The Statistical Measurement of Agreement

W. S. Robinson

American Sociological Review, Vol. 22, No. 1 (Feb., 1957), 17-25.

Stable URL:
http://links.jstor.org/sici?sici=0003-1224%28195702%2922%3A1%3C17%3ATSMOA%3E2.0.CO%3B2-A

American Sociological Review is currently published by American Sociological Association.

theorems of O'M•.[41] This means that von Wright's system is a subsystem of O'M•; and that some account is also taken of Bohnert's suggestions. It would be pointless to try to characterize O'M• further in such a brief compass; formal systems always require intensive study if their explanatory power is to be fully appreciated.

If the "explanatory power" of such constructs seems highly abstract or remote, the situation should be compared with contemporary physical theories, where mathematical theories much more recondite than those presented here have a direct and immediate bearing on empirical research. And just as many mathematical developments were initiated by empirical problems, so the development of O'M• was motivated by the hope of throwing light on some empirical problems in small-group research. More specifically, in experiments currently being conducted in the Yale Interaction Laboratory, opportunities are available for studying, under well-controlled conditions, small groups in the process of developing normative structures. A clear and accurate account of these interactional processess calls for a precise and rigorous conceptual framework within which to characterize the behavior of the group.

Small-groups research is of course only one, and not necessarily the most important, area of application of such formal systems. It is hoped that the foregoing discussion will have suggestive value for workers in other sociological fields, and will stimulate interest in current research in mathematical logic, especially modal logic. This research is not remote from the daily affairs of sociologists. For instance, whenever instructions are given for filling out a questionnaire, commands expressing obligations are involved. More generally, any adequate sociological theory must encompass, in our opinion, the concepts *norm, obligation*, etc. It is therefore a matter of importance to develop sound techniques for analyzing norms and systems of norms.

[41] As applied to contingent propositions, that is; this qualification applies elsewhere as well.

SUMMARY

The development of an adequate theoretical structure for sociology will in all likelihood require interdisciplinary co-operation between sociologists and those working in the formal sciences if it is to proceed in a maximally fruitful way. The purpose of this article is to bring to the attention of sociologists recent work in mathematical logic which has direct relevance for their research. Von Wright's deontic logic offers promising leads for the analysis of normative structures; the family of systems of which O'M• is a member provides a more comprehensive framework, taking account of the role of penalties and possibilities for action *vis-a-vis* norms.

THE STATISTICAL MEASUREMENT OF AGREEMENT

W. S. ROBINSON
University of California, Los Angeles

A NUMBER of statistical problems of importance in social research logically require the measurement of the *agreement* (rather than the correlation) between two or more variables. These problems are often handled individually as they arise and on an *ad hoc* basis. Appropriate techniques, however, are available for handling them, though they have hitherto seen use mainly in non-sociological fields. The purpose of this paper is to discuss these techniques in a general setting, and to indicate the importance of agreement as a statistical concept in its own right.

The idea of agreement appears in a variety of situations, among which are problems involving: the interchangeability of measures, as in the substitution of an index for a more fundamental measure; measurement of the reliability of an instrument such as a test or scale; measurement of the "objectivity" or lack of bias of a concept as applied to observational material by different observers; comparison of observed with theo-

retically deduced values of a variable; and measurement of the degree of homogeneity within "families" of observations.

A definition of agreement from first principles is shown in this paper to lead in the case of two variables to the intraclass correlation as a measure of agreement comparable to the Pearsonian correlation as a measure of correlation. The discussion leads to a precise and meaningful interpretation for the difference between a Pearsonian and an intraclass correlation computed from the same data. It also leads to the suggestion that a logically correct estimate of the reliability of a test is given by the intraclass rather than the Pearsonian correlation, and hence that the concept of agreement rather than correlation is the basis of reliability theory. It leads for the case of more than two variables to the suggestion of a more appropriate measure of agreement than that provided by the intraclass correlation. In all, by its repeated distinction between agreement and correlation, the paper is intended as an essay in conceptual clarification.

An Occasion for a Measure of Agreement. It will be useful to begin by describing a research problem in which the need for a measure of agreement arises naturally and obviously from the purpose for which data were gathered. In 1940, Lundberg [1] presented the results of an experiment to answer the question, "What is the degree of agreement in 'commonsense' judgments of socioeconomic status of the population of a community by two persons who are themselves of radically different socioeconomic status, that is, how are informal ratings of socioeconomic status influenced by the socioeconomic status of the rater?" [2] Lundberg chose a New England village of about 300 families as the field for his experiment. He asked a local banker of the community and a local janitor, both residents for not less than forty years, to rate on a six-point scale the socioeconomic status of the families they knew in common. The resulting 196 pairs of ratings constitute the data by which Lundberg answered his question.

[1] George A. Lundberg, "The Measurement of Socioeconomic Status," *American Sociological Review*, 5 (February, 1940), pp. 29–39.

[2] *Op. Cit.*, p. 29.

Inadequacy of Common ad hoc *Measures of Agreement.* All information supplied by the results of an experiment to measure the "objectivity" of a concept as applied by two raters is given by a correlation table relating the scores of the two observers.[3] Such a table is given here as Table 1, which was reconstructed (though not uniquely) from the data in Lundberg's article. Table 1 gives the joint distribution of janitor's and banker's ratings, showing how many families were assigned each possible combination of janitor's and banker's ratings.[4]

TABLE 1. RELATION BETWEEN JANITOR'S AND BANKER'S RATINGS OF THE SOCIOECONOMIC STATUS OF 196 FAMILIES

Banker's Rating	Janitor's Rating					
	1	2	3	4	5	6
6	–	–	–	–	–	–
5	–	–	–	6	**8**	8
4	–	1	–	**21**	27	–
3	–	1	**25**	47	13	2
2	–	**4**	6	4	1	–
1	**3**	4	11	3	1	–

Two methods of measuring agreement in a situation like this are commonly used:

(1) One is to find the percentage of judgments for which there is agreement, or for which there is agreement within a specified range. For example, every family recorded in the bold-face diagonal of Table 1 was rated identically by janitor and banker. There are 61 such families, and therefore in $(100)(61)/196 = 31$ per cent of the judgments there was agreement. Or sometimes an acceptable range of agreement is defined, and a percentage computed for it. For example, each family recorded in the diagonals immediately above and below the bold-faced one was rated only one unit differently by janitor and banker. If one

[3] Extension of the argument to three or more observers is made later.

[4] This example was chosen for its sociological pertinence but with some question, at least for the statistical purist, of the legitimacy of applying the methods discussed. For complete rigor of application, the variables of Table 1 should be strictly metricized variables, and not ratings in which the unit is not known to be constant. This question is discussed under the heading *Tests of Significance.*

relaxes the definition of agreement to include differing by one unit, then the percentage of judgments in Table 1 for which there was agreement becomes 82.

The inadequacy of this method is that it arbitrarily selects certain judgments and ignores others. Moreover, the percentage of judgments in which there is agreement depends upon the range of deviation from perfect agreement that one is willing to define as "agreement." It is the entire table that truly expresses the degree to which janitor and banker agree.

(2) A second method is to use the Pearsonian correlation between the ratings as a measure of their agreement. This method has the advantage that it uses all of the observations, and also that it allows one to talk in terms of variance allocation, e.g., to say that 42 per cent of the variation in the janitor's scores is allocable to variation in the banker's scores, and *vice versa*,[5] but it also obscures the true nature of agreement.

The Pearsonian correlation is an inadequate measure of agreement because it measures the degree to which the paired values of the two variables are proportional (when expressed as deviations from their means) rather than identical. Agreement would be perfect in Table 1 if all observations fell in the bold-faced diagonal, and, to be sure, the Pearsonian correlation would then be unity. However, the Pearsonian correlation would also be unity if all observations fell in the diagonal directly below the bold-faced one, indicating that the janitor had rated each family one unit higher than the banker. A Pearsonian correlation of unity thus would not support by itself the claim that the janitor's and banker's ratings were in perfect agreement.

Correlation is measured by the degree to which values of a dependent variable are approximated by (or agree with) the best-predicting linear adjustment of the corresponding values of an independent variable. There is no point so far as agreement is concerned in the distinction between a dependent and an independent variable. There is even no sense in restricting the problem to two variables; one might have three or a dozen raters. The problem in measuring agreement is to find the degree to which the values of one variable approximate the corresponding values of another (or several others) *without adjustment*. Agreement, that is, is correlation without regression, and fundamentally a multivariate rather than a bivariate relation.

Agreement requires that paired values be identical, while correlation requires only that paired values be linked by a linear relationship, or, if one defines correlation more broadly, that the paired values be linked according to some mathematical function. Perfect agreement has but one form, $X_1 = X_2$, whereas correlation may variously be written $X_1 = a + bX_2$, $X_1 = a + bX_2 + cX_2^2$, $X_1 = \log X_2$, etc. Thus agreement is a special case of correlation, since two variables that agree must be correlated, but variables which are correlated do not necessarily agree. The research significance of the distinction is that agreement must be measured against the model $X_1 = X_2$, while correlation may be measured against any functional relationship as model, often the rectilinear one $X_1 = a + bX_2$.

A Fundamental Measure of Agreement. It will be useful to start from first principles in discussing the measurement of agreement. The degree to which two observations fail to be identical, or fail to agree, is measurable by their variance, or more conveniently by their sum of squares (of deviations from their mean). If the two values are identical, their sum of squares, or disagreement, is zero, and agreement is perfect. If the difference between the two values is large, their sum of squares, or disagreement, will likewise be large. In fact, *the sum of squares for any pair of values is half the square of the difference between them*, which is the primitive evidence of disagreement. That is

$$(X_{1j} - \overline{X}_j)^2 + (X_{2j} - \overline{X}_j)^2 = (X_{1j} - X_{2j})^2/2, \quad (1)$$

where

X_{1j} = the value of X_1 for the j-th pair;
X_{2j} = the value of X_2 for the j-th pair;
$\overline{X}_j$ = the mean of X_1 and X_2 for the j-th pair.

For a single pair, then, $(X_{1j} - \overline{X}_j)^2 + (X_{2j} - \overline{X}_j)^2$, or the sum of squares, meas-

[5] The "42 per cent" is the square of the Pearsonian correlation for Table 1 multiplied by 100.

ures the lack of agreement between the values of X_1 and X_2. The sum of these N sums of squares, one for each pair, is thus a sum of squares measuring the total within-pair lack of agreement for the N pairs of values. Let this sum be D (for Disagreement), so that

$$D = \Sigma(X_{1j} - \overline{X}_j)^2 + \Sigma(X_{2j} - \overline{X}_j)^2. \quad (2)$$

While D is a measure of lack of agreement for the entire sample of N pairs, it is not a very useful measure because it involves the units of X_1 and X_2. To find a relative rather than an absolute measure of disagreement, it is necessary to relate D to its range of possible variation. The least value D can have is zero, which occurs when the pair members agree perfectly for all pairs. Let the greatest possible value of D for a given sample of N pairs be D_{max}. Then D/D_{max} is the fraction the observed D is of its maximum possible value, i.e., a relative measure of disagreement. D/D_{max} will be zero when agreement is perfect, and unity when maximum possible disagreement is present. A coefficient of agreement defined as

$$A = 1 - D/D_{max} \quad (3)$$

will then be unity when agreement is perfect and zero when the paired values disagree maximally.

The greatest possible value of D for N pairs of observations is the sum of the squares of the deviations of all 2N observations from their common mean,[6] or

$$\Sigma(X_{1j} - \overline{X})^2 + \Sigma(X_{2j} - \overline{X})^2, \quad (4)$$

where

$$\overline{X} = (\Sigma X_{1j} + \Sigma X_{2j})/2N. \quad (5)$$

The coefficient of agreement defined in (3) is then

$$A = 1 - \frac{\Sigma(X_{1j} - \overline{X}_j)^2 + \Sigma(X_{2j} - \overline{X}_j)^2}{\Sigma(X_{1j} - \overline{X})^2 + \Sigma(X_{2j} - \overline{X})^2}. \quad (6)$$

In the special case of two variables, the agreement coefficient of (3) and (6) is very simply related to the intraclass correlation coefficient.[7] The relation is not so simple in the case of three or more variables, however, and it will be argued later that the coefficient of agreement defined here is for general use preferable to the intraclass correlation when the number of variables exceeds two.

The intraclass correlation between N pairs of observations on two variables X_1 and X_2 is by definition[8] the ordinary Pearsonian (interclass) correlation between 2N pairs of observations, the first N of which are the original observations, and the second N the original observations with X_1 replacing X_2 and *vice versa*. Thus if one has three pairs of values of X_1 and X_2, e.g.,

X_1	X_2
1	2
3	7
8	12

the intraclass correlation between X_1 and X_2 is the ordinary Pearsonian correlation between the six pairs of values:

X_1	X_2
1	2
3	7
8	12
2	1
7	3
12	8

Certain computational simplifications follow from this reversal of the variables, mainly because it makes the marginal distributions for the new variables the same, and therefore the means and variances of the new variables the same.

[6] This follows from the analysis of variance. There are N pairs of observations on X_1 and X_2, or 2N observations in all. The fundamental equation of the analysis of variance in this case is

$$\begin{bmatrix}\text{Total}\\ \text{Sum of}\\ \text{Squares}\end{bmatrix} = \begin{bmatrix}\text{Sum}\\ \text{of Squares}\\ \text{Within Pairs}\end{bmatrix} + \begin{bmatrix}\text{Weighted Sum}\\ \text{of Squares Be-}\\ \text{tween Pairs}\end{bmatrix}$$

D is the sum of squares within pairs. The maximum value D can have is therefore the total sum of squares (of all 2N observations from their common mean), and this value occurs, of course, when the sum of squares between pairs is zero, i.e., when the pair means are all equal and all the variation comes from within-pair differences or disagreement.

[7] Simple proofs for the relations discussed are available from the writer upon request.

[8] R. A. Fisher, *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd, 7th ed., 1938, Section 38.

For the special case of two variables ($k = 2$), the relation between the coefficient of agreement of (3) and (6) and the coefficient of intraclass correlation is

$$r_i = 2A - 1, \qquad (7)$$

where r_i denotes the intraclass correlation. In the case of two variables, the intraclass correlation is thus a simple linear function of the coefficient of agreement. The only difference between the two is that the range of values for the intraclass correlation is from -1 to $+1$, while the range for the agreement coefficient is from 0 to 1. For three or more variables the relation is also linear, but the range of the intraclass correlation then becomes a function of the number of variables (k), and r_i is not so useful a measure as A.[9]

The Relation Between Pearsonian and Intraclass Correlation. In the case of two variables, the value of the intraclass correlation depends in part upon the corresponding Pearsonian correlation, but it also depends upon the differences between the means and the standard deviations of the two variables:

$$r_i = \frac{[(s_1^2+s_2^2)-(s_1-s_2)^2]r-(\overline{X}_1-\overline{X}_2)^2/2}{(s_1^2+s_2^2)+(\overline{X}_1-\overline{X}_2)^2/2}, \qquad (8)$$

where $\overline{X}_1$ and $\overline{X}_2$ denote the means of X_1 and X_2, s_1 and s_2 the standard deviations, and r the Pearsonian correlation between X_1 and X_2.

When the means of the two variables are equal, and the standard deviations are equal as well, then the Pearsonian and intraclass correlations are also equal.[10] When the means and/or the standard deviations of the two variables differ, the intraclass correlation is less than the Pearsonian.[11]

The intraclass correlation, which measures agreement, thus penalizes the Pearsonian for a difference in origin, or level, between the two variables. If one variable has consistently higher or lower values than the other, this fact diminishes the agreement between the variables.

The standard deviation is a measure of the scale or unit of the distributed variable. The intraclass correlation, which measures agreement, thus also penalizes the Pearsonian for a difference in unit between the two variables. If a given difference between scores is differently assessed by the two variables, then the agreement between the variables is diminished.

For the example of Table 1, the Pearsonian correlation is .649, while the intraclass correlation is only .429.[12] The square of the difference between the standard deviations of janitor's and banker's ratings, $(s_1 - s_2)^2$, which functions to diminish the numerator of the right member of (8), is only .007, in comparison with the sum of the variances of janitor's and banker's ratings, $(s_1^2 + s_2^2) = 2.210$, from which it is subtracted. The difference between the standard deviations of the ratings by janitor and banker thus has little effect on the difference between agreement and correlation.

The major reason for the difference between the intraclass and the Pearsonian correlations for Table 1 rests in the difference between the means (or levels) of janitor's and banker's ratings. A glance at Table 1 will show that the janitor consistently rated families higher in socioeconomic status than did the banker. There are only eight families that the janitor rated lower than the banker (the sum of the frequencies above the bold-faced diagonal of agreement), while there are 127 families the janitor rated higher than the banker (the sum of the frequencies below the bold-faced diagonal of agreement). The janitor's ratings are thus biased with respect to the banker's, or *vice versa*, and it is this bias that is of primary importance in reducing the Pearsonian correlation of .649 to the intraclass correlation of .429. The bias is measured by the difference between the means of janitor's and banker's ratings, which in this instance is .821, and which functions in both the numerator and the denominator of the right member of (8) in reducing the value of the Pearsonian correlation.

[9] See Figure 1.

[10] If one sets $(\overline{X}_1 - \overline{X}_2)$ and $(s_1 - s_2)$ both equal to zero in (8), then $r_i = r$.

[11] If $(s_1 - s_2) \neq 0$ in (8), the numerator is diminished. If $(\overline{X}_1 - \overline{X}_2) \neq 0$ in (8), the numerator is diminished and the denominator increased.

[12] The means of janitor's and banker's ratings are 3.995 and 3.173 respectively, and the standard deviations are 1.0076 and 1.0929.

Agreement in the Case of More than Two Variables. Extension of the idea of agreement to more than two variables involves no difficulty. For example, had Lundberg in his socioeconomic status experiment chosen a banker, a janitor, *and* a clerk as raters, he would have been dealing with *families* of three observations each, and he would have been concerned to determine how closely the three values in each family agreed.

The argument in the k-variable situation is the same as in the bivariate. The problem is to measure how closely on the average the observations agree within families. As in the case of two variables, one can measure the disagreement of the k observations in a family by their sum of squares (of deviations from the family mean).[13] If the k values for a family are identical, their sum of squares or disagreement will be zero. If the values differ somewhat, their sum of squares or disagreement will be greater than zero; and the greater the degree to which the values of the family fail to agree, the greater will be their sum of squares.

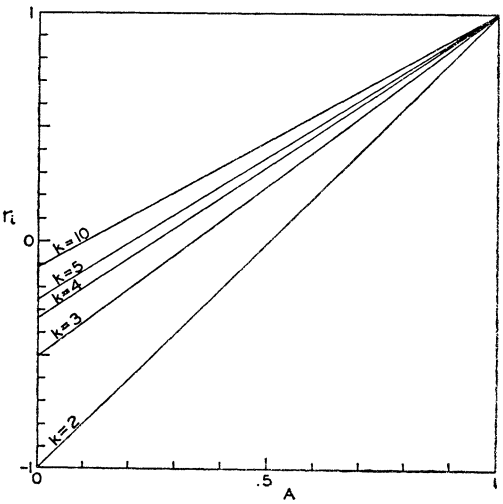
As in the case of two variables, the sum of the N sums of squares, one for each family, is a sum of squares measuring the total within-family disagreement. To find a relative rather than an absolute measure of disagreement, one must relate the total disagreement sum of squares to its maximum possible value, and this is, analogously to the bivariate case, the sum of the squares of the deviations of all kN observations from their common mean. Then the formula for the coefficient of agreement given in equation (3) holds for the k-variable case as well.

It was pointed out for the two-variable case that the intraclass correlation coefficient (r_i) and the coefficient of agreement (A) were very simply related, the only difference between them being that one had a range from -1 to $+1$ and the other a range from 0 to 1. No advantage was claimed for the coefficient of agreement over the coefficient of intraclass correlation. In the k-variable case, however, the relation is not so simple, and the coefficient of agreement does have an advantage. The relation is

$$r_i = \frac{kA - 1}{k - 1}, \qquad (9)$$

and is shown in Figure 1.

FIGURE 1. The Relation Between the Coefficient of Agreement (A) and the Coefficient of Intraclass Correlation (r_i) for Selected Values of k



It is apparent from Figure 1 that the value of the intraclass correlation depends not only upon A but also upon k, the number of observations per family. The range of A is always from zero to unity regardless of the number of observations per family, and therefore comparisons between agreement coefficients based upon different numbers of variables are commensurable. The upper limit of the intraclass correlation is always unity, but its lower limit is $-1/(k-1)$.[14] For two variables the lower limit of r_i is -1, but for three variables it is $-\frac{1}{2}$, for four variables $-\frac{1}{3}$, and for five variables $-\frac{1}{4}$.

For two or three variables, the intraclass correlation is most easily computed by computing the Pearsonian correlation for the intraclass correlation table. The intraclass scatter diagram (which defines the intraclass correlation table) has already been described

[13] For any set of k values, the sum of squares of deviations from the mean is proportional to the sum of squares of all possible differences between the k values. It is the sum of the squares of these *differences* which is pertinent here, and the sum of squares of deviations from the mean is used merely for convenience. See Maurice G. Kendall, *The Advanced Theory of Statistics*, London: Griffin and Co., 1947, Section 2.20.

[14] Fisher, *loc. cit.*

for the two-variable case. When three variables are involved, the construction of the table is more laborious. The three values for each family are entered in the table as six observations, each one being one of the six permutations of two values which can be made from the original three. That is, the values of the three variables X_1, X_2, and X_3 for each family are entered in a bivariate correlation table with coordinates (X_1, X_2), (X_1, X_3), (X_2, X_3), (X_2, X_1), (X_3, X_1), and (X_3, X_2), and the Pearsonian correlation is computed for the resulting table. If the value of A is desired, it can be computed from the relation

$$A = \frac{k-1}{k} r_i + \frac{1}{k}. \qquad (10)$$

For more than three variables, it is easier computationally to go directly to the analysis of variance; the details are well discussed by Fisher and by Yule and Kendall.[15]

Tests of Significance. Tests of significance for the intraclass correlation are exhaustively discussed by Fisher.[16] One may easily test the significance of a given value of the intraclass correlation, set up confidence limits for it, or test the significance of the difference between values of the intraclass correlation derived from independent samples. If one wants to deal in terms of the coefficient of agreement (A) instead of the intraclass correlation coefficient (r_i), the easier way is to make the tests in terms of r_i and to translate the results into terms of A by means of equation (10). Any probability statement about r_i involves a corresponding statement about A because of the linear relation between r_i and A shown in Figure 1. However, stating the problem in terms of intraclass correlation permits one when $k = 2$ to use the easily available table in Fisher [17] rather than a detailed table of the exponential function.

Both the intraclass correlation coefficient and the coefficient of agreement involve the assumption that the variables in the problem are metricized. Furthermore, probability statements involving either coefficient are derived on the assumption that the variables are normally distributed in the parent population. In a strict sense, therefore, the coefficient should not be applied to variables for which the unit is not known to be constant or for which non-normality can be demonstrated. For example, application of the coefficients in assessing the agreement among judges' ratings (as in Table 1) may be suspect because of possible inequalities amoung the units over different portions of the scales.

However, these two assumptions underlying the application of measures of agreement are identical with the assumptions underlying the application of the Pearsonian correlation coefficient. There thus seems no reason why measures of agreement should not be treated with the same degree of relaxation as the Pearsonian correlation is, providing the person so treating them accepts the risks involved. Ratings, for example, are often correlated Pearsonian style for reasons of convenience, or because inequality of units is thought to be of minor importance, or because no alternative technique is available. Sets of variables that are obviously not samples from bivariate or multivariate normal populations are also correlated for similar reasons.

One might wish for a measure of agreement free of assumptions as to metric or the distribution of the parent population, but this idea seems inconsistent with the idea of agreement. For example, Kendall has proposed a "coefficient of concordance" W,[18] based upon rankings, which is similar in certain respects to the coefficient of agreement of this paper. Kendall's W measures the agreement between m (in our notation k) observers who have ranked the same objects or individuals. Of course numerical ratings or scores can be reduced to ranks so as to make Kendall's coefficient more widely applicable. The range of possible values for W is from 0 to 1, as is the range

[15] Fisher, *op. cit.*, Sections 38 and 40; and G. Udny Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, London: Griffin and Co., 1937, Sections 13.27–13.33.

[16] Fisher, *op. cit.*, Sections 39–41.

[17] Fisher, *op. cit.*, Table V.B. Fisher's tests of significance involve transforming r_i into z. The relation between r_i and z for $k = 2$, i.e., for r_i computed from *pairs* of observations, is given by Table V.B. For $k > 2$, z has to be computed. When Fisher writes "log" he means "natural log."

[18] Maurice G. Kendall, *Rank Correlation Methods*, London: Griffin and Co., 1948, Chapters 6 and 7.

of the coefficient of agreement proposed here for the metricized case.[19] When W = O, the rankings disagree maximally. As the rankings come closer to agreement, W increases in value, and when the rankings are identical, W = 1. Since only rankings are involved in the computation of W, the coefficient does not involve the assumption of metric or of a specified parent population distribution.

However, while Kendall's W is quite legitimately used on rankings, it does not measure agreement as defined by the intraclass correlation or the coefficient of agreement.[20] When objects or individuals are ranked instead of being assigned scores, or when assigned scores themselves are ranked, the resulting rankings are distribution-free in the sense that the initial variables have been forced into the same distribution, *viz.*, a rectangular one composed of the numbers 1, 2, 3, . . . The rankings of three judges, for example, of necessity have the same distribution and therefore the same means and variances. Thus Kendall's coefficient of concordance W measures agreement only in the sense of agreement among the *serial orders* assigned to the observations, not in a sense in which different levels (means) or scales (standard deviations) of the ratings affect the coefficient. Thus W would ignore a judge's tendency to rate individuals consistently higher or lower than other judges. W, in other words, is a measure of what might be called "serial agreement," and is more closely akin to correlation than to the idea of agreement proposed in this paper. In fact, W is a linear function of the average Spearman rank correlation between all possible pairs of rankings, in an equation in which means and variances of course do not occur at all.

Other Applications. Once the distinction between correlation and agreement is acknowledged, many problems in which correlation has hitherto been used but in which agreement is the appropriate concept will be recognized. Three of these problems are briefly discussed below.

1. Much can be said in favor of the idea that the determination of test reliability involves the concept of agreement rather than correlation. It is true that usually only two variables are involved in a reliability computation, whether the variables represent split-halves, parallel forms, or test and retest; but there is no reason why three or more parallel forms, or test and two or more retests, should not be used. Moreover, the idea of regression (absent in agreement) is not involved in reliability determination at all. There is no point to estimating scores on one form of a test from scores on a second or third form, or of estimating scores on a retest from scores on an initial test. What is really wanted is a measure of the degree to which scores derived from one trial are identical with scores from another trial or other trials.

The limited scope of this paper, which is primarily expository, and the present confused state of reliability theory,[21] do not permit the extended discussion the topic really warrants. However, papers by Hoyt [22] and Guttman [23] have emphasized not only the multivariate nature of reliability theory but also theoretical reasons for defining reliability in terms of the ratio of the within-pair or within-family variation to the total variation of the test, in a way similar to that used in defining the coefficient of agreement

[19] Kendall points out that for more than two variables agreement and disagreement are not exact opposites. Two variables can disagree completely, but three cannot disagree completely in the same sense. For example, if a variable X disagrees with *both* Y and Z, then Y and Z must agree to some extent. In other words, the number of variables sets a lower limit to the amount of disagreement possible. This is illustrated by the change in the lower bound of the intraclass correlation coefficient as the number of variables changes, as shown in Figure 1. Both Kendall's W and the coefficient of agreement proposed here relate the actual amount of disagreement to the maximum possible amount for the number of variables under consideration. This results in a uniform scale of agreement from 0 to 1 regardless of the number of variables. See Kendall, *op. cit.*, p. 81.

[20] See Equation (8).

[21] See, for example, Louis Guttman, "A Special Review of Harold Gulliksen, *Theory of Mental Tests*," Psychometrika, 18 (June, 1953), pp. 123–130; and Harold Gulliksen, "Comments on Guttman's Review of *Theory of Mental Tests*," *op. cit.*, pp. 131–133.

[22] Cyril Hoyt, "Test Reliability Estimated by Analysis of Variance," *Psychometrika*, 6 (June, 1941), pp. 153–160.

[23] Louis Guttman, "A Basis for Analyzing Test-Retest Reliability," *Psychometrika*, 10 (December, 1945), pp. 255–282.

in Equations (3) and (6). In fact, the primitive evidence of unreliability, by which reliability can in turn be defined as in Equation (3), is the discrepancy between the test score and the retest score, or between the score on one form and the score on another form.

The foregoing statement will have to stand here merely as a suggestion, but it does agree with a commonsense approach to reliability. In reliability theory each trial of a test is regarded as an independent trial from one and the same universe of trials, i.e., the test is assumed to be measuring the same thing on different trials. It follows that the means, or the variances, of two trials should differ only randomly, and that for large samples the means, and the variances, should be effectively identical. If these conditions are fulfilled, then the Pearsonian correlation between the two trials will be effectively equal to the intraclass correlation by virtue of Equation (8). If these conditions are not fulfilled, then the intraclass correlation will quite properly reduce the value of the Pearsonian.

2. A measure of agreement is required in some problems in which observed and theoretically deduced values of a variable are compared. It is often required, for example, as an index of closeness of fit of a model to data. In some problems of this kind the model variable is constrained to have the same mean and variance as the observed data, and here of course the Pearsonian and intraclass correlations are identical. In situations in which model and data are not forced to this correspondence, the idea of agreement is usually called for rather than correlation. The use of intraclass correlation in such instances does away with the need for statements such as, "The theoretically deduced values correlate so-and-so with the experimentally observed values, but some bias is evident as the theoretical values tend to underestimate the observed ones in the lower portion of the range and to overestimate the observed ones in the upper portion of the range." In statements of this kind, which appear repeatedly in the literature, it is obvious that the concept of agreement is involved. In these situations, only the coefficient of agreement or the coefficient of intraclass correlation properly assesses the situation by taking into account not only the correlation between the two sets of values but also discrepancies in level and scale. Analysis of the numerical effects of the components $(\overline{X}_1 - \overline{X}_2)$, $(s_1 - s_2)$, and r in Equation (8) then provides an explanation for the inadequacy of the model.

3. The idea of agreement is particularly useful to sociologists because many sociological problems call for the measurement of within-group "likeness." Intraclass correlation would be appropriate, for example, in assessing the likeness (or agreement) of husbands' and wives' attitude scores, or in comparing their agreement at two or more times during a political campaign. Again, analysis of the effects of the components $(\overline{X}_1 - \overline{X}_2)$, $(s_1 - s_2)$, and r in Equation (8) would provide an explanation for the discrepancies between observed scores. The easy extension of the idea of agreement to groups of three or more observations makes the idea applicable in measuring relative degrees of consistency among group members regardless of the size of the group. The coefficient of agreement should thus be useful, for example, in small-group studies in which a measure of consistency of member response is wanted.

Summary. This paper has pointed out and illustrated the distinction between Pearsonian correlation and intraclass correlation. It has proposed that what is inconveniently measured by the coefficient of intraclass correlation be termed *agreement* to reflect its true meaning, and has proposed a logically defensible coefficient for assessing agreement.

There is a difference between research situations in which the concern is to measure the agreement between two or more variables and situations in which the purpose is to estimate one variable from another or from others. This paper proposes that the distinction is an essential part of the conceptual apparatus of sociological research.