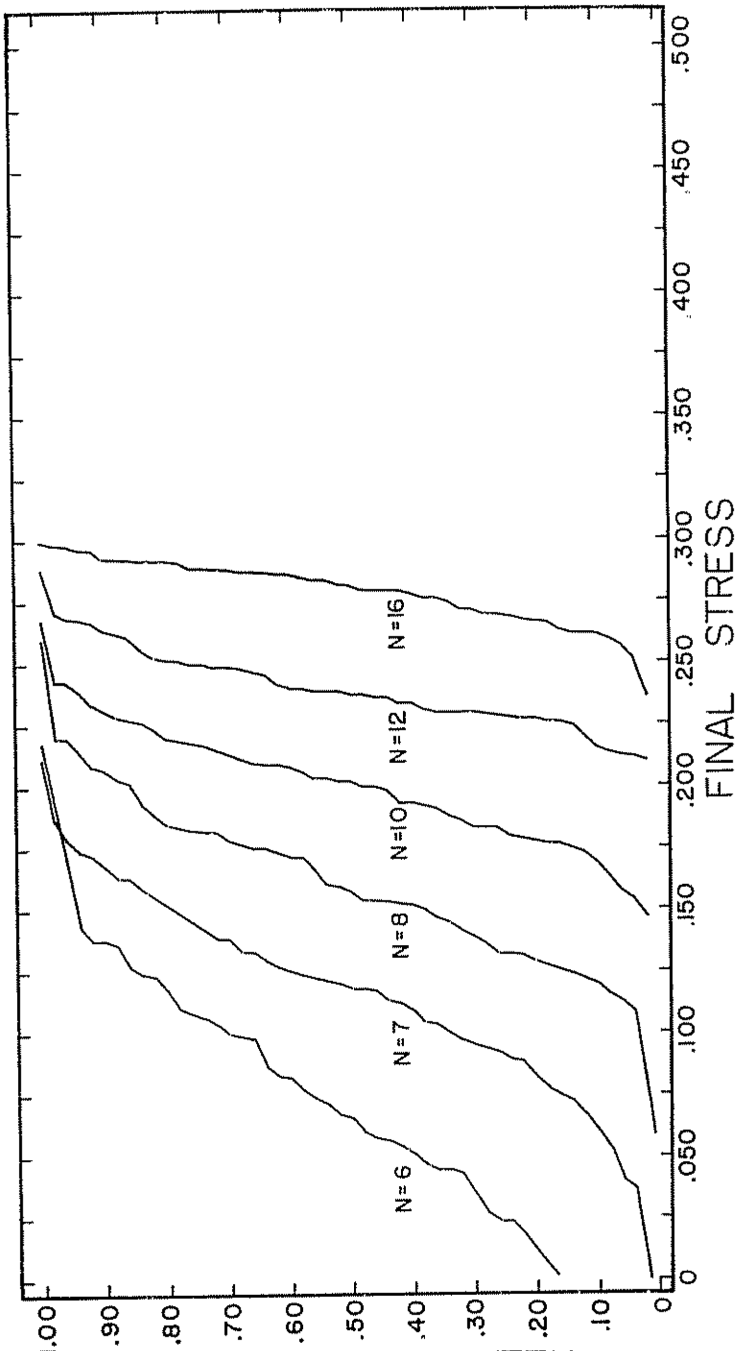

Fig. 2.--Cumulative distributions of relative frequency of final stress in a 2 dimensional configuration for 6, 7, 8, 10, 12, and 16 points.

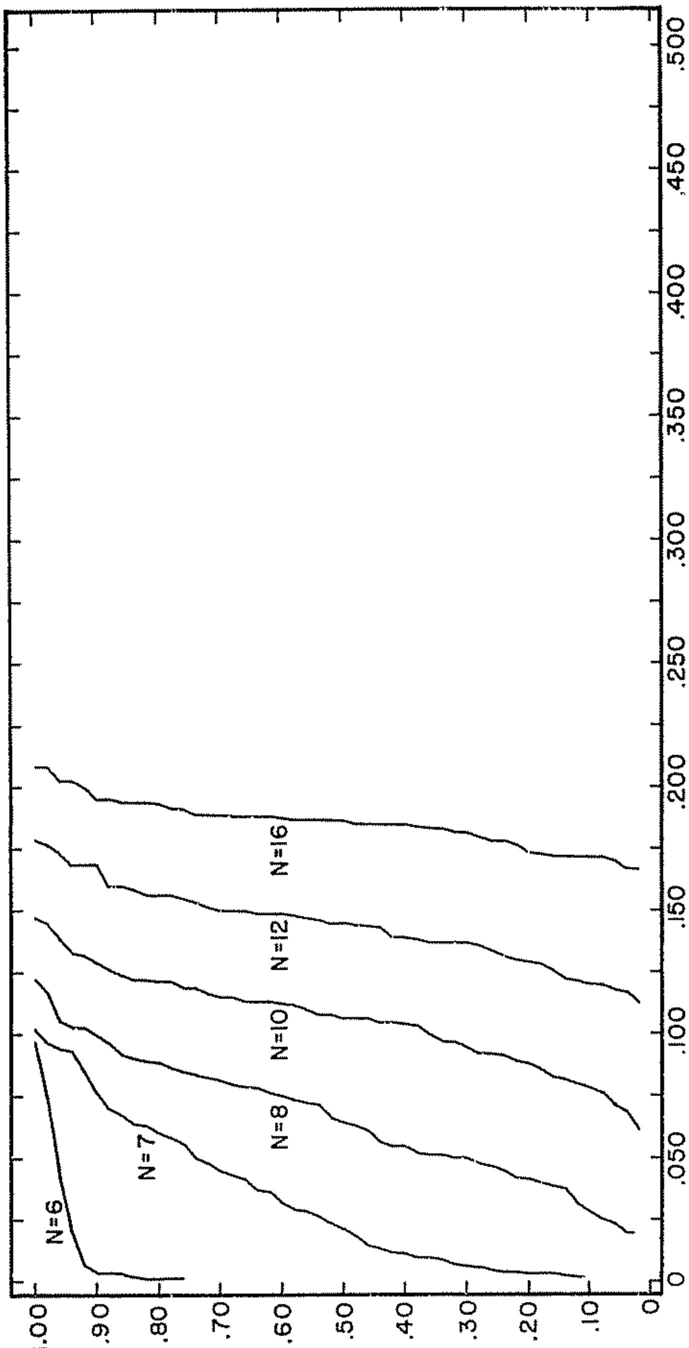

Fig. 3.--Cumulative distributions of relative frequency of final stress in a 3 dimensional configuration for 6, 7, 8, 10, 12, and 16 points.

TABLE 1

Summary Statistics From Monte Carlo Runs: Average, Standard Deviation, Max, Min, Good and Excellent Stress

	Number of Dimensions					
	1	2	3	4	5	
Number of "Good" Final Stresses	4	39	96	100	100	
Number of "Excellent" Final Stresses	4	27	95	100	100	
Average Final Stress	0.288	0.070	0.005	0.	0.	6 points
Standard Deviation of Final Stresses	0.097	0.053	0.017	0.	0.	100 sets
Maximum Final Stress	0.471	0.218	0.096	0.001	0.	
Minimum Final Stress	0.	0.	0.	0.	0.	
Number of "Good" Final Stresses	0	7	74	100	100	
Number of "Excellent" Final Stresses	0	3	53	100	100	
Average Final Stress	0.332	0.114	0.031	0.002	0.	7 points
Standard Deviation of Final Stresses	0.066	0.043	0.030	0.005	0.	100 sets
Maximum Final Stress	0.477	0.212	0.101	0.025	0.001	
Minimum Final Stress	0.177	0.	0.	0.	0.	
Number of "Good" Final Stresses	0	0	33	95	99	
Number of "Excellent" Final Stresses	0	0	8	78	99	
Average Final Stress	0.368	0.160	0.065	0.016	0.002	8 Points
Standard Deviation of Final Stresses	0.055	0.034	0.027	0.018	0.008	100 sets
Maximum Final Stress	0.470	0.260	0.122	0.079	0.073	
Minimum Final Stress	0.222	0.059	0.	0.	0.	
Number of "Good" Final Stresses	0	0	0	36	95	
Number of "Excellent" Final Stresses	0	0	0	3	52	
Average Final Stress	0.400	0.201	0.105	0.055	0.025	10 points
Standard Deviation of Final Stresses	0.038	0.025	0.020	0.017	0.013	100 sets
Maximum Final Stress	0.498	0.267	0.147	0.096	0.057	
Minimum Final Stress	0.314	0.142	0.054	0.008	0.	

TABLE 1--Continued

	Number of Dimensions					
	1	2	3	4	5	
Number of "Good" Final Stresses	0	0	0	1	16	
Number of "Excellent" Final Stresses	0	0	0	0	0	
Average Final Stress	0.444	0.240	0.144	0.088	0.057	12 points
Standard Deviation of Final Stresses	0.025	0.017	0.016	0.016	0.015	50 sets
Maximum Final Stress	0.496	0.288	0.178	0.131	0.089	
Minimum Final Stress	0.387	0.210	0.112	0.049	0.027	
Number of "Good" Final Stresses	0	0	0	0	0	
Number of "Excellent" Final Stresses	0	0	0	0	0	
Average Final Stress	0.473	0.279	0.185	0.130	0.096	16 points
Standard Deviation of Final Stresses	0.018	0.014	0.010	0.011	0.011	50 sets
Maximum Final Stress	0.505	0.300	0.207	0.152	0.127	
Minimum Final Stress	0.404	0.237	0.166	0.106	0.074	

TABLE 2

Selected Percentile Points From Cumulative Distributions of Final Stress

		Percentile							
		5	10	25	50	75	90	95	
Dimensions	1	.121	.171	.227	.296	.363	.410	.415	6 points 100 sets
	2	.000	.000	.023	.065	.107	.138	.148	
	3	.000	.000	.000	.000	.000	.004	.020	
	4	.000	.000	.000	.000	.000	.000	.000	
	5	.000	.000	.000	.000	.000	.000	.000	
Dimensions	1	.208	.230	.285	.334	.381	.417	.427	7 points 100 sets
	2	.040	.060	.086	.118	.142	.168	.176	
	3	.000	.000	.004	.021	.056	.077	.094	
	4	.000	.000	.000	.000	.000	.006	.009	
	5	.000	.000	.000	.000	.000	.000	.000	
Dimensions	1	.276	.298	.326	.396	.412	.439	.456	8 points 100 sets
	2	.111	.120	.132	.157	.183	.206	.215	
	3	.022	.028	.046	.064	.084	.100	.104	
	4	.000	.000	.002	.010	.023	.042	.050	
	5	.000	.000	.000	.000	.001	.005	.007	
Dimensions	1	.333	.349	.373	.401	.427	.448	.461	10 points 100 sets
	2	.153	.169	.180	.202	.216	.232	.246	
	3	.068	.078	.092	.107	.119	.129	.136	
	4	.027	.035	.044	.054	.067	.077	.079	
	5	.004	.009	.018	.024	.032	.040	.050	
Dimensions	1	.402	.408	.429	.441	.465	.481	.485	12 points 50 sets
	2	.211	.215	.228	.237	.250	.264	.269	
	3	.118	.120	.134	.145	.153	.169	.174	
	4	.071	.072	.075	.086	.094	.109	.118	
	5	.032	.037	.047	.057	.063	.078	.085	
Dimensions	1	.454	.455	.463	.474	.486	.492	.498	16 points 50 sets
	2	.257	.261	.270	.282	.289	.293	.298	
	3	.170	.171	.178	.186	.191	.195	.203	
	4	.111	.118	.121	.129	.138	.144	.147	
	5	.077	.083	.089	.096	.102	.112	.119	

averaging of many curves, some with quite distinctive elbows at different dimensionalities. Thus Figure 4 should not obscure the fact that pure noise may give the spurious appearance of a true dimensionality.

The importance of these findings to a user of Kruskal's program rests upon the manner in which the scaling procedure is being used. If it is being used to test *a priori* hypotheses about the dimensionality or spatial arrange-

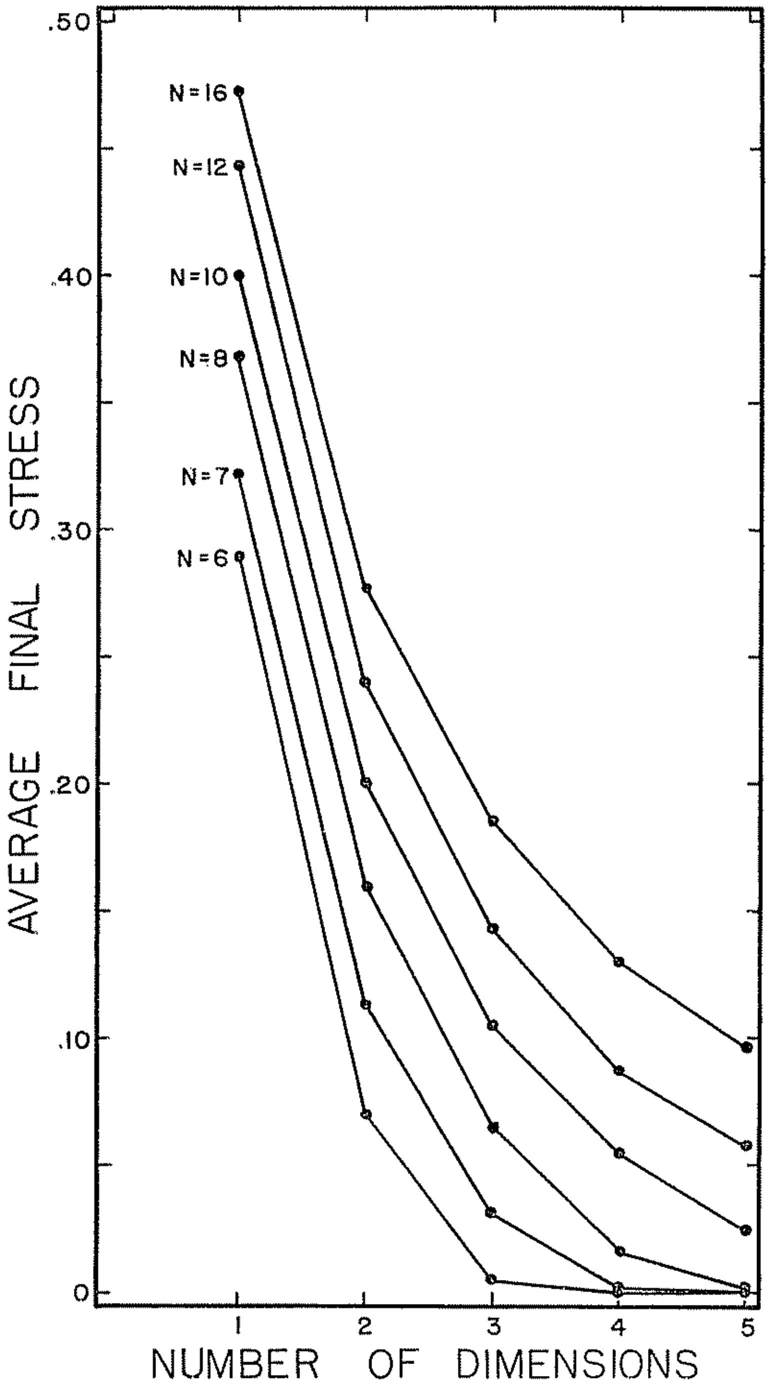

Fig. 4.--Average final stress vs. number of dimensions for 6, 7, 8, 10, 12, and 16 points.

ment of a stimulus set, then the significance criterion is but one of many pieces of evidence that can be used in interpreting results. However, if the scaling procedure is used in an exploratory study, where there are no *a priori* notions about the configuration, then any results for only 8 or 9 points in 2 or 3 dimensions cannot be considered very convincing evidence for the existence of structure. For example, from Figure 3 it is evident that with eight points there are about 2 chances in 3 that pure noise could be accounted for in 3 dimensions with a stress less than .075. If the experimenter is at the design stage, then he can use these results as a lower bound on the number of stimuli he will need in a situation where he suspects a certain number of dimensions.

6. Conclusion

The scaling procedure discussed in this paper is one of a class of new techniques that are so powerful and convenient that they are being used as exploratory devices to see if any structure exists in a set of proximity measures. If n is small, and if a low stress constitutes the only evidence of structure, then any results may be meaningless. The estimates of significance presented here can be used as a bench mark against which to assess the meaningfulness of a wide variety of multidimensional scaling applications.

REFERENCES

Green, B. F. The computer revolution in psychometrics. *Psychometrika*, 1966, **31**, 437–445.

Green, P. E., Carmone, F. J., & Robinson, P. S. Nonmetric scaling: An exposition and overview. *Wharton Quarterly*, 1968, **3**.

Klahr, D. Decision making in a complex environment. *Management Science*, 1969 (forthcoming).

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

Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 1964, **29**, 28–42. (b)

Lingoes, J. C. An IBM 7090 program for Guttman-Lingoes smallest space analysis—I. *Behavioral Science*, 1965, **10**, 183–184.

Russell, P. N., & Gregson, R. A. A comparison of intermodal and intramodal methods in multidimensional scaling of three-component taste mixtures. *Australian Journal of Psychology*, 1967, **18**, 244–254.

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

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function, II. *Psychometrika*, 1962, **27**, 219–264. (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.

Stenson, H. H., & Knoll, R. L. Goodness of fit for random rankings in Kruskal's nonmetric scaling procedure. *Psychological Bulletin*, 1969, **71**, 122–126.

Torgerson, W. S. Multidimensional scaling of similarity. *Psychometrika*, 1965, **30**, 379–393.

Manuscript received: 1/4/68
Revised manuscript received: 9/13/68