INFORMATION TO USERS

This dissertation was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.
2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.
3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.
4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

University Microfilms

300 North Zeeb Road
Ann Arbor, Michigan 48106
A Xerox Education Company

72-27,085

POOR, David Deas Sinkler, 1945-
INVARIANCE OF MULTIDIMENSIONAL CONFIGURATIONS.

The Ohio State University, Ph.D., 1972
Psychology, general

University Microfilms, A XEROX Company, Ann Arbor, Michigan

INVARIANCE OF MULTIDIMENSIONAL CONFIGURATIONS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy in the Graduate
School of The Ohio State University

By

David Deas Sinkler Poor, B.A., M.A.

* * * * *

The Ohio State University
1972

Approved by
Robert J. Wherry

Advisor
Department of Psychology

PLEASE NOTE:

**Some pages may have
indistinct print.**

Filmed as received.

University Microfilms, A Xerox Education Company

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS.....	ii
VITA.....	iv
LIST OF TABLES.....	vii
LIST OF FIGURES.....	ix
INTRODUCTION.....	1
Chapter	
I. INVARIANCE REVIEWED: APPROACHES FROM FACTOR ANALYSIS.....	6
II. INDEX OF INVARIANCE.....	11
III. TESTING FOR INVARIANCE: SOME EXAMPLES.....	49
CONCLUSIONS.....	62
BIBLIOGRAPHY.....	67

ACKNOWLEDGMENTS

I gratefully acknowledge the consistent encouragement and helpful suggestions given me by the members of my committee, especially that given by my advisor, Professor Wherry. Without his support, this dissertation would never have been completed.

I would also like to thank my many colleagues and friends who gave constructive criticism and suggestions, especially for their encouragement during several times of depression. Special thanks go to Professor Thomas A. Willke from the Statistics Laboratory for his comments on the distribution of distances and for suggesting the use of subsets of data used for Index II (Chapter II). For permission to use their data, I would like to thank Susan Mattingly and Dean Owen. Special thanks go to Professor Paul Isaac, whose investigations into multidimensional scaling inspired this dissertation, for well thought and constructive criticisms, especially of the preliminary draft of this dissertation. For their continued support in supplying resources and tolerating my absence, I wish to thank the Merston Center, especially Professors Snyder and Herman. I owe a large debt of

gratitude to the Instruction and Research Computer Center which supplied welcome suggestions as well as the free computer time for this project.

Most of all, however, for their continued moral support, I am indebted to my wife and children who tolerated my long hours away from home.

VITA

David D. D. Poor

February 24, 1945

EDUCATION

B.A. cum laude in Social Relations, Harvard College,
1967.

M.A. in Social Psychology, The Ohio State University,
1970.

Ph.D. in Quantitative Psychology, The Ohio State
University, 1972.

PUBLICATIONS

Familiarity and attraction to social and non-social ob-
jects by rats. With Bibb Latané, and Lloyd Sloan.
Psychonomic Science, 1972, 26, pp. 171-172.

RATLOVE: Analysis of social behavior from two-dimension-
al position data. Behavioral Science, Computer
Program Abstract, in press.

ACADEMIC AWARDS, STIPENDS, AND APPOINTMENTS

1971-1972 . . Research Associate, Merson Center, The
Ohio State University.

. . Research Associate, The Department of
Psychology, The Ohio State University.

- 1971 . . Invited guest, Multivariate Analysis
Training Session, University of Chicago,
August 1971.
- 1970-1971 . . Research Associate, Community Health
Research Group, College of Pharmacy, The
Ohio State University.
- 1967-1970 . . Merston Fellowship, Merston Center, The
Ohio State University.
- 1966 . . National Science Foundation Undergraduate
Research Participation Grant, Harvard
College.

AREAS OF INTEREST

Computer applications for statistical analysis.

Statistical data analysis.

Research design and implementation: data collection
techniques.

Data reduction and verification.

RESEARCH EXPERIENCE

- 1966-1967 . . Determinants of responses to survey
questionnaires. With Robert Rosenthal.
- 1967-1970 . . Cognitive responses to persuasive situations.
With Anthony Greenwald.

- 1970 . . Effects of enriched housing on social attraction between laboratory rats. With Bibb Latené and Lloyd Sloan.
- 1971-1972 . . Measuring the invariance of multidimensional configurations. With Robert Wherry.
- 1972 . . Comparison of multivariate and univariate analysis of repeated measures data.

LIST OF TABLES

TABLE	CONTENTS	PAGE
1.	SOLUTION CONFIGURATIONS FOR EIGHT POINTS IN TWO DIMENSIONS.....	15
2.	INTER-POINT DISTANCES FROM SOLUTION CON- FIGURATIONS FOR EIGHT POINTS IN TWO DIMENSIONS.....	16
3.	DIFFERENCES BETWEEN EMPIRICAL DENSITY FUNCTION AND ESTIMATED DENSITY FUNC- TION BASED ON EMPIRICAL CHARACTERISTICS.....	32
4.	TESTS FOR THE HYPOTHESIS: $\bar{X} = 0.0$	34
5.	PREDICTION EQUATIONS FOR THE CHARACTERISTICS OF INDEX I.....	35
6.	DIFFERENCES BETWEEN EMPIRICAL DENSITY FUNCTION AND PREDICTED DENSITY FUNC- TION BASED ON PREDICTED CHARACTERISTICS.....	37
7.	DIFFERENCES BETWEEN EMPIRICAL DENSITY FUNCTION AND PREDICTED DENSITY FUNC- TION BASED ON INDEPENDENTLY PREDICTED CHARACTERISTICS.....	38
8.	SIGNIFICANT VALUES FOR INDEX OF INVARIANCE: $p = .005$	40
9.	SIGNIFICANT VALUES FOR INDEX OF INVARIANCE: $p = .01$	41
10.	SIGNIFICANT VALUES FOR INDEX OF INVARIANCE: $p = .02$	42
11.	SIGNIFICANT VALUES FOR INDEX OF INVARIANCE: $p = .05$	43
12.	SIGNIFICANT VALUES FOR INDEX OF INVARIANCE: $p = .10$	44

	PAGE
13. SIGNIFICANT VALUES FOR INDEX OF INVARIANCE: $p = .20$	45
14. INDEX OF INVARIANCE FOR PHONEMIC CONFUSIONS....	52
15. INDEX OF INVARIANCE FOR SIMILARITY OF RANDOM FORMS.....	58
16. INDEX OF INVARIANCE BETWEEN TRUE CONFIGUR- ATION AND SOLUTIONS WITH VARYING ERROR.....	61

LIST OF FIGURES

FIGURE	CONTENTS	PAGE
1.	GEOMETRIC REPRESENTATION OF DISTANCES.....	17
2.	GRAM-CHARLIER FAMILY OF CURVES USING $F'(x)$	28
3.	TRAPEZOIDAL APPROXIMATION OF AREA.....	30
4.	RANDOM FORMS USED.....	55
5.	OBSERVED STRESS FOR SOLUTIONS OF RANDOM FORMS DATA.....	57

INTRODUCTION

The goal of this dissertation is to define a statistical test for the invariance of multidimensional configurations. A multidimensional configuration is a matrix of coordinates in which the rows of the matrix correspond to identifiable points or objects, and the columns correspond to different dimensions.¹ Any entry

1. Matrix notation will be used for brevity in notation. The basic principles of the procedure can be understood without knowledge of matrix manipulations.

in the matrix, c_{ik} , for row i , column k , will hold the coordinate corresponding to the value of point i on dimension k . This paper concerns itself with assessing the invariance, or the similarity, between two multidimensional configuration matrices in which both matrices have rows corresponding to the same set of points, and both have the same number of orthogonal dimensions or columns.

Configuration matrices as defined above are commonplace in statistical analysis. The most common of such matrices are factor loading matrices from factor analytic

solutions and solution matrices from multidimensional scaling. The goal of both statistical processes is to place a number of points in a multidimensional space defined by a small number of dimensions. Although the techniques proposed in this paper are equally applicable to both factor analytic and multidimensional scaling results, the primary impetus for this work has been from research in ordinal multidimensional scaling (ODMS), and the primary exposition of the technique will be in terms of ordinal multidimensional scaling.

When the raw data for an ODMS problem are collected from a single source, the solution, in the form of a multidimensional configuration, can be found from a straightforward application of the particular ODMS algorithm being employed. The basic form of the ODMS solution is the multidimensional configuration matrix, i.e. a solution matrix. ODMS algorithms are not precisely analytic by nature of the iterative approximation process, and may introduce some error into the final solution (e.g. identification of local minima as absolute minima). In addition, data scaled by ODMS algorithms are typically fallible data. Thus, any scaling solution can be assumed to be an estimate of the correct solution, with the observed solution deviating from the correct solution

because of error in the data and the ODMS algorithm being employed.

When, however, the data for an ODMS application are two or more complete sets of data (from independent sources or from different conditions for the same source), the experimenter must decide whether to combine all of the data to yield a single solution, or to scale each set of data separately. Before combining sets of data, the experimenter should be able to show that the data are, in some manner, significantly related. The index of invariance is designed to provide a rational decision rule that indicates whether or not such sets of data should be combined.

The multidimensional configuration of an ODMS solution can be interpreted as a spatial representation of the underlying processes (unusually psychological) from which the raw data were generated. The index of invariance, then, attempts to identify those sets of data which have underlying processes that are significantly related. Since the multidimensional configurations from ODMS procedures represent underlying processes, the index of invariance is based on comparisons of multidimensional configurations which result from applying the ODMS procedure to the several sets of data.

Assuming that two sets of data do, in fact, derive from the same or equivalent underlying processes, and assuming that such processes can be given a geometric representation in a multidimensional configuration matrix, the scaling solutions will still deviate from the "true" solutions because of error in the two solutions. One function of the index of invariance is to measure the magnitude and significance of the relationship between two multidimensional configurations, and thereby estimate the magnitude and significance of the relationship between the processes from which the two sets of data derive.

Two orthogonal multidimensional configurations can be called "invariant" if, for their common set of points, one configuration matrix can be rotated orthogonally to map onto the other configuration matrix. However, the presence of error in the solution matrices will tend to make such a mapping impossible. The index of invariance, therefore, will indicate an "invariant" relationship only when such a mapping is possible. When such a mapping is not possible, the index of invariance will show whether the two configurations are more similar than would be expected by chance.

One important application of the technique will be in providing a quantitative decision rule for combining or

not combining data. In general, when two configuration matrices are found to be significantly related, a new solution, based on the combined data of both sets, may be a better estimate of the underlying processes than either initial solution. (It should be noted that, although two configurations are statistically related, the two initial solution configurations may be better than the composite solution. Such a situation may arise, for example, when there are a large number of points so that a relatively low value for the index is significant.) On the other hand, when two configuration matrices are not significantly related, the solution from the combined data may well be meaningless.

The present paper proposes an "index of invariance" based on the inter-point distances derived from the respective multidimensional configurations. If the null hypothesis of independent underlying psychological processes is rejected, through the significance of the index of invariance, then the two sets of data can be said to be more related than would be expected by chance. In such cases, the data can be combined to yield a single solution which may be a better estimate of the underlying processes than either of the solutions from the two sets of data alone.

CHAPTER I

INVARIANCE REVIEWED: APPROACHES FROM FACTOR ANALYSIS

The question of configuration invariance is not new to the solutions from factor analytic studies. Indeed, a number of researchers have attacked the problem from a number of different approaches. The two primary approaches are (1) the salient variables approach, and (2) the configurative matching approach. The two approaches are generally used for different applications because they serve to answer different questions. Neither approach, however, indicates the significance or non-significance of observed relationships.

Salient Variables

In the salient variables approach, two independent factor solutions are assumed to have some factors in common and some factors that are unique to the particular solution. It is the goal of the salient variables approach to identify, simultaneously, both the common and unique factors. The process is best conceived as a stepwise process to find successive factors common to the two configurations. At each step, a new factor is found

such that the new factor in one configuration is as similar as possible to the same factor in the second configuration, and an index is computed to indicate the similarity between the factor in one and the factor in the other. In practice, however, the entire process is done in a single analytic process which yields factors ordered from the most similar to the least similar.

The experimenter must decide at which value of the index the factors no longer match, i.e. they are unique in each configuration rather than being common (Taylor, 1967). Factors with an index above the selected value are said to be "congruent." The most common index for such similarity is the "coefficient of congruence", but there are problems in its use.

The "coefficient of congruence" is similar to a correlation between the factor loadings for the factor in the two configurations after they have been rotated as described above. The coefficient does not correct for differences in means, however, although it varies from -1.0 to +1.0. If c_{ik} and d_{ik} are the respective factor loadings for variable i on factor k in the two configurations, then the coefficient of congruence for factor k is defined as:

$$\frac{\sum_{j=1}^n c_{jk} d_{jk}}{\left[\left(\sum_{j=1}^n c_{jk}^2 \right) \left(\sum_{j=1}^n d_{jk}^2 \right) \right]^{1/2}}$$

where n is the number of variables (Tucker, 1951).¹ The

1. The coefficient of congruence is presented and discussed by Harman (1967, pp. 269-272), and Rummel (1970, pp. 461-462).

main problem with the coefficient is that the distribution is not known, and therefore, even for a given factor, there can be no simple statement of the probability that the factor is equivalent in the two structures. Indeed, using random data, Nesselroade and Baltes (1970) have shown that the coefficient of congruence, by chance alone, can be expected to have relatively high values. Rules of thumb can sometimes be applied: Norman Cliff suggests that, using correlations between the factors rather than the coefficient of congruence, a "correlation of .75 is minimal if the factors are to have recognizably the same interpretations." (1966, p. 41). Pinneau and Newhouse (1964) present data showing the problems with interpreting the coefficient of congruence or the correlation between factor loadings, and propose the correlation between factor scores as the "coefficient of invariance."

Even for the coefficient of invariance, however, the distribution of the index is not known. One possible resolution to the problem of unknown distributions is to find empirical distributions of the indices to use as base lines in testing obtained coefficients.

Even if the distributions of the coefficients were known, however, the salient variables approach can not directly provide an overall measure of invariance for the two configurations. Such an overall measure is provided by the second approach, that of configurative matching.

Configurative Matching

The second major approach is that of configurative matching, or the "Procrustes" approach. As in the salient variables approach, the two factor loading matrices are orthogonally rotated to arrive at a single solution that is, in some quantitative sense, a best fit or match between the two configurations. A number of different solutions have been proposed with more or less similar minimization criteria.

Some solutions are based on solutions related to the salient variables approach. Both Cliff (1966) and Pinneau and Newhouse (1964) propose indices to assess the overall relationship based on the techniques derived from their salient variables approach. Others, e.g.

Schöneman (1966), have developed techniques independently of the salient variables approach. In most applications, the matching involves one configuration being specified as the target matrix and the other being specified as the matrix to be rotated. Schöneman and Carrol (1970) present a technique which can yield asymmetrical indices (depending on which matrix is specified as the target), or a symmetric index which is analogous to the index of invariance presented in the next chapter. In no case, however, is the distribution of the final index known.

The index presented in the remainder of this dissertation differs from all of the above approaches in that it is based on the distances between points (as derived from the configuration matrices) rather than being based on rotations of the configuration matrices.

CHAPTER II

INDEX OF INVARIANCE

One goal of this research was to find an easily applied index, and therefore two initial decisions were made: (1) the index would be based on distances rather than on the configuration matrices, and (2) the index would employ the product-moment correlation computation.

The first decision was made because, in a standard Euclidian space (see below), distances remain invariant under all permissible transformations of the configuration matrix. Thus, under all orthogonal rotations of the configuration matrix, the distances will remain invariant.

The second decision was made partially because of the computational simplicity of the correlation coefficient. A primary consideration was that, with a null hypothesis of zero correlation, the null hypothesis is rejected when the two samples are similar (e.g. invariant), unlike tests on means ($\bar{t}$ or $\bar{F}$) which reject the null hypothesis when samples are dissimilar.

Furthermore, when the configuration matrix is expanded or reduced (scalar multiplication of the

configuration matrix), the corresponding distances are also multiplied by a constant. Thus, scale changes of the configuration matrix will alter the scale of the distances, but will not alter an index based on correlations.

Thus, the correlation provides an index which is invariant under scale changes and permissible transformations of the configuration matrices. The correlation provides an index which will yield a significant result when the two configurations are more similar than would be expected by chance. Thus, significance will be indicated when two configurations are invariant, or when, while not invariant, the configurations are more similar than would be expected by chance.

A final advantage from using the correlation coefficient is that, unlike some more esoteric techniques, almost every researcher is accustomed to using the correlation coefficient.

Within the limitations of the above two decisions, there were a number of different possible indices that could be constructed: of the many possible, two were investigated in some detail. Before examining these indices, however, the distances themselves should be understood.

The distance matrix. Within the limitations of the

above, before an index can be computed, the configuration matrix must be converted to a matrix of distances. In the present application, only the standard, Euclidian distance will be considered. For a configuration matrix with entries c_{ik} for row i , column k , the distance between any two points, i and j (i.e. the distance between i and j d_{ij}) is:

$$d_{ij} = \left[\sum_{k=1}^m (c_{ik} - c_{jk})^2 \right]^{1/2}$$

where m is the number of dimensions or columns in the configuration matrix. Note that the distance from i to j (d_{ij}) is the same as the distance from j to i (d_{ji}), and that the distance from a point to itself is zero. Taking all possible distances for n points, there would be n^2 distances. Since, however, n of these distances would be from a point to itself, these distances will be ignored leaving $n^2 - n$ or $n(n-1)$ distances. Furthermore, since $d_{ij} = d_{ji}$, (i.e. the distance matrix is symmetric) there are only $n(n-1)/2$ distances of interest and these will completely describe the configuration matrix.¹

1. To avoid confusion, it is prudent to always reference a distance between two points with the higher numbered point first, e.g. d_{32} , thus guaranteeing that the $n(n-1)/2$ distances will have the same designations for any configuration matrix.

Table 1 contains two solution configurations for two independent sets of data for the same eight stimuli (points).² Using the distance formula as noted above,

2. These data are from the experiment on random forms described in the next chapter.

the entire set of 28 inter-point distances was computed for each configuration and is presented in Table 2. Note that, in order to compare the distances between configuration matrices, the points in the two matrices must be given the same identification so that their corresponding inter-point distances will have identical identification in the two sets of distances.

Within this set of 28 distances, however, there are not 27 degrees of freedom because the distances are highly interrelated. The basic cause of the interrelationships is the "triangular inequality" which states that, for any three distinct points, i , j , and k , their distances are restrained by the inequality: $d_{ij} \leq d_{ik} + d_{jk}$ (see Torgerson, 1958, Ch. 11; and Shepard, 1966). This inequality can best be understood in terms of a geometric representation in which the points are the three vertices of a triangle and the distances are the sides of the triangle as in Figure 1. The inequality merely states that no one side may exceed the sum of the other two sides.

TABLE 1

SOLUTION CONFIGURATIONS
FOR EIGHT POINTS IN
TWO DIMENSIONS

POINT NUMBER	CONFIGURATION I		CONFIGURATION II	
	DIMENSION		DIMENSION	
	1	2	1	2
=====	=====	=====	=====	=====
1	-0.336	0.376	-0.869	-0.053
2	-0.749	0.093	-1.111	-0.068
3	1.047	-0.790	1.113	-0.113
4	-0.243	-0.866	0.673	-0.453
5	-0.628	0.328	-0.897	-0.406
6	0.645	1.191	0.330	1.051
7	-0.844	-0.563	0.371	-0.879
8	1.108	0.230	0.339	0.920

TABLE 2

16

INTER-POINT DISTANCES FROM
SOLUTION CONFIGURATIONS
FOR EIGHT POINTS IN
TWO DIMENSIONS

CONFIGURATION I

POINT NUMBER	1	2	3	4	5	6	7
=====	=====	=====	=====	=====	=====	=====	=====
2	0.501						
3	1.809	2.001					
4	1.245	1.094	1.292				
5	0.296	0.264	2.014	1.255			
6	1.275	1.774	2.021	2.240	1.538		
7	1.068	0.663	1.905	0.673	0.917	2.301	
8	1.451	1.862	1.022	1.740	1.739	1.067	2.107

CONFIGURATION II

POINT NUMBER	1	2	3	4	5	6	7
=====	=====	=====	=====	=====	=====	=====	=====
2	0.242						
3	1.983	2.224					
4	1.593	1.825	0.556				
5	0.354	0.400	2.031	1.571			
6	1.630	1.824	1.403	1.543	1.905		
7	1.490	1.689	1.066	0.522	1.353	1.930	
8	1.590	1.796	1.261	1.402	1.947	0.144	1.799

FIGURE 1

17

GEOMETRIC REPRESENTATION
OF DISTANCES

ONE DIMENSION

$$\begin{aligned} d_{AB} &= 1 \\ d_{BC} &= 2 \end{aligned}$$

$$\begin{aligned} d_{AC} &= 3 \\ d_{BD} &= 6 \end{aligned}$$

$$\begin{aligned} d_{AD} &= 7 \\ d_{CD} &= 4 \end{aligned}$$

TWO DIMENSIONS

Note that the inequality may be re-written: $d_{ik} \geq d_{ij} - d_{jk}$. Since the above holds for all three points, it follows that the distance between any two points must be at least as large as the absolute difference between the distances from the two points to any third point. The more points in the entire set, the more restrictions the triangular inequality will impose on the distance between any pair of points: for any dyad of points, i and j , there will be $n-2$ additional points, each of which can be designated k in the above inequalities, thus imposing $n-2$ restrictions on the maximum and minimum possible distance between i and j .

The actual number of degrees of freedom for the distances is difficult to specify. For ODMS solutions, Young (1970) proposes the ratio of the degrees of freedom for the raw dissimilarities to the degrees of freedom for the configuration matrix as a useful descriptive tool: i.e. $[n(n-1)/2]/[m(n-1)-m(m-1)/2]$ for n points in m dimensions. Young points out, however, that "this ratio is purely descriptive and is not meant to be used in a formal statistical sense, because the monotonic transformation makes it difficult to determine the precise expression for the degrees of freedom." (Pp. 466-468.) The total degrees of freedom for the distances, then, should be more than $n-1$, less than $n(n-1)/2$, and should

depend, in some manner, on m .

Index I: The first index to be examined was merely the correlation between the entire distance matrices: i.e. the correlation between the two vectors of $n(n-1)/2$ elements each. (Note that the order of the distances in the vectors is arbitrary, but must be the same for both vectors of distances.) Index I, then, is a product-moment correlation based on $D=n(n-1)/2$ data values. The distribution of the product-moment correlation is known under the assumptions of independently sampled, normally distributed data values in each data vector. When the assumptions are met, each vector has $D-1$ degrees of freedom, and the correlation coefficient can be tested using standard procedures.³

3. To test the hypothesis that the population correlation, ρ , is zero, a test based on the t -distribution is often used (Hays, 1963, p. 529). A more general procedure, which permits testing for $\rho=x$ (where x is "not too close to 1 or -1"), transforms obtained correlations into values that can be compared to a normal distribution $ND(0,1)$ by use of Fisher's " r to z " transformation (Hays, 1963, p. 531).

In the previous section, however, it was noted that distances are highly interdependent so that the assumption of independent data values within the vectors of distances is clearly violated. In addition, distances are not

normally distributed: distances are always positive and tend to be positively skewed. Despite these two violations of the assumptions, it was felt that the distribution of Index I might well closely conform to the distribution found in a standard correlation table, albeit with fewer than $D-1$ degrees of freedom. It did not.

To find the distribution of Index I, a number of configuration matrices were empirically generated. Each configuration matrix was identified by the number of points (rows) and dimensions (columns) (NP/ND), and there were NP times ND coordinates for the matrix. The coordinates for each matrix were independent data values sampled from a normal distribution $c_{ij} \stackrel{d}{=} N(0,1)$ by means of a pseudo-random number generator. For each (NP/ND) combination, 210 such configurations were generated, each thus having coordinates sampled from a multivariate normal population. As a preliminary step, thirteen combinations of NP and ND were selected, ranging from (6/1) through (45/6). For each (NP/ND) combination, 1000 values for Index I were computed out of the 21,945 different combinations that could have been used. These values therefore yielded thirteen different empirical distributions for Index I.

One striking finding from the empirical distributions was the lack of extreme negative correlations. In

general, it is impossible to exactly reverse the rank order of the distances as would be necessary to obtain a correlation of -1.0. The reader can verify this fact by spending a few moments with the one dimensional distances shown in Figure 1. A second problem with the empirical distributions was that they were positively skewed: the median correlation was negative while the mean was near zero. The biggest problem, however, was that there was no systematic pattern for the degrees of freedom which would have caused the correlation to conform to standard correlation tables.

Index II: The second index was selected because it took into account the interdependencies of distances. Instead of using all D distances, a subset was selected so that each distance in the subset would be independent of other distances in the subset. There are at most $n-1$ independent distances, where independence implies no connected loops such as $d_{21} d_{32} d_{43} d_{41}$. Working with any independent subset of $n-1$ distances, the relationship between that subset and an identical subset from another set of distances should estimate, with $n-2$ degrees of freedom for each subset, the relationship between the two sets of $n(n-1)/2$ distances.

The second index, then, was a correlation index based on $n-1$ data values, where the data values are a

subset of independent distances from the set of D distances. To correct for the skewed distribution of distances, a correction for skewness was applied to each distance value, i.e. taking the logarithm of the distance.

An initial subset was arbitrarily selected, namely $\{d_{21}, d_{32}, d_{43}, \dots, d_{n \ n-1}\}$. Each distance in the subset was transformed: $d'_{ij} = \log_e(d_{ij} + 1.0)$, to yield a set of independent data values that would more closely approximate the normal distribution than the untransformed distances. The second index, then, was defined as the correlation between two such vectors of transformed distances. It was felt that the distribution of Index II should closely approximate the distribution of the correlation coefficient in a standard table with $n-1$ degrees of freedom.

To be a valid index, however, the second index would also have to provide the same index value, no matter which subset of independent distances was selected. Thus there were two questions to be asked: (1) does the distribution of Index II approximate the distribution of the correlation coefficient on $n-1$ independent bivariate normal values, and (2) is the obtained index stable under selection of different independent subsets of distances?

To answer the above questions, a number of configuration matrices were generated using the same procedures as

described above for Index I. Seven combinations of NP and ND were selected, ranging from (6/1) through (30/3). For each (NP/ND) combination, 160 configurations were generated, and Index II was computed for 608 of the possible 12,720 possible dyads of configurations. In general, the index closely approximated the predicted distribution using normal correlation tables to predict the distribution of the index.

The index failed, however, when different subsets were used to estimate the relationship between the same two configurations. As an example, for two (8/3) configurations, the initial relationship was estimated at -0.002 . Using different subsets, however, the estimated index varied from below $-.30$ to above $.40$!

Because of the instability of the index, Index II was dropped, and attention was focused on describing the distribution of Index I.

The Gram-Charlier technique to describe distributions:

As noted above, the distribution of Index I was different from the standard distribution of the product-moment correlation. It was felt, however, that it might be possible to describe the distribution of the index, using two parameters to describe the distribution, with parameters: the number of points and the dimensionality of the configuration matrix (NP/ND). Essentially, then it was

hoped that a two-parameter distribution could be found for the index.

As a preliminary step, the thirteen empirical distributions were generated, each based on 1000 values of Index I as described above. It was noted that the mean of the empirical distributions were close to zero and that the first three moments about the mean tended to vary systematically as a function of the two tentative parameters. On the basis of this apparent relationship between the two parameters and the moments, it was decided to attempt to describe the distributions using two steps: (1) estimating the moments by the two parameters NP and ND, and (2) estimating the distribution using the Gram-Charlier technique.

The Gram-Charlier technique describes a family of curves based on the normal distribution and deviations from normality as a function of the moments. Any distribution is considered to have important "characteristics", namely the mean, standard deviation, skewness, and kurtosis (additional higher moments may also be included but were not considered in the characteristics as defined by Charlier (1947).) Using the characteristics, then, a curve can be found to approximate the density function for the data.

Although the technique can be used with raw data, it is most easily applied if the data are first converted to z-scores. Note that the conversion to z-scores uses the first two characteristics, the mean and the standard deviation:

$$x = \frac{X - \bar{X}}{s}$$

where x is the z-score, X is the raw score, $\bar{X}$ is the mean raw data value, and s is the observed standard deviation. Once the z-scores have been computed, the ordinate of the density function at any point is found by correcting the ordinate from the normal distribution as a function of the higher order characteristics.

For any z-score, the ordinate of the normal density function is:

$$y = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Correcting the above for higher moments in general, the Gram-Charlier technique predicts the ordinate to be:

$$f(x) = \phi(x) - \frac{1}{3!}u_3\phi^3(x) + \frac{1}{4!}(u_4-3)\phi^4(x)$$

$$- \frac{1}{5!}(u_5-10u_3)\phi^5(x) + \dots = \sum_{j=0}^{\infty} A_j \phi^j(x)$$

where $\phi^n(x) = \frac{\partial^n}{\partial x^n} \phi(x)$; u_n is the n^{th} moment of z-scores

about the mean; and:

$$A_n = \frac{\sigma(-1)^n}{n!} \int_{-\infty}^{\infty} f(x) H_n(x) dx$$

where $H_n(x)$ is a Hermite polynomial:

$$H_n(x) = x^n - \frac{n(n-1)}{2} x^{n-2} + \frac{n(n-1)(n-2)}{2 \cdot 4} x^{n-4} \dots$$

(Rietz, 1936, p. 66). Thus, using the above, one can generate a density function for z-scores by, in effect, starting with the density function for the normal curve and distorting it as a function of the moments.

Following Charlier, it was decided to use only the first three terms as an approximation technique for the distribution of Index I. When x is measured in z-scores, the first three terms require:

$$\phi(x); \phi^3(x) = (3x-x^3)\phi(x); \text{ and } \phi^4(x) = (x^4-6x^2-3)\phi(x)$$

(Elderton and Johnson, 1969, p. 112).

From the above, with the first three terms only, the ordinate for any z-score is estimated to be:

$$f(x) = \phi(x) [1 - u_3(3x-x^3)/6 + (u_4-3)(x^4-6x^2+3)/24].$$

The above ordinate function can then be converted back into a density function for raw scores by converting the z-scores to raw scores and correcting the ordinate so that the area under the curve remains unity:

$$X = x\sigma + \bar{X} \quad \text{and} \quad f'(X) = f(x)/\sigma .$$

Using the above to generate an approximation for a density function, the characteristics will operate in the following manner: the mean ($\bar{X}$) will merely shift the curve up or down on the X-axis; the standard deviation (σ , estimated from $\underline{s}$) will expand or contract the curve on the X-axis and will correspondingly cause the ordinate to decrease or increase at every point; the skewness (u_3) will cause the curve to be asymmetric if it is non-zero, and the kurtosis (u_4-3) will increase or decrease the peakedness of the curve if non-zero. Figure 2 shows curves generated from the Gram-Charlier technique using the first three terms of the expression. More extensive coverage of the technique can be found in Rietz (1936) or Elderton and Johnson (1960).

Estimating the area under the density function. The formula for the function $f'(X)$ is a non-integrable function, so some approximation technique must be used to compute the area under the curve. Although more complex methods exist, it was decided to use a simple trapezoidal approximation to yield the area under the curve. For any two points X_i and X_j , the area under the curve between them can be estimated by:

$$|X_i - X_j| [f'(X_i) + f'(X_j)] / 2 .$$

FIGURE 2

GRAM-CHARLIER FAMILY OF CURVES USING $F'(X)$

	MEAN	S.D.	SKEN	KUR.	
—————	0.	1.0	0.	0.	NORMAL CURVE
-----	0.	1.0	0.6	0.	POSITIVE SKEW
- - - - -	0.	1.0	0.	-.25	NEGATIVE KURTOSIS

	MEAN	S.D.	SKEN	KUR.	
—————	1.0	1.0	0.	0.	SHIFTED MEAN
-----	1.0	0.6	0.	0.	REDUCED S.D.
- - - - -	1.0	0.6	0.6	0.	S.D. AND SKEW.

The closer X_i is to X_j , the more accurate the approximation will be. The area under the total curve, then, can be obtained by finding a series of X values along the X -axis, computing the estimated area between each successive pair of values, and adding up each subset of the area successively to obtain the total area up to any point. This approximation may be made to be as accurate as desired by increasing the frequency of points along the X -axis and is represented in Figure 3.

Estimating the distribution of Index I with empirical moments: The initial approximation of Index I was performed to insure that the Gram-Charlier technique would, in fact, offer a close approximation to the empirical distributions. For the thirteen empirical distributions of Index I, an initial comparison was made between the observed curve and a curve predicted by the Gram-Charlier technique using the observed moments in the empirical distribution as the characteristics.

For each distribution of the index, fifteen points along the X -axis were identified from the cumulative frequency distribution corresponding the cumulative frequency cutoffs ranging from .5% through 95.5%. For each of these points, $f'(X)$ was computed to yield an estimated density function curve from the Gram-Charlier technique. Although no observed values for the ordinate

FIGURE 3

TRAPEZOIDAL APPROXIMATION
OF AREA

of the density function were computed (to compare to $f'(X)$), it was possible to compare the estimated curve and the observed data in terms of area under the curves. At each of the observed points, then, the area under the estimated curve was estimated (for X less than or equal to the observed point) using the trapezoidal rule.

The expected area at any point was given by the cumulative frequency cutoff corresponding to the observed point, and the estimated area at that point was given by applying the trapezoidal rule to the curve estimated by the Gram-Charlier technique. The fit between the estimated area and the expected area was excellent, and is presented in Table 3. Chi-square tests for equality of distributions were all negligible and failed to reject the hypothesis of equality between empirical and estimated density functions.⁴

4. Chi-square tests used the expected and estimated area between successive points as the probability of the index falling in sixteen categories (bounded by -1.0, the fifteen cutoff points, and 1.0). The test compared the expected and estimated frequencies (see Siegel, 1956, pp. 104-110). For two distributions with 16 categories, the test had 15 degrees of freedom. The largest observed χ^2 was 11.5 (for 6 points in 1 dimension) with a mean χ^2 of 2.8, all non-significant.

Predicting characteristics of Index I: Since the above findings indicated that the Gram-Charlier technique

TABLE 3

DIFFERENCES BETWEEN EMPIRICAL DENSITY FUNCTION
AND ESTIMATED DENSITY FUNCTION BASED
ON EMPIRICAL CHARACTERISTICS

EMPIRICAL DENSITY FUNCTION	CONFIGURATION CATEGORY (NP/ND) ^a						
	(6/ 1)	(8/ 1)	(8/ 3)	(12/ 1)	(12/ 3)	(16/ 1)	(16/ 3)
0.5	-1.88 ^b	-1.38	-0.20	-0.50	-0.19	-0.95	-0.73
1.0	-2.52	-1.59	-0.66	-0.32	-0.14	-1.21	-0.98
2.0	-3.41	-1.19	-0.69	-0.93	-0.16	-1.05	-1.10
5.0	-2.98	-1.27	-0.90	-0.15	-0.48	-0.94	-0.71
10.0	-2.62	-0.97	-0.37	0.78	0.21	-1.91	-0.12
20.0	0.53	0.23	0.15	2.95	0.14	0.72	0.89
30.0	1.76	3.43	-0.45	3.71	0.38	1.78	0.37
50.0	3.74	1.87	1.35	1.40	4.63	1.70	-0.59
70.0	2.18	-0.16	-0.49	0.07	0.80	0.35	0.11
80.0	-0.09	-1.21	0.47	-1.19	0.82	0.19	0.91
90.0	-1.30	-0.24	0.38	0.24	0.03	0.03	0.37
95.0	0.05	1.02	0.81	0.87	0.49	0.55	-0.27
98.0	0.33	-0.31	-0.17	0.45	0.46	-0.19	0.16
99.0	0.13	-0.21	0.31	-0.09	0.16	-0.16	0.03
99.5	-0.07	0.02	0.09	0.00	-0.09	0.04	0.02
AVERAGE ABSOLUTE DIFFERENCE	1.57	1.01	0.49	0.92	0.61	0.78	0.49

EMPIRICAL DENSITY FUNCTION	CONFIGURATION CATEGORY (NP/ND)					
	(16/ 6)	(30/ 1)	(30/ 3)	(30/ 6)	(45/ 3)	(45/ 6)
0.5	-0.15	-0.39	-0.01	-0.14	0.11	-0.10
1.0	-0.29	-0.65	0.02	0.03	0.08	-0.12
2.0	-0.89	-1.09	0.19	0.32	-0.11	-0.15
5.0	-0.45	-0.65	-0.07	0.60	-0.27	-0.10
10.0	0.23	-0.25	-0.67	0.24	0.31	0.60
20.0	-0.32	0.24	0.16	-0.69	-0.64	-0.17
30.0	0.10	0.11	-1.22	0.29	-1.05	-0.78
50.0	0.17	0.77	-0.38	0.00	-0.73	-0.50
70.0	0.70	-0.95	0.33	-0.84	0.95	0.92
80.0	-0.04	-1.06	0.79	0.57	0.66	0.55
90.0	-0.02	0.75	0.93	0.74	-0.10	0.03
95.0	0.25	1.11	-0.44	0.75	0.22	-0.71
98.0	-0.08	0.51	-0.06	0.30	0.46	0.09
99.0	-0.09	0.21	-0.02	0.25	0.31	0.21
99.5	0.07	0.10	0.00	0.21	0.03	0.19
AVERAGE ABSOLUTE DIFFERENCE	0.26	0.59	0.38	0.40	0.40	0.37

- a. NP IS THE NUMBER OF POINTS IN THE CONFIGURATION, ND IS THE NUMBER OF DIMENSIONS IN THE CONFIGURATION. EACH EMPIRICAL DISTRIBUTION IS BASED ON 1000 VALUES OF INDEX I.
- b. TABLE ENTRIES ARE PERCENT AREA UNDER EMPIRICAL CURVE MINUS THAT UNDER EXPECTED CURVE (FOR POINTS LOCATED IN THE EMPIRICAL DISTRIBUTION).

would approximate the empirical distribution of Index I, it remained to be shown that the two parameters (NP/ND) could be used to estimate the characteristics of the distribution.

As noted above, the mean of each of the empirical distributions was near zero. The null hypothesis of mean zero was tested for each of the thirteen distributions as shown in Table 4. Although the null hypothesis was rejected more than would be expected by chance, there was no apparent pattern for the means, and an overall test for zero mean failed to reject the null hypothesis that the average mean was zero. Thus, the prediction for the mean was accepted as mean equal to zero.

Prediction equations for the remaining three characteristics were made by applying a step-wise regression to predict each characteristic in terms of variables based on NP and ND.⁵ Both regression with and without

5. Stepwise regression performed by "BMD02R, Stepwise regression program" (Dixon, 1967, pp. 233-257d).

the constant term were tried with the final equations selected as shown in Table 5.

Estimating the distribution of Index I with predicted moments: The utility of the predicted characteristics was then tested on two different sets of distributions of

TABLE 4
TESTS FOR THE HYPOTHESIS: $\bar{X} = 0.0$

<u>DISTRIBUTION POINTS</u>	<u>DIMENSIONS</u>	<u>OBSERVED SAMPLE SD</u>	<u>ESTIMATED STANDARD ERROR OF THE MEAN</u>	<u>Z-SCORE FOR MEAN WITH $u=0$</u>
6	1	.2972	.0094	.568
8	1	.2222	.0070	.598
8	3	.2232	.0071	-2.300
12	1	.1680	.0053	.273
12	3	.1713	.0054	.880
16	1	.1325	.0042	-.975
16	3	.1423	.0045	2.473
16	6	.1417	.0045	1.079
30	1	.8046	.0027	-1.319
30	3	.0904	.0029	-.210
30	6	.0932	.0030	1.930
45	3	.0749	.0024	.351
45	6	.0740	.0023	-.636

Test of constant term (test of difference scores: mean - 0):

$t = .25$

df = 12

H_0 : mean=0 not rejected

TABLE 5
PREDICTION EQUATIONS FOR
THE CHARACTERISTICS
FOR INDEX I

PREDICTIONS BASED ON 13 DISTRIBUTIONS ^a .	MULTIPLE R	STD. ERROR
MEAN = 0.0	----	----
S.D. = $0.41 + 1.5/NP$	.997	.005
SKEW = $0.0 + .73/ND$	.977	.109
KUR. = $0.0 + .660/ND^4 - 983.7/NP^4$	.856	.220
PREDICTIONS BASED ON 20 DISTRIBUTIONS	MULTIPLE R	STD. ERROR
MEAN = 0.0	----	----
S.D. = $0.41 + 1.5/NP$	.998	.005
SKEW = $0.0 + .72/ND$	.965	.120
KUR. = $0.0 + .635/ND^4 - 109.8/NP^3$	.808	.205

a. NP is number of points in the configuration. ND is the number of dimensions in the configuration.

Index I. First, the predicted characteristics were used to generate $f'(X)$ using the Gram-Charlier technique for each of the thirteen distributions of Index I. Once again, $f'(X)$ was computed for the thirteen points identified by the cutoffs on the empirical cumulative frequency curves. Using the trapezoidal rule, then, the predicted area was computed as the area under the predicted density function and below the observed X value. As shown in Table 6, the fit between the expected and predicted areas was extremely good and chi-square tests failed to reject the hypothesis that the predicted and empirical distributions were the same. (The largest χ^2 value was 17.4 for 8 points in one dimension, and the average χ^2 value was 4.4.)

Secondly, seven new empirical distributions of Index I were generated to increase the range of the empirical data base and to test the validity of the prediction equations derived from the first thirteen distributions. The new configuration matrices were generated as described for the initial set, and ranged from (5/1) through (60/10). The characteristics were predicted using the prediction equations from the initial thirteen distributions, and the predicted area was compared to the expected area as shown in Table 7. The fit was again extremely close, and chi-square tests for differences in

TABLE 5

DIFFERENCES BETWEEN EMPIRICAL DENSITY FUNCTION
AND PREDICTED DENSITY FUNCTION BASED
ON PREDICTED CHARACTERISTICS

EMPIRICAL DENSITY FUNCTION	CONFIGURATION CATEGORY (NP/ND) ^a						
	(6/ 1)	(8/ 1)	(8/ 3)	(12/ 1)	(12/ 3)	(16/ 1)	(16/ 3)
0.5	-1.62 ^b	-2.68	-0.24	-1.22	-0.08	-1.17	-1.06
1.0	-2.25	-2.98	-0.65	-1.08	0.00	-1.40	-1.33
2.0	-3.15	-2.64	-0.63	-1.92	0.02	-1.20	-1.46
5.0	-2.74	-3.04	-0.52	-1.14	-0.27	-0.93	-1.12
10.0	-2.47	-2.78	0.26	-0.25	0.36	-1.57	-0.66
20.0	0.52	-1.51	1.50	2.08	0.01	1.42	-0.01
30.0	1.51	1.88	1.59	3.21	-0.10	2.85	-1.06
50.0	3.08	1.22	4.17	2.22	3.62	3.16	-3.12
70.0	1.32	0.02	2.45	1.42	-0.63	1.59	-3.05
80.0	-0.79	-0.93	3.05	-0.14	-0.53	1.02	-2.19
90.0	-1.69	-0.14	2.20	0.18	-0.94	0.45	-2.05
95.0	-0.20	1.05	2.06	0.10	-0.17	0.91	-1.80
98.0	0.18	-0.26	0.39	-0.17	0.09	0.09	-0.69
99.0	0.05	-0.16	0.74	-0.37	-0.04	0.02	-0.44
99.5	-0.11	0.06	0.31	-0.16	-0.18	0.18	-0.23
AVERAGE ABSOLUTE DIFFERENCE	1.45	1.42	1.39	1.04	0.47	1.20	1.35

EMPIRICAL DENSITY FUNCTION	CONFIGURATION CATEGORY (NP/ND)					
	(16/ 6)	(30/ 1)	(30/ 3)	(30/ 6)	(45/ 3)	(45/ 6)
0.5	-0.13	-0.54	0.28	0.42	0.45	-0.43
1.0	-0.17	-0.87	0.36	0.64	0.50	-0.50
2.0	-0.55	-1.39	0.59	1.03	0.39	-0.56
5.0	0.12	-1.06	0.33	1.53	0.24	-0.56
10.0	0.98	-0.74	-0.38	1.49	0.68	0.14
20.0	0.25	-0.19	0.15	1.15	-0.79	-0.55
30.0	0.09	-0.13	-1.55	2.58	-1.68	-1.01
50.0	-1.38	1.04	-1.42	2.36	-1.86	-0.38
70.0	-1.87	-0.03	0.08	1.92	-0.04	1.51
80.0	-2.51	0.21	0.98	2.79	0.05	1.73
90.0	-1.73	2.40	1.31	1.93	-0.14	1.13
95.0	-0.76	2.93	0.15	1.04	0.42	0.27
98.0	-0.46	2.09	0.41	0.09	0.71	0.85
99.0	-0.26	1.34	0.32	0.01	0.52	0.79
99.5	-0.02	0.83	0.22	0.02	0.15	0.61
AVERAGE ABSOLUTE DIFFERENCE	0.75	1.05	0.56	1.30	0.58	0.74

a. NP IS THE NUMBER OF POINTS IN THE CONFIGURATION, ND IS THE NUMBER OF DIMENSIONS IN THE CONFIGURATION. EACH EMPIRICAL DISTRIBUTION IS BASED ON 1000 VALUES OF INDEX I.

b. TABLE ENTRIES ARE PERCENT AREA UNDER EMPIRICAL CURVE MINUS THAT UNDER EXPECTED CURVE (FOR POINTS LOCATED IN THE EMPIRICAL DISTRIBUTION). CHARACTERISTIC PREDICTIONS BASED ON THIRTEEN CONFIGURATION CATEGORIES SHOWN.

DIFFERENCES BETWEEN EMPIRICAL DENSITY FUNCTION
AND PREDICTED DENSITY FUNCTION BASED ON
INDEPENDENTLY PREDICTED CHARACTERISTICS

EMPIRICAL DENSITY FUNCTION	CONFIGURATION CATEGORY (NP/ND)A.						
	(5/ 1)	(10/ 2)	(10/ 4)	(20/10)	(30/10)	(60/ 3)	(60/10)
0.5	-1.64 ^b	0.67	-1.81	-0.09	0.38	-0.17	-0.26
1.0	-2.26	0.55	-2.43	0.12	0.68	-0.10	-0.69
2.0	-3.71	0.26	-3.10	0.23	0.96	-0.25	-0.32
5.0	-4.70	-0.69	-3.43	1.18	1.64	-0.78	-1.12
10.0	-4.88	-0.45	-2.50	1.33	2.49	0.66	-1.33
20.0	-2.19	0.25	-2.38	1.85	2.38	0.54	-3.05
30.0	1.68	-1.33	-1.52	2.14	2.24	-0.45	-4.87
50.0	4.57	0.04	1.23	2.38	-0.15	-0.79	-4.63
70.0	2.30	2.35	-0.60	1.70	-2.59	1.24	-2.65
80.0	1.90	2.83	-2.81	0.86	-1.29	1.92	-1.11
90.0	-1.08	1.60	-2.58	0.25	-0.33	1.15	-0.15
95.0	-0.92	1.74	-0.78	0.09	0.02	1.85	0.24
98.0	-0.52	1.27	-0.18	-0.18	0.15	1.13	0.84
99.0	-0.47	0.63	-0.19	-0.27	0.17	0.68	0.02
99.5	-0.28	0.17	-0.12	-0.11	-0.01	0.45	0.02
AVERAGE ABSOLUTE DIFFERENCE	2.21	0.99	1.71	0.85	1.04	0.81	1.42

- a. NP IS THE NUMBER OF POINTS IN THE CONFIGURATION, ND IS THE NUMBER OF DIMENSIONS IN THE CONFIGURATION. EACH EMPIRICAL DISTRIBUTION IS BASED ON 1000 VALUES OF INDEX I.
- b. TABLE ENTRIES ARE PERCENT AREA UNDER EMPIRICAL CURVE MINUS THAT UNDER EXPECTED CURVE (FOR POINTS LOCATED IN THE EMPIRICAL DISTRIBUTION). CHARACTERISTIC PREDICTIONS BASED ON THE THIRTEEN DIFFERENT CONFIGURATION CATEGORIES.

distributions were negligible. (The largest χ^2 value was 13.3 with an average χ^2 value of 5.1.)

On the basis of the above, the entire set of twenty empirical distributions were used to predict the characteristics as shown in Table 5. On the basis of these revised predictions, then, the Gram-Charlier technique can be used to generate density functions for Index I using the number of points and the number of dimensions as two parameters.

Tables for Index I: Using the predictions, then, the density function curves were computed from (5/1) through (300/80). For these curves, there were 1000 values for $f'(X)$ computed to increase accuracy in estimating the area under the curve (note that there had been only 15 values of $f'(X)$ in the preliminary tests). For selected areas, values of X were found such that, using the trapezoidal approximation for area:

$$\text{AREA} = \int_{-1.0}^X f'(X) dx$$

(note that the lower limit of integration, -1.0, is equivalent to $-\infty$ since the index is bounded by ± 1.0). For selected areas, Tables 8-13 present the value of X below which Index I would be expected to fall for a specified percentage of time under the null hypothesis of no relationship between the two configurations. Unity minus

SIGNIFICANT VALUES FOR INDEX OF INVARIANCE: $P=0.005$

TABLE 8

NUMBER OF DIMENSIONS	NUMBER		POINTS	
	1	2	3	4
5	900 810 757 790 785	816 736 722 715 708	740 670 657 651 645 641 639	674 611 600 594 589 586 582 581
6				557 498 478 471 468 465 461 459 457 457
7				521 458 449 443 439 436 434 432 431 430 428 427
8				575 522 512 506 501 499 496 494 491 490
9				620 563 552 545 541 538 535 534 531
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
25				
30				
35				
40				
45				
50				
55				
60				
70				
80				

TABLE ENTRIES LISTED WITHOUT DECIMAL POINT. NUMBER PRINTED REPRESENTS THREE DIGITS BEYOND THE DECIMAL POINT: *XXX

TABLE 12

SIGNIFICANT VALUES FOR INDEX OF INVARIANCE: $p=0.10^a$

NUMBER POINTS	NUMBER OF DIMENSIONS																																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	35	40	45	50	55	60	70	80					
5	461	462	460	457	458																														
6	390	391	388	387	386	385																													
7	340	340	338	337	336	335	336																												
8	303	304	302	301	300	299	298	298																											
9	275	275	274	272	271	271	271	270	270																										
10	252	253	251	250	249	249	248	248	248	248																									
11	238	235	233	232	231	230	230	230	230	230	230																								
12	222	220	218	217	216	216	216	215	215	215	215	215																							
13	209	207	205	204	203	203	203	203	202	202	202	202	202																						
14	198	196	194	194	192	192	192	192	192	192	192	191	191	191																					
15	188	186	185	184	183	183	183	182	183	182	182	182	182	182	182																				
16	180	178	176	176	175	175	174	174	174	174	174	174	174	174	174	174																			
17	172	170	169	169	169	168	167	167	167	167	167	167	167	167	166	166	166																		
18	166	164	163	162	162	161	161	161	161	160	160	160	160	160	160	160	160	160																	
19	160	158	157	156	156	155	155	155	155	155	155	155	155	154	154	155	154	154	155	155															
20	154	153	152	151	151	150	150	150	150	150	150	150	149	149	149	149	149	149	149	149	149														
22	145	144	143	142	142	141	141	141	141	141	141	141	141	141	141	141	141	140	141	140															
24	139	136	136	135	134	134	134	134	134	133	133	133	133	133	133	133	133	133	133	133	133														
26	131	130	129	128	128	128	127	128	127	127	127	127	127	127	127	127	127	127	127	127	127	127													
28	126	125	124	123	123	122	122	122	122	122	122	122	122	122	122	122	122	122	121	122	122	121													
30	121	120	119	118	118	118	118	118	117	117	117	117	117	117	117	117	117	117	117	117	117	117	117												
40	104	103	102	102	102	102	101	101	101	101	101	101	101	101	101	101	101	101	101	101	101	101	101												
50	94	94	93	92	92	92	92	92	92	92	92	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	
60	88	87	86	86	86	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	
70	83	82	82	81	81	81	81	81	81	81	81	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	
80	79	79	78	78	78	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	
90	77	76	75	75	75	75	75	74	74	74	74	74	74	74	74	74	74	74	74	74	74	74	74	74	74	74	74	74	74	74	74	74	74	74	
100	74	74	73	73	73	73	72	72	72	72	72	72	72	72	72	72	72	72	72	72	72	72	72	72	72	72	72	72	72	72	72	72	72	72	
125	70	70	69	69	69	69	69	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	
150	68	67	67	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	
175	66	66	65	65	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	
200	64	64	63	63	63	63	63	63	63	63	63	63	63	63	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	
300	61	61	60	60	60	60	59	59	59	59	59	59	59	59	59	59	59	59	59	59	59	59	59	59	59	59	59	59	59	59	59	59	59	59	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	35	40	45	50	55	60	70	80					

^a.TABLE ENTRIES LISTED WITHOUT DECIMAL POINT. NUMBER PRINTED REPRESENTS THREE DIGITS BEYOND THE DECIMAL POINT: .xxx

TABLE 13
SIGNIFICANT VALUES FOR INDEX OF INVARIANCE: $p=0.20$

NUMBER POINTS	NUMBER OF DIMENSIONS																																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	35	40	45	50	55	60	70	80					
5	271	200	304	305	306																														
6	223	248	252	253	254	254																													
7	192	214	217	218	219	220	220																												
8	170	190	192	194	194	195	195	195	195																										
9	154	171	174	175	175	176	176	176	177	177																									
10	141	157	160	161	161	161	161	161	162	162	162																								
11	132	146	148	149	149	149	150	150	150	150	150	150																							
12	123	136	138	139	139	140	140	140	140	140	140	140	140																						
13	116	128	130	131	131	131	131	131	132	132	132	132	132	132																					
14	109	121	122	124	124	124	124	125	125	125	125	125	125	125	125																				
15	104	115	117	118	118	118	118	118	118	118	119	119	119	119	119	119																			
16	99	110	112	112	113	113	113	113	113	113	113	113	113	113	113	113	113																		
17	95	106	107	108	108	108	108	108	108	109	109	109	109	109	109	109	109	109																	
18	92	102	103	104	104	104	104	104	104	104	104	104	104	104	105	105	105	105	105	105															
19	88	98	99	100	100	100	100	100	101	101	101	101	101	101	101	101	101	101	101	101	101														
20	85	95	96	96	97	97	97	97	97	97	97	97	97	97	97	97	97	97	98	97	98	98	98												
22	80	89	90	91	91	91	91	91	91	92	92	92	92	92	92	92	92	92	92	92	92	92	92												
24	76	85	86	86	86	87	87	87	87	87	87	87	87	87	87	87	87	87	87	87	87	87	87												
26	73	81	82	82	82	83	83	83	83	83	83	83	83	83	83	83	83	83	83	83	83	83	83	83											
28	69	77	78	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	80										
30	67	74	75	76	76	76	76	76	76	76	76	76	76	76	76	76	76	76	76	76	76	76	76	76	77	77									
40	58	64	65	65	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66	66								
50	52	58	59	59	59	59	59	59	59	59	59	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60						
60	48	54	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55	55		
70	46	51	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52		
80	44	49	49	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50		
90	42	47	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48	48		
100	41	46	46	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47		
125	39	43	44	44	44	44	44	44	44	44	44	44	44	44	44	45	44	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43		
150	37	42	42	42	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43	43		
175	36	41	41	41	41	41	41	41	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42		
200	36	40	40	40	40	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41		
300	34	38	38	38	38	38	38	38	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	35	40	45	50	55	60	70	80					

TABLE ENTRIES LISTED WITHOUT DECIMAL POINT. NUMBER PRINTED REPRESENTS THREE DIGITS BEYOND THE DECIMAL POINT: .xxx

the area, yields the probability of rejecting the null hypothesis so that Tables 8-13 can be used for rejecting the null hypothesis of no relationship between two configurations.

Using the tables: The tables are based on the assumption that the empirical distributions represent configurations comparable to configurations obtained with real data. The empirical configurations were distributed multivariate normal, so configurations should be assumed to be sampled from a multivariate normal population.

The tables further assume that the configurations being compared have meaningful spatial representations of the data being analyzed. Such procedures as ODMS algorithms will always yield a configuration solution although the data may not be accurately represented by such a model.

On the basis of the findings reported above, Index I was adopted as the Index of Invariance, and it shall be referred to as such throughout the remainder of this dissertation. When, by reference to the tables, the observed Index of Invariance is more extreme than that given in the table, the null hypothesis can be rejected at the given probability level. Rejection of the null hypothesis does not mean that the two configurations are

strictly invariant, i.e. that one can be exactly mapped onto the other through orthogonal rotation, but the rejection indicates that the underlying processes for the two sets of data are more similar than would be expected by chance.

It should be clear that the Index of Invariance is strictly defined only when both configurations have the same parameters (NP/ND). Although no data exist, some speculations can be made for situations in which the parameters vary. When the number of dimensions are unequal, but the number of points are the same, the distances can still be computed for each configuration. The distribution of the Index of Invariance for such data is unknown but would probably fall between the distribution of the index for the two values of ND. A conservative approach, therefore, would be to use the tabled values for the larger of the two values of ND.

When the two configurations are not made up of the identical points, additional problems exist. One approach would be to reduce both configurations to those points (rows) which are in common. Solution configurations based on all rows, however, might not be the same or similar to the solution configurations based on data for the subset of rows, (i.e. analyzed without the points that are removed). Thus, the experimenter can analyze

subsets of the data (i.e. only the common points), or compare subsets of his solutions.

Error levels in the tables: The tables differ from the empirically derived curves, partially as a function of random variation in the empirical distributions, and partially because the tables are based on empirical approximations. On the basis of data presented in Tables 6 and 7, it seems unlikely that the tables would ever be off by more than two percent. It might, therefore, be prudent to adopt a conservative approach and add one or two percent to the obtained significance estimate.

CHAPTER III

TESTING FOR INVARIANCE: SOME EXAMPLES

Tests for invariance using the Index of Invariance can be applied to a broad range of data and situations. The examples in this chapter are included to demonstrate the utility of the index and should not delimit the range of applications. The examples all deal with applications of an ODMS technique: J. B. Kruskal's M-D-SCAL, but examples could have been found using factor analytic data, discriminant function data, etc.¹

1. The technique is described by Kruskal (1964a, 1964b), and is operationalized in his computer program MDSCAL. The program used here, MDSCAL-6MP, includes modifications by the author. The program and appropriate documentation are available from the Program Librarian, Instruction and Research Computer Center, The Ohio State University.

The first example deals with comparing configurations from the same source of data under different experimental conditions. The second deals with comparing configurations from different sources. The third deals with the relationship between a true configuration and one recovered after error had been introduced into the data.

Example 1: Confusions among phonemes: The first example is an application in which the same source supplied data for three different solution configurations. The Index of Invariance is used to test whether the data for the three different solutions can be collapsed to form a single set of data and, subsequently, a single, composit solution. The points in the experiment were sixteen English phonemes, and the data were collected by Mattingly to access the asymmetrical confusions among the sixteen (1972).

The basic paradigm was a Peterson and Peterson interference - no rehearsal paradigm (1959). The data of interest involve the incorrect identification of one phoneme when another was presented. Subjects were presented with four phonemes, occurring at a one second interval, followed by distractor instructions. After a distraction interval of 4, 8, or 16 seconds, subjects were asked to recall the phonemes and their order. There were 32 sets of four phonemes for each time condition: each phoneme thus served as a stimulus eight times in each condition for each subject. The actual phoneme given for each stimulus was recorded, and the misidentifications were summed over all twelve subjects for each time period. These "confusion matrix" data, then, served as similarities ratings since responding i given j, or responding j

given i would both serve to indicate subjective proximity between phonemes i and j. These similarities ranged from 0 through 24.

Separate scaling solutions were computed for each of the three confusion matrices corresponding to the three distraction intervals. Solutions were obtained for one through four dimensions, inclusive.² The Index

2. Scaling was performed with a maximum of 75 iterations, used Stress formula II, a random starting configuration, and computed a separate monotone regression for each row (splitting byrows).

of Invariance was computed between the three solutions for each of the four dimensions and the obtained values are presented in Table 14. All obtained values were highly significant beyond the $p = .005$ level except the 8 and 16 second configurations in one dimension which was significant at the $p = .05$ level. For these data, then, it was concluded that the underlying processes for the confusion data were more similar over the time periods than would be expected by chance, and the data were combined to yield a single solution. Mattingly further found that the two dimensional composite solution closely corresponds to previously published results.

TABLE 14
INDEX OF INVARIANCE FOR
PHONEMIC CONFUSIONS

TIME DELAYS	SOLUTION DIMENSIONALITY			
	1	2	3	4
4-8 ^a	.468 ^b	.639	.498	.519
4-16	.468	.579	.634	.534
8-16	.294	.552	.615	.604

- a. Time delays in seconds: i.e. distractor period.
Index of invariance on each row based on configurations from time delays as indicated.
- b. Each index based on 256 inter-point distances.
Data from Mattingly, 1972.

Example 2: Similarity between random forms: The second example uses the Index of Invariance to compare solution matrices between two independent sets of data. Furthermore, it demonstrates the use of the index to show changes in the relationship between configurations in different conditions.

The author ran an experiment to assess two effects: first the experiment was designed to measure the effect of establishing a psychological "set" on subjects' judgments of random forms, and secondly the experiment was designed to see whether the Index of Invariance was sensitive to whatever effects were established by the "set" manipulation. The results are based on data from two subjects, each of whom was run in two conditions. The basic experimental task was to rank all possible (28) pairs of eight random forms from the "most similar pair" to the "most dissimilar pair". In the first session, the subjects were not given any criterion or definition for "similarity" but were permitted to use any strategy they wanted. Approximately three days later, each subject was again given the task of judging the 28 pairs, but was told:

It is obvious that figures can vary on a large number of dimensions, but I have selected these eight figures because I think that they probably

vary on the two dimensions which are most important for my study: elongation and rectangular area:

ELONGATION: By elongation I mean the length of a line drawn through the center of the figure to the farthest edges.

RECTANGULAR AREA: By rectangular area I mean the size (area) of the minimul outer border which will still contain the figure, i.e. move in the edges of the outer box until the edges just touch the figure.

(Neither the box nor the figure can be rotated.)

It was hypothesized that the judgements from the two subjects from the second condition, i.e. the "set" condition, would be more similar than the judgements from the first condition, i.e. the "no-set" condition. The Index of Invariance was used to compute the relationship between the two subjects' data under the two conditions.

Eight, four-sided random forms were selected from a large pool in an attempt to select figures in varying combinations of the two dimensions used in the set manipulation (figures from Brown and Owen, 1967). The forms used and their placement on the dimensions are shown in

FIGURE 4

RANDOM FORMS USED

Figure 4. The data were collected as described above, and the rank order of the rankings was recorded as the "dissimilarity" or ordinal rating of distance between the pairs of random forms. The four sets of rankings were then scaled to yield separate solution configurations in dimensions one through six, inclusive.³

3. Scaling was performed with a maximum of 75 iterations, used Stress formula II, a random starting configuration, and scaled only the 28 dissimilarities so that diagonal elements were absent.

Obtained stress values are presented in Figure 5 for dimensions one through four. Using the criterion of Constraint (Isaac and Poor, 1972), the correct solutions were identified as the one-dimensional solutions for the "no-set" condition and as the two-dimensional solutions for both subjects in the "set" condition.

The Index of Invariance was then computed using both one and two dimensional configurations as shown in Table 15. Neither index approached significance in the "no-set" condition. In the "set" condition, the one dimensional index was significant at the $p = .10$ level, and the two dimensional index was significant at the $p = .005$ level. On the basis of these data and the Constraint measure, then, it was concluded that the two-dimensional solutions

FIGURE 5

57

OBSERVED STRESS FOR SOLUTIONS
OF RANDOM FORMS DATA

x-----x RANDOM STRESS (FROM ISARC AND POOR)
 ●-----● SUBJECT 1: NO-SET CONDITION
 ○-----○ SUBJECT 2: NO-SET CONDITION
 □-----□ SUBJECT 1: SET CONDITION
 ○-----○ SUBJECT 2: SET CONDITION

TABLE 15

INDEX OF INVARIANCE
FOR SIMILARITY OF
RANDOM FORMS

<u>CONDITION</u>	<u>SOLUTION DIMENSIONALITY</u>	
	<u>1</u>	<u>2</u>
NO SET	.078 ^a .	-.024
SET	.412	.682

a. Each index based on 28 inter-point distances.

were more similar than would be expected by chance under the set condition and thereby that the set manipulation had been successful and that the Index of Invariance was sensitive to such manipulations.

Example 3: Significance of metric determinacy: The ability of ODMS algorithms to recover "true configurations": has been assessed by "metric determinacy" (M) which is the square of the Index of Invariance when one configuration is a known "true configuration" (Shepard, 1966; Young, 1970; Isaac and Poor, 1972). Investigations using M have scaled error-perturbed distances, and then used M to indicate the extent to which the ODMS algorithm was able to recover the true configuration despite the error. To the extent that the Index of Invariance is valid in error-perturbed data, the index between the true configuration and solution configurations should be significant with small and moderate amounts of error.

Using data from Isaac and Poor (1972), the index of invariance was obtained in the true dimension for configurations varying factorially in (a) number of points (6, 8, 12, 16, and 30 points), (b) dimensionality (1, 2, and 3), and (c) error level ($c = \{0, .1, .2, .4, .7, 1, 2\}$; where the standard deviation of the error equalled

the standard deviation of the true coordinates times $\underline{c}$). The authors found, as did Young (1970), that M increased with increasing number of points, and decreased with increasing error level.

Using the Index of Invariance (i.e. the positive square root of M) for the Isaac and Poor data, the significance level was measured using Tables 8-13 below. The pattern of significance is shown in Table 16. (Note that there are five configurations in each cell.) From an examination of Table 16, it can be seen that the algorithm (MDSCAL) is consistently able to find a solution which is significantly related to the true configuration for low and moderate levels of error ($c \leq .4$). For larger levels of error, the results become less significant with increasing error, and more significant with increasing number of points.

TABLE 16

INDEX OF INVARIANCE BETWEEN TRUE CONFIGURATION
AND SOLUTIONS WITH VARYING ERROR

DIMENSION	ERROR LEVEL	NUMBER OF POINTS																			
		SIX				EIGHT				TWELVE				SIXTEEN				THIRTY			
		SIGNIFICANCE LEVEL				SIGNIFICANCE LEVEL				SIGNIFICANCE LEVEL				SIGNIFICANCE LEVEL				SIGNIFICANCE LEVEL			
		.005	.05	.20	ns.	.005	.05	.20	ns.	.005	.05	.20	ns.	.005	.05	.20	ns.	.005	.05	.20	ns.
1	NONE	5 ^a				5				5				5				5			
	.1	5				5				5				5				5			
	.2	5				5				5				5				5			
	.4	5				5				5				5				5			
	.7	1	3	1		2	1	1		5				5				5			
	1.0	1	2	1	1	1	4			5				5				5			
	2.0		1		4		2	3		1		2	2			2	3			2	3
2	NONE	5				5				5				5				5			
	.1	5				5				5				5				5			
	.2	5				5				5				5				5			
	.4	3	2			5				5				5				5			
	.7	2	1	2		2	3			5				5				5			
	1.0		1	3	1	1	2	2		2	2	1		4	1			5			
	2.0		1	2	2			1	4			2	3	1	1	2	1	1	3	1	
3	NONE	5				5				5				5				5			
	.1	5				5				5				5				5			
	.2	5				5				5				5				5			
	.4	5				5				5				5				5			
	.7	1	3	1		3		2		5				5				5			
	1.0		1	3	1		3	2		2	3			4	1			5			
	2.0		1		4			2	3		1		4			2	3		2	2	1

a. Table entries are the number of configurations resulting in index of invariance at indicated significance level. For each cell (number of points, dimension, and error level) there were five configurations. Data from Isaac and Poor, 1972.

CONCLUSIONS

The primary goal of this research was to define an index for the invariance of multidimensional configurations. As shown in the Chapter I, the existing techniques for such a comparison all suffer from the problems of unknown distributions of the indices. The present effort was designed to find an index and its distribution under the null hypothesis of no relationship between the two configurations.

Unlike previous measures for comparing factor analytic solutions, the present index of invariance is computed on the inter-point distances rather than on the configuration matrix itself. For purposes of computational simplicity, familiarity, and other reasons discussed in Chapter II, it was decided to base the index on the correlation coefficient. Following unsuccessful attempts to analytically describe the distribution of the resulting correlation, empirical distributions of the correlations were obtained for a number of different configurations. The results from the empirical data were then generalized, using the Gram-Charlier technique to

generate a density function on the basis of predicted characteristics. As a result of this generalization, significance tables were prepared on a two-parameter index of invariance: the parameters being the number of points in the configuration and the number of dimensions in the configuration.

A major advantage of the present technique over other possible techniques (discussed in the Chapter II) is that it does not require an inverse to exist for the configuration matrix. Correspondingly, a disadvantage is that it requires that both configurations have the same number of dimensions as well as the same number of points.

Applications of the index of invariance to ordinal multidimensional scaling techniques is straight-forward since the process of scaling works directly with the solution distances. Note, however, that such applications assume the standard, Euclidian metric when testing for significance. Other metrics, e.g. the City Block metric or the Dominance metric, will most likely yield distributions of the index that do not conform to the distributions tabled. Within the Euclidian metric, applications are straightforward, as shown in Chapter IV.

Applications of the index of invariance to other multivariate analytical techniques are also possible:

the index can be applied whenever a solution yields some number of points in a multidimensional space. Two obvious techniques are factor analysis and multivariate discriminant analysis.

The factor loadings matrix of a factor analysis can be interpreted as a configuration matrix: the factors are the dimensions and the variables are the points. Two such factor loading matrices can be compared if the variables for each are the same. The index of invariance, then, can be used to compare factor analytic solutions from the same subjects over two or more time periods, and can be used to compare solutions from different sets of subjects. Note, however, that the configurations are assumed to be orthogonal so that only orthogonal, not oblique, factor rotations may be used prior to computation of the index. Note also that the index will be invariant under orthoganoal rotations of either or both of the factor solutions.

A second application to factor analysis involves factor scores rather than factor loadings. In such a case, the same units must have two sets of factor scores: e.g. the same subjects under two different conditions or the same objects under two different conditions. Suppose, as an example, that one subject judged the same set of

objects on the same variables under two different conditions. Using the objects within condition as the units for the factor analysis two separate factor analyses could be computed and the factor loadings then compared as discussed above. Alternatively, the two sets of data might be put into a single factor analysis and factor scores computed for each object under each condition. The factor scores of the objects under one condition would be considered the first configuration, the factor scores under the second condition would be considered the second configuration: factors would be the dimensions of the configurations. Computing the index of invariance between these two configurations would test the null hypothesis that judgement of the points in one condition was not different than the judgement of the points in the other condition. Differences in the two factor analytic approaches for such data may or may not exist: the relationship between the two approaches has not yet been examined.

For multivariate discriminant analysis, procedures are similar to factor analysis using factor loadings. The discriminant functions are the dimensions of the coordinate space and the variables are considered the points in the coordinate matrix. Comparisons can again

be made between solutions for different subjects, or between solutions for the same subjects under different conditions.

In summary, the index of invariance has possible applications in many areas of multivariate data. Although applications have only been made for ordinal multidimensional scaling, the use in other techniques should be easily accomplished. The index of invariance permits rejection or non-rejection of the null hypothesis of independent multidimensional configurations of points. When, using the index of invariance, such configurations are accepted as being significantly related, then the data may profitably be combined to yield a single solution, and thus a single configuration.

BIBLIOGRAPHY

- Brown, D. R., and D. H. Owen. "The Metrics of Visual Form: Methodological Dyspesia." Psychological Bulletin, 1967, 68, 243-259.
- Charlier, C. V. L. Elements of Mathematical Statistics. Edited and translated by J. A. Greenwood. Cambridge, Massachusetts, 1947.
- Cliff, N. "Orthogonal Rotation to Congruence." Psychometrika, 1966, 31, 33-42.
- Dixon, W. J. BMD: Biomedical Computer Programs. Berkeley: Univ. of California Press, 1967.
- Elderton, W. P., and N. L. Johnson. Systems of Frequency Curves. Cambridge: Cambridge Univ. Press, 1969.
- Harman, H. H. Modern Factor Analysis. (Second edition, Revised.) Chicago: Univ. of Chicago Press, 1967.
- Hays, W. L. Statistics for Psychologists. New York: Holt, 1963.

- Isaac, P. D., and D. D. S. Poor. "On the Determination of Appropriate Dimensionality in Data with Error." Unpublished manuscript, 1972.
- Kruskal, J. B. "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis." Psychometrika, 1964(a), 29, 1-27.
- Mattingly, S. C. "Phonemic Parameter Processing of Sixteen English Consonants." Unpublished doctoral dissertation, in progress, Ohio State University, 1972.
- Nesselroade, J. B., and P. B. Baltes. "On a Dilemma of Comparative Factor Analysis: A Study of Factor Matching Based on Random Data. Educational and Psychological Measurement, 1970, 30, 935-948.
- Peterson, L., and M. J. Peterson. "Short-term Retention of Individual Verbal items." Journal of Experimental Psychology, 1959, 58, 193-198.
- Pinneau, S. R., and A. Newhouse. "Measures of Invariance and Comparability in Factor Analysis for Fixed Variables." Psychometrika, 1964, 29, Pp. 271-281.

Rietz, Henry Lewis. Mathematical Statistics. (The Carus Mathematical Monographs Number Three.) Chicago: Open Court, 1936.

Rummel, R. J. Applied Factor Analysis. Evanston: Northwestern Univ. Press, 1970.

Schönemann, P. H. "A Generalized Solution of the Orthogonal Procrustese Problem." Psychometrika, 1966, 31, 1-10.

Schönemann, P. H., and R. M. Carroll. "Fitting one Matrix to Another Under Choice of a Central Dilation and a Rigid Motion." Psychometrika, 1970, 35, 245-255.

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

Taylor, P. A. "The Use of Factor Models in Curriculum Evaluation: A Mathematical Model Relating Two Factor Structures." Educational and Psychological Measurement, 1967, 27, 305-321.

Torgerson, W. S. Theory and Methods of Scaling. New York: Wiley, 1958.

Tucker, L. A. "A Method for Synthesis of Factor Analytic Studies." Personnel Research Section Report No. 984. Washington D.C.: Department of the Army, 1951.

Young, F. W. "Nonmetric Multidimensional Scaling: Recovery of Metric Information." Psychometrika, 1970, 35, 475-492.