EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1978, 38

DIFFERENTIATING THE CONTRIBUTION OF ELEVATION, SCATTER AND SHAPE IN PROFILE SIMILARITY

HARVEY A. SKINNER
Addiction Research Foundation, Toronto

The aims of this paper are: (1) to present a similarity generating function composed of elevation, scatter and shape parameters; (2) to describe linear models for integrating these parameters either for euclidean distance or vector-product association indices; and (3) to suggest a computational strategy based upon the Eckart-Young (1936) theorem that has certain advantages for minimizing the effects of measurement error in estimating profile similarity. Given these developments, the investigator may differentiate the *independent* contribution of each parameter to more global indices of resemblance. A brief example from the classification of psychopathology is discussed.

IN their classic paper on assessing similarity between profiles, Cronbach and Gleser (1953) made the distinction among shape, scatter, and elevation. Shape denotes the actual pattern of "ups and downs" across variables in the profile, scatter describes how dispersed scores are from the average, while elevation is the mean score of the individual over all attribute measures in the profile. Many applications of numerical taxonomy use as a similarity index either an euclidean distance measure that confounds all three profile components, or correlation that emphasizes only shape. Clearly, a desirable research strategy should

Portions of this paper were presented at the joint meeting of the Psychometric Society and the Classification Society (NAB), April, 1975, University of Iowa, Iowa City. This research was supported by the Defence Research Board of Canada Grant Number 9435-75 (UG), and Canada Council Grant Number S74-0761. The author wishes to thank Douglas N. Jackson and Philip L. Reed for their helpful suggestions.

Reprint requests should be directed to Dr. Harvey A. Skinner, Addiction Research Foundation, 33 Russell Street, Toronto, Ontario, Canada. M5S 2S1.

allow one to differentiate the *independent* contribution of each similarity component. Then, the investigator may choose the appropriate set of parameters for his specific research problem.

Some investigators advocate the use of correlation in the study of individual differences in profile pattern (Gollob, 1968; Lorr, 1966; Overall and Klett, 1972). Lorr (1966) criticizes euclidean distance measures (D^2) because "a D^2 can represent a large difference between two individuals on only one dimension, or the sum of many small differences on all of the dimensions involved" (p. 25). Other researchers, such as Fleiss and Zubin (1969), favour distance measures because of problems with correlation. For example, how does one interpret a correlation of zero between two profiles? In an attempt to capitalize upon the advantages of alternative similarity indices, Guertin (1966) has suggested a two-stage strategy. First, correlation is employed to form homogeneous groups based on profile shape. Then, using a distance measure, each "shape" group is divided into subclasses having similar elevation and scatter.

When choosing among alternative resemblance coefficients, the important point to keep clear is the specific research goal. If the investigator is interested primarily in defining resemblance due to profile shape, then correlation or a generalized coefficient (Cohen, 1969) would be the statistic of choice. On the other hand, if the researcher wants to preserve the most information in his index, then euclidean distance or some variant such as Cattell's (1949) r_p would be appropriate. A third investigator, in the tradition of Guertin (1966), may wish to differentiate the independent contribution of elevation, scatter and shape parameters in more global resemblance measures, using the similarity generating function described below. Thus, the decision to employ a given resemblance index should be made with a full knowledge of what aspects of similarity are being assessed.

The aims of this paper are threefold. Firstly, a similarity generating function composed of elevation, scatter and shape parameters is developed. This function should provide a useful heuristic for the conceptual and operational organization of classification research. Secondly, linear models are presented integrating the similarity parameters either for euclidean distance or vector-product association indices. Finally, a computational strategy is suggested based upon the Eckart and Young (1936) theorem that has certain advantages for minimizing effects of measurement error in estimating profile similarity.

Similarity Generating Function

Consider a data matrix X of n entities by k attributes. For convenience, assume $n > k$. These entities may represent, for example, 200 psychiatric patients with scores on the 13 MMPI clinical scales. To

simplify the discussion, assume that the k attribute measures are of compatible scale (e.g., standard scores or column centered). Let $\bar{X}$ denote the row centering of X, that is, each entity's mean is subtracted from attributes in its profile row vector, while Z represents the row standardization of X. Note, in moving from X to $\bar{X}$ one is eliminating information due to elevation (entity mean), and in transforming $\bar{X}$ to Z one is equating all profiles with respect to scatter. Thus, Z contains information pertaining only to shape. The entity means m may be stored in column vector M, and standard deviations s stored in column vector S.

It is possible to express several quantitative measures of similarity ω_{12} between Entity 1 and Entity 2 as functions of the following parameters

$$\omega_{12} = f(m_1, m_2, s_1, s_2, r_{12}), \tag{1}$$

where r_{12} is the product-moment correlation coefficient between Entity 1 and Entity 2 profiles. The similarity generating function may be expressed in matrix terms as

$$\Omega = f(M, S, R), \tag{2}$$

where Ω is a symmetric matrix of association indices among entities, M is a column vector of elevation parameters, S is a column vector of scatter parameters, and R is a symmetric matrix of shape parameters (correlations). One final generalization is to consider the principal components factoring of the entity intercorrelations (Q-technique), that is,

$$R = A\,A' + E, \tag{3}$$

where A is an n by i component weighting matrix and E is an n by n residual matrix. Let $\hat{R}$ denote a rank i ($i < k$) approximation to R. Thus, the similarity generating function may be expressed formally as

$$\Omega = f(M, S, \hat{R}), \tag{4}$$

and specifically for two entities as

$$\omega_{12} = f\left(m_1, m_2, s_1, s_2, \hat{r}_{12} = \sum^{i} a_{1j}\, a_{2j}\right). \tag{5}$$

Potential advantages of considering the component weightings A and a reduced rank approximation $\hat{R}$ will be discussed below.

Vector-Product Indices

The average raw cross-products CP among entities is given by

$$CP = \frac{X\,X'}{k} \tag{6}$$

As discussed by Nunnally (1962), a factoring of this matrix yields results similar to euclidean distance measures. Consider the average cross-product entry for Entity 1 and Entity 2 over the k attributes. Following Horn (1969), this term may be expressed as

$$cp_{12} = m_1 m_2 + s_1 s_2 r_{12} \tag{7}$$
$$= f(\text{elevation} + (\text{scatter} \cdot \text{shape})).$$

Given this development, one may see why the raw cross-product term confounds all three profile components. For example a $cp_{12} = 100$ for Entities 1 and 2 could be due primarily to elevation, while a $cp_{13} = 100$ for Entities 1 and 3 may reflect a predominant similarity due to covariance (scatter · shape). No univocal interpretation regarding the similarity parameters may be attached to a single entry in CP. If covariances C are computed as a vector-product index among entities, that is,

$$C = \frac{\bar{X}\,\bar{X}'}{k}, \tag{8}$$

each entry would confound profile scatter with shape, viz,

$$c_{12} = s_1 s_2 r_{12} \tag{9}$$
$$= f(\text{scatter} \cdot \text{shape}).$$

Finally, an examination of correlations R among entities emphasizes similarity with respect to shape. One writes

$$R = \frac{Z\,Z'}{k} \tag{10}$$
$$= f(\text{shape}).$$

Euclidean Distance Indices

A similar development may be given for euclidean distance measures of profile similarity based on Cronbach and Gleser (1953). Given their equation 3 (p. 461), the following parameters may be defined (where the middle term is in their notation):

$$\text{elevation} = k\,\Delta^2 El_{12} = k(m_1 - m_2)^2; \tag{11}$$

$$\text{scatter} = \Delta^2 S_{12} = (\sqrt{k}\,s_1 - \sqrt{k}\,s_2)^2 \tag{12}$$
$$= k(s_1 - s_2)^2;$$

and

$$\text{shape} = s_1 s_2 D''_{12} = s_1 s_2 \sum^{k} (Z_{1j} - Z_{2j})^2$$

$$= s_1 s_2 \left(\sum^{k} Z_{1j}^2 + \sum^{k} Z_{2j}^2 - 2 \sum^{k} Z_{1j} Z_{2j} \right)$$

$$= s_1 s_2 (k + k - 2k r_{12})$$

$$= 2k s_1 s_2 (1 - r_{12})$$

$$= 2k (1 - r_{12}) . \tag{13}$$

since $s_1 = s_2 = 1$.

Thus, the euclidean distance measure D_{12}^2 between two entities averaged over the k attributes may be expressed as

$$D_{12}^2 = \frac{1}{k} \sum^{k} (X_{1j} - X_{2j})^2$$

$$= (m_1 - m_2)^2 + (s_1 - s_2)^2 + 2s_1 s_2 (1 - r_{12})$$

$$= f[\text{elevation} + \text{scatter} + (\text{scatter} \cdot \text{shape})]. \tag{14}$$

Note that each D^2 entry is complexly determined by components due to (1) elevation, (2) scatter, plus (3) a scatter by shape interaction. Penrose (1954) describes an analogous solution for differentiating D_{12}^2 into a 'size' component (our elevation) and a 'shape' component (our scatter and shape). To complete this development, observe that if the average euclidean distance is computed using deviation scores $\bar{X}$, then scatter and shape components are confounded similar to equation 9 for covariances. One writes

$$D_{12}^2 = \frac{1}{k} \sum^{k} (\bar{X}_{1j} - \bar{X}_{2j})^2 \tag{15}$$

$$= (s_1 - s_2)^2 + 2s_1 s_2 (1 - r_{12})$$

$$= f(\text{scatter} + (\text{scatter} \cdot \text{shape})).$$

Finally, euclidean distance measures based on Z emphasize similarity solely as a function of shape, that is,

$$D_{12}^2 = \frac{1}{k} \sum^{k} (Z_{1j} - Z_{2j})^2 \qquad (16)$$

$$= 2(1 - r_{12})$$

$$= f(\text{shape}).$$

Thus, given the relationships among the elevation, scatter and shape parameters summarized in equation 7 for vector-product indices and equation 14 for euclidean distance measures, the global similarity index may be decomposed into independent components reflecting the contribution of each parameter. The similarity generating function (equation 4) provides an efficient method of calculating a desired association index from among the set of euclidean distance and vector-product coefficients (cf. Wishart, 1972). Note that two other popular association indices, Cattell's (1949) r_p and the Mahalanobis distance function (Sneath and Sokal, 1973), confound elevation, scatter and shape since both coefficients are derived from calculations based on the original data matrix X (rather than $\bar{X}$ or Z).

Consider a research problem, such as the classification of psychopathology, where the investigator is concerned with differentiating each similarity parameter. A logical strategy is first to place patient profiles into relatively homogeneous subgroups on the basis of shape, through application of some clustering or ordination algorithm upon either R (equation 10) or D^2 (equation 16). Then, at a second stage, one may potentially differentiate subjects *within* each shape subgroup with respect to scatter, and at a third stage with respect to profile elevation. This strategy is similar in rationale to Morrison's (1967, p. 141) discussion of profile analysis for independent groups following a significant multivariate test (e.g., Hotelling T^2).

For example, given a sample of 300 subjects depicted in Figure 1, the first stage may delineate two shape subgroups. By plotting the distribution of scatter parameters, the A shape subset may be further differentiated into two (shape + scatter) subgroups, while subjects in the B subset may be homogeneous with respect to scatter. Finally, the third stage would potentially distinguish within subgroups on the basis of profile elevation. Thus, psychiatric and normal individuals may be placed in the same group when considering profile shape, for example, a Code 2-7-4 on the MMPI. However, the psychiatric patients may prove distinguishable from normals when examining profile elevation, where the psychiatric patient highpoints would presumably exceed a T score of 70. This strategy directly achieves the same objectives as Guertin's (1966) proposed two-stage analysis.

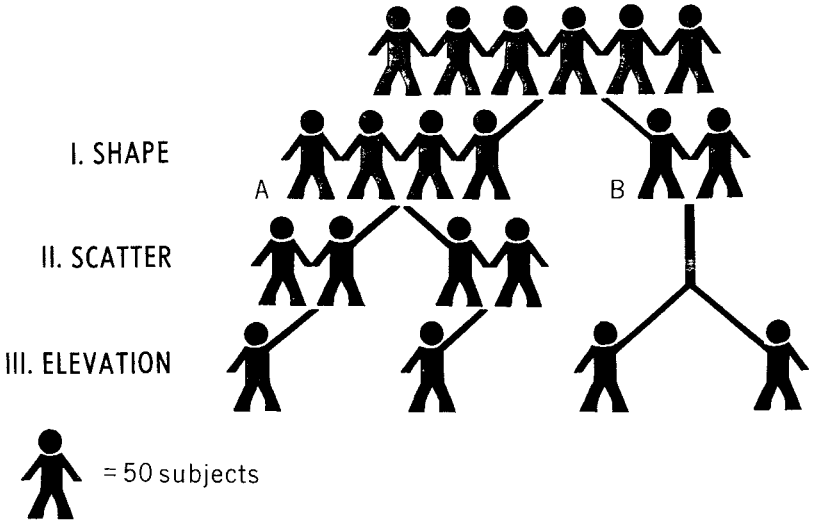

Figure 1. Differentiating individuals according to profile shape, scatter, and elevation.

Computational Strategy

The first step is to standardize the data matrix X by row to yield Z, while saving the vector of entity means M and standard deviation S for the similarity generating function. The Z matrix (scaled $1/\sqrt{k}$) may be decomposed by the Eckart-Young (1936) theorem into

$$Z = P \Delta Q', \tag{17}$$

where the left-hand eigenvectors P of order n by k describe the configuration among entities, the right-hand eigenvectors Q of order k by k display the configuration among attributes, and Δ is a k by k diagonal matrix of singular values whose nonzero entries determine the rank of Z. The principal component weightings A among entities analogous to equation 3 are given by

$$A = P \Delta. \tag{18}$$

At this point one could enter the complete n by k principal component weightings A into the similarity generating function. However, a more interesting possibility is to consider only the first i components of A using some criterion such as the scree test, since the residual components should largely reflect measurement error (Gleason and Staelin, 1973). From the Eckart-Young (1936) theorem, A_i provides a rank i least-squares estimation of R, designated $\hat{R}$. Thus, the reduced rank approximation to the correlation between two entity profiles $\hat{r}_{12}$ de-

picted in equation 5 should generally provide a better estimate of the "true" shape parameter.

For some research problems, the investigator may decide to examine these entity principal components as an ordination method of classification (Skinner, 1977). To enhance interpretability of the data, the i entity components could be transformed (rotated) to some criterion of simple structure such as varimax. To compute the n by i matrix B_i of transformed weightings, one writes

$$B_i = A_i T, \tag{19}$$

where T is an i by i orthonormal transformation matrix. Corresponding component scores may be readily generated relating each attribute measure to the first i transformed entity components, that is,

$$Y_i = Q_i T, \tag{20}$$

where Y_i is a k by i matrix of component scores. Observe that the model in Equation 17 may be represented as

$$Z = B_i Y_i' + E, \tag{21}$$

where the trace of the residual matrix, Trace $E'E$, is a minimum for a given rank i solution.

These component scores describe the projection of each profile attribute on principal components of the entity factor space (Skinner, 1977). With increasing availability of computer programs for the Eckart-Young (1936) theorem, or singular value decomposition (Stewart, 1973), decomposing the data matrix itself (equation 17) is a more direct strategy for computing component weightings and scores than the traditional principal components factoring of cross-product matrices (Horst, 1965, chapter 4).

Example

A brief research example should consolidate the major points of this paper. The data are taken from Lanyon (1968) who compiled mean Minnesota Multiphasic Personality Inventory profiles (T scores) for 233 diverse psychiatric and normal groups. In this example, each entity actually represents the mean MMPI profile for a group of subjects. The 233 by 13 data matrix was row standardized and decomposed as described in equations 17 to 20. Four entity components accounting for 82.7% of the total variance were transformed to a varimax criterion. Corresponding component scores, listed in Table 1, may be interpreted as modal MMPI profiles (Skinner, 1977) that are characteristic of frequently occurring profiles or subsets among the

TABLE 1
MMPI Component Scores (Modal Profiles)

Component	MMPI Clinical Scale												
	L	*F*	*K*	*Hs*	*D*	*Hy*	*Pd*	*Mf*	*Pa*	*Pt*	*Sc*	*Ma*	*Si*
I	50	69	50	49	54	53	76	49	65	65	75	71	53
II	47	55	49	73	80	71	64	54	57	66	64	50	50
III	42	49	71	55	56	66	67	80	51	60	59	70	54
IV	61	51	77	62	53	72	74	39	63	57	54	58	60

Note—The component scores have been scaled to have a mean = 60 and standard deviation = 10.

233 Lanyon (1968) groups. The component weightings in A_i specify the degree of similarity (shape only) of each group profile to the MMPI component score profiles in Table 1.

For illustrative purposes, Figure 2 presents the MMPI profiles for two male Delinquent groups, while Figure 3 displays MMPI profiles of two Suicide groups. The similarity generating function parameters and various euclidean distance and vector-product association indices are given in Table 2. Although both Delinquent groups have a shape (.95 and .87 respectively) markedly similar to the first MMPI component score profile of Table 1, the Social group has lower scatter and elevation parameters. Similarly, both Suicide groups resemble the shape (.82 and .96 respectively) of the second MMPI component score profile of Table 1, while the Attempt group is lower on scatter and elevation. In this example, one might hypothesize that the shape

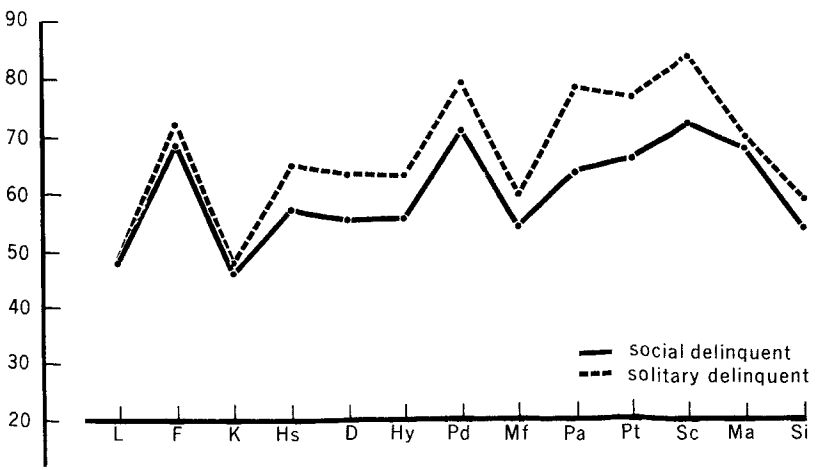

Figure 2. Mean MMPI profile for a group of Social Delinquents (N = 39) and a group of Solitary Delinquents (N = 18) from Lanyon (1968).

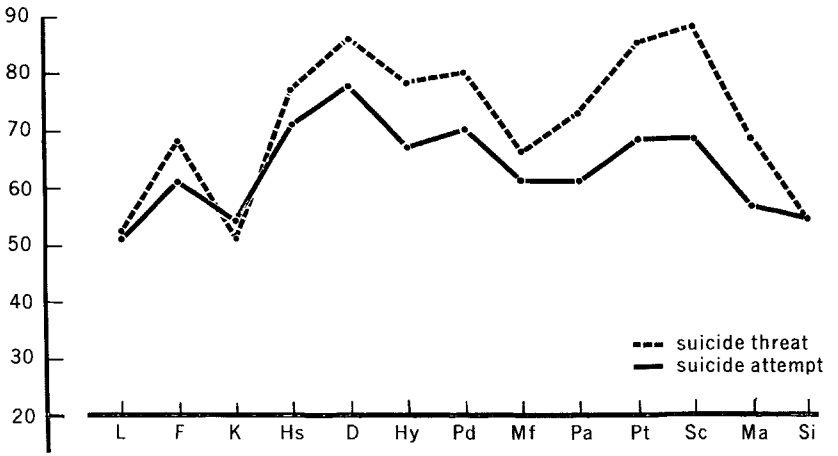

Figure 3. Mean MMPI profile for a Suicide Threat group (N = 32) and a Suicide Attempt group (N = 32) from Lanyon (1968).

parameters indicate more enduring personality dispositions, whereas scatter and elevation reflect temporary factors influencing the degree of symptom severity. Further research could evaluate the temporal stability of these parameters following treatment or therapy programs.

Note that the cross-product term cp in Table 2 is almost totally

TABLE 2
Profile Similarity Data

Similarity Parameters		Delinquent Groups		Suicide Groups	
		Social	Solitary	Threat	Attempt
Mean (M)		59.54	65.69	70.89	63.13
Standard Deviation (S)		8.26	10.43	12.38	7.62
Component Weightings (A)	I	.95	.87	.51	.20
	II	.17	.38	.82	.96
	III	−.01	.05	.08	.07
	IV	−.15	−.15	−.18	−.09
Similarity Indices					
Correlation r		.96		.91	
(estimated) $\hat{r}$		.91		.90	
Covariance c		78.40		84.90	
Cross-Product cp		3989.58		4560.18	
D^2 (shape)		.18		.20	
D^2 (shape + scatter)		20.22		41.53	
D^2 (shape, scatter and elevation)		58.04		101.75	

Note—The c, cp and D^2 incides were computed using $\hat{r}$ instead of r.

TABLE 3
Elements of the Euclidean Distance D^2 Measure (Equation 14)

	Elevation $(m_1 - m_2)^2$	Scatter $(s_1 - s_2)^2$	Scatter · Shape $2s_1s_2(1 - \hat{r}_{12})$	Total D^2
Delinquent Groups	37.82 (65.16%)	4.71 (8.12%)	15.51 (26.72%)	58.04
Suicide Groups	60.22 (59.18%)	22.66 (22.27%)	18.87 (18.55%)	101.75

Note—The D^2 index is based on the estimated shape parameter $\hat{r}$ instead of r.

dominated by the elevation parameter (equation 7), a fact that is generally reflected in the first principal axes component of *CP* (Gollob, 1968). Table 3 presents a breakdown of the overall D^2 index (equation 14) according to the various parameters. In both the Delinquent groups and Suicide groups comparisons, the elevation component clearly dominates, contributing approximately 60 per cent. The shape parameter, on the other hand, tends to be overwhelmed by the elevation and scatter parameters in determining the overall D^2 measure. This may be an undesirable property since in most psychological research profile shape assumes primary importance. One suggestion is that researchers should examine the contribution of each similarity component to the D^2 index (Table 2), before the D^2 matrix is input to a clustering algorithm. Then, the investigator would have some "feel" for the relative importance of each similarity parameter in determining the clusters.

Although the differentiation among elevation, scatter and shape may be inappropriate for many research applications, investigators should be cognizant of this potential distinction. Aside from its heuristic value, the similarity generating function of equation 4 could be routinely incorporated as a preliminary stage of any clustering or ordination algorithm. Then, researchers may be encouraged to run several analyses on the same entities, highlighting alternative aspects of profile resemblance. This concluding remark, urging the assessment of varied similarity parameters, is consistent with the observation of Sneath and Sokal (1973) that "it is not at all clear at this point that a unique measure of similarity . . . is possible or even desirable" (p. 31).

REFERENCES

Cattell, R. B. r_p and other coefficients of pattern similarity. *Psychometrika,* 1949, 14, 279–298.

Cohen, J. r_c: A profile similarity coefficient invariant over variable reflection, *Psychological Bulletin,* 1969, 71, 281–284.

Cronbach, L. J. and Gleser, G. Assessing similarity between profiles. *Psychological Bulletin,* 1953, 6, 456–473.

Eckart, C. and Young, G. The appoximation of one matrix by another of lower rank. *Psychometrika,* 1936, 1, 211–218.

Fleiss, J. L. and Zubin, J. On the methods and theory of clustering. *Multivariate Behavioral Research,* 1969, 4, 235–250.

Gleason, T. C. and Staelin, R. Improving the metric quality of questionnaire data. *Psychometrika,* 1973, 38, 393–410.

Gollob, H. F. Confounding of sources of variation in factor-analytic techniques. *Psychological Bulletin,* 1968, 70, 330–344.

Guertin, W. H. The search for recurring patterns among individual profiles. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1966, 26, 151–165.

Horn, J. L. Factor analysis with variables of different metric. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1969, 29, 753–762.

Horst, P. *Factor analysis of data matrices.* New York: Holt, Rinehart & Winston, 1965.

Lanyon, R. I. *A handbook of MMPI group profiles.* Minneapolis: University of Minnesota Press, 1968.

Lorr, M. *Explorations in typing psychotics.* Oxford: Pergamon Press, 1966.

Morrison, D. F. *Multivariate statistical methods.* New York: McGraw-Hill, 1967.

Nunnally, J. C. The analysis of profile data. *Psychological Bulletin,* 1962, 59, 311–319.

Overall, J. E. and Klett, C. J. *Applied multivariate analysis.* New York: McGraw-Hill, 1972.

Penrose, L. S. Distance, size and shape. *Annals of Eugenics,* 1954, 18, 337–343.

Skinner, H. A. The eyes that fix you: A model for classification research. *Canadian Psychological Review,* 1977, 18, 142–151.

Sneath, P.H.A. and Sokal, R. R. *Numerical taxonomy.* San Francisco: Freeman, 1973.

Stewart, G. W. *Introduction to matrix computations.* New York: Academic Press, 1973.

Wishart, D. A general tripartite clustering method and similarity generating function. Report No. R-31, Statistics 1 Division, Civil Service Department, Whitehall, London, England, April, 1972.