EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, *26*, 825–831.

SIMILARITY ANALYSIS BY RECIPROCAL PAIRS FOR DISCRETE AND CONTINUOUS DATA

LOUIS L. McQUITTY
Michigan State University

THE similarity index was developed as a central technique in an early method of hierarchical pattern analysis, called similarity analysis (McQuitty, 1955).

Similarity analysis has the advantage of being applicable to both discrete and continuous data. However, in its original form, it has the disadvantages of being complicated and laborious, sometimes requires iteration, and can lead to inconsistencies (McQuitty, 1955). By application of a particular theory of types (McQuitty, 1966) all of these problems can be solved, and a very simple method of hierarchical analysis can be developed, called Similarity Analysis by Reciprocal Pairs, applicable to both discrete and continuous data.

Theory

The theory says that every individual represents a succession of types, first an individual type, then types analogous to a species, a genus, a family, etc. As more and more individual types are classified together to represent higher and higher orders of hierarchical types, the successive categories become better representatives of pure types, which exist only in theory.

Individuals and hierarchical categories of the above kind are jointly characterized as typal representatives.

Every typal representative at any level x is best classified at the next higher level if it is classified with the typal representative most like it at level x and if the two representatives are reciprocal, i.e., Typal Representative i is most like j, and j is in turn most like i (McQuitty, 1957).

The Basic Equation

Let:

ij = any typal representative formed by combining the two typal representatives i and j from the next lower level, level x.

k = any typal representative other than i and j from level x.

a_{ik} = an index of association between i and k.

a_{jk} = an index of association between j and k.

a_{ij-k} = an index of association between ij and k.

Then:

$$a_{ij-k} = \frac{a_{ik} + a_{jk}}{2}. \tag{1}$$

An Illustration

In order to illustrate the method, it was applied to the data of Table 1, coefficients of correlation between people for the second through the sixth persons of a matrix from Stephenson (1953, p. 169).

TABLE 1

Coefficients of Correlation between People

	B	C	D	E	F
B		60	49	24	43
C	60		62	46	51
D	49	62		40	57
E	24	46	40		67
F	43	51	57	56	

Note—Data from Stephenson (1953, p. 169); entries have been rounded from two to three places and decimal points have been omitted.

The coefficients of correlation were first converted to variances by squaring them and are shown in Matrix 1 of Table 2. This is not essential in most cases but does help increase the accuracy in details to the extent that variances represent better units than do coefficients of correlation.

The first step is to underline the highest entry in each column of Matrix 1, Table 2; it is 36 for Column B, 38 for C, 38 for D, and 31 for each E and F. Then, select the highest entry in the entire matrix; it is 38 and mediates between individuals C and D. In-

TABLE 2

A Similarity Analysis of Variance Indices Computed from the Correlation Coefficients of Table 1

	B	C	D	E	F		B	CD	E	F		BCD	E	F		BCD	EF
B		36	27	06	18	B		$\underline{32}$	06	18	BCD		13	24	BCD		19
C	$\underline{36}$		$\underline{38}$	21	26	CD	$\underline{32}$		19	29	E	13		$\underline{31}$	EF	19	
D	27	$\underline{38}$		16	32	E	06	19		$\underline{31}$	F	$\underline{24}$	$\underline{31}$				
E	06	21	16		$\underline{31}$	F	18	29	$\underline{31}$								
F	18	26	32	$\underline{31}$													
			Matrix 1					Matrix 2					Matrix 3			Matrix 4	

dividuals C and D jointly are accepted as a better representative of a hierarchical type than is either separately. They are joined as a type and placed in both Row CD and Column CD of Matrix 2, Table 2.

Using the basic equation, the entry for both Row B—Column CD and Row CD—Column B is:

$$a_{CD-B} = \frac{a_{B-C} + a_{B-D}}{2} = \frac{36 + 27}{2} = \frac{63}{2} = 31.50, \text{ rounded to } 32. \quad (2)$$

The other entries of Row and Column CD are computed in an analogous fashion. The entries for the remaining cells of Matrix 2 are taken from the corresponding cells of Matrix 1.

The highest entry in each column of Matrix 2 is underlined and the highest entry in the matrix is chosen, 32 for Row B—Column CD and Column CD—Row B.

Typal Representatives B and CD are joined to form a higher Typal Representative, BCD, and are entered as a row and column of Matrix 3.

The basic equation is used to compute the entry for Row E—Column BCD and Row BCD—Column E.

$$a_{BCD-E} = \frac{a_{E-B} + a_{E-CD}}{2} = \frac{6 + 19}{2} = \frac{25}{2} = 12.50, \text{ rounded to } 13. \quad (3)$$

The entry for Row F—Column BCD and Row BCD—Column F is computed analogously. The other entries of Matrix 3 are taken from their corresponding cells of Matrix 2.

The highest entry in each column of Matrix 3 is underlined, and the highest entry in the entire matrix is determined, 31 for Row E—Column F and Row F—Column E. Typal Representatives E and F are therefore combined to form a new type at a higher level, Typal Representative EF. This completes the analysis. However, for purpose of further illustration, we show how the similarity index for Row BCD—Column EF and Row EF—Column BCD is computed:

$$a_{EF-BCD} = \frac{a_{E-BCD} + a_{F-BCD}}{2}$$

$$= \frac{13 + 24}{2} = \frac{37}{2} = 18.50, \text{ rounded to } 19. \quad (4)$$

A Critique

All reciprocal pairs of any matrix could have been operated on in preparing a new matrix. For example, Matrix 1 contains two reciprocal pairs, CD and EF. Each of these could have been combined to yield Matrix 5 as shown in Table 3.

TABLE 3

An Alternative Similarity Analysis of Matrix 1, Table 2

	B	CD	EF		BCD	EF
B		32	12	BCD		18
CD	32		24	EF	18	
EF	12	24				
	Matrix 5				Matrix 6	

The entry for Row B—Column CD and Row CD—Column B is computed in the same fashion as it was for Matrix 2, Table 2, and likewise for the entry of Row B—Column EF and Column EF—Row B.

The entry of Row CD—Column EF and Row EF—Column CD was, however, computed in a slightly different fashion; two applications of the basic equation were involved:

$$a_{CD-E} = \frac{21 + 16}{2} = 18.50, \text{ rounded to } 19. \tag{5}$$

$$a_{CD-F} = \frac{26 + 32}{2} = 29. \tag{6}$$

$$a_{CD-EF} = \frac{21 + 16 + 26 + 32}{4} = 23.75, \text{ rounded to } 24 \tag{7}$$

or

$$a_{CD-EF} = \frac{19 + 29}{2} = 24. \tag{8}$$

Stated in general notation, Equation 7 states,

$$a_{ij-kl} = \frac{a_{ik} + a_{il} + a_{jk} + a_{jl}}{2}. \tag{9}$$

This entire equation can be generalized to:

$$s_{xy} = \frac{\sum_{x=1}^{x=n} \sum_{y=1}^{y=m} a_{xy}}{n \cdot m}. \tag{10}$$

s_{xy} = the original similarity index (McQuitty, 1955).
x = Type x represented by Individuals $x_1, x_2, x_3, \cdots n$.
y = Type y represented by Individuals $y_1, y_2, y_3, \cdots m$.

The difference in Equations 1 and 10 derives from a difference in theory. Equation 1 says that each type should be given equal weight when they are joined to form an hierarchical type of the next higher level. Equation 10, on the other hand, says that the two types being joined should be given differential weights in relation to the number of individuals in each of them. Equation 7 (a special case of Equation 10) gave the same results as Equation 1, because in the special case of Equation 7 each CD and EF were composed of two individuals and were therefore given equal weights.

The theory of this paper recommends Equation 1; Equation 1 was derived from the theory of this paper and has the advantage of being very simple to apply to both discrete and continuous data.

An alternative theory of types could require differential weights in relation to the number of individuals in each of two types being joined to form a new type at the next higher level. An equation simpler to apply than Equation 10 can be developed for the purpose.

Let:

i = a type just realized by combining Types j and k; it will replace Types j and k in the next matrix.
j = a type composed of Individuals $b_1, b_2, b_3, \cdots b_n$.
k = a type composed of Individuals $c_1, c_2, c_3, \cdots c_m$.
l = any other type realized earlier by combining Types x and y.
x = a type composed of Individuals $d_1, d_2, d_3, \cdots d_r$.
y = a type composed of Individuals $e_1, e_2, e_3, \cdots e_s$.
a_{il} = the similarity index between Types i and l.
a_{jl} = the similarity index between Types j and l.
a_{kl} = the similarity index between Types k and l.

Then:

$$a_{il} = \frac{n(a_{jl}) + m(a_{kl})}{n + m} \tag{11}$$

Differential weights are used for Type i in the above application of Equation 11; they are represented by n and m, showing the numbers of individuals in each Type j and k which combined to yield Type i. Differential weights are not, however, used for Type l in the above application of Equation 11; they were used earlier in the analysis when Equation 11 was applied to compute each a_{jl} and a_{kl}; they are represented by r and s.

Similarity Analysis by Reciprocal Pairs can be used to analyze any matrix (even one generated out of chance) into statistical types, for every matrix has at least one reciprocal pair. Consequently, the investigator should be especially aware of the need to find other evidence before interpreting the results as indicative of substantive types.

Summary

This paper generates an especially simple method for analyzing both continuous and discrete data into hierarchical types.

REFERENCES

McQuitty, L. L. A Method of Pattern Analysis for Isolating Typological and Dimensional Constructs, Lackland Air Force Base, Texas. *AFPTRC Research Bulletin*, 1955, *TN-55-62.*

McQuitty, L. L. Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1957, 17, 207-229.

McQuitty, L. L. Single and Multiple Hierarchical Classification by Reciprocal Pairs and Rank Order Types. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1966, 26, 253-265.

Stephenson, W. *The Study of Behavior*. Chicago: The University of Chicago Press, 1953.