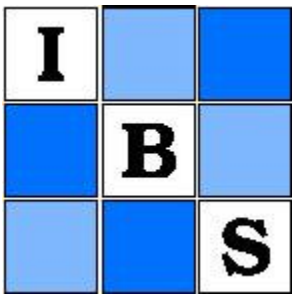

BIOMETRICS *31*, 651–659
September 1975

MEASURING AGREEMENT BETWEEN TWO JUDGES ON THE PRESENCE OR ABSENCE OF A TRAIT

JOSEPH L. FLEISS

Biometrics Research Unit, New York State Psychiatric Institute, 722 W. 168th St., New York, N. Y. 10032, U. S. A., and Division of Biostatistics, Columbia University

SUMMARY

At least a dozen indexes have been proposed for measuring agreement between two judges on a categorical scale. Using the binary (positive-negative) case as a model, this paper presents and critically evaluates some of these proposed measures. The importance of correcting for chance-expected agreement is emphasized, and identities with intraclass correlation coefficients are pointed out.

1. INTRODUCTION

Many research and clinical endeavors rely on an observer's judgment as to whether a trait, attribute or symptom is present or absent. No matter what the basis is for the judgment, some degree of error is inevitable (Fleiss [1973]). In the absence of a laboratory or other test that might provide a standard against which to assess the correctness of the judgment, one must rely on the degree of agreement between different judges for information about error.

Attention is restricted in this paper to reliability studies in which two observers independently judge each of a sample of n subjects on a binary (positive-negative) trait. The sample size is assumed to be large, so that the emphasis may be placed on the measurement of agreement rather than on statistical inference. A number of proposed indexes of agreement are presented and criticized, and the importance is stressed of building into the measure of agreement the value of the index expected by chance alone. Some chance-corrected measures of agreement are shown to be interpretable as intraclass correlation coefficients.

2. SOME INDEXES OF AGREEMENT

One may crossclassify the judgments by two observers on the presence or absence of a trait as in Table 1. Each entry in the table is a proportion.

The most obvious index of agreement is the proportion of all subjects on whom the two observers agree, $a + d$. The proportion of agreement, called the "index of crude agreement" by Rogot and Goldberg [1966], appears as a special case of a number of indexes of agreement. Thus, two of the three indexes proposed by Armitage *et al.*, [1966] become, when there are only two observers, identically equal to $a + d$. One is the "mean majority agreement index," which measures the frequency with which observers agree with the majority judgment. The second is the "mean pair agreement index," which is the proportion of agreeing pairs of judgments out of all possible pairs of judgments.

TABLE 1

RELATIVE FREQUENCIES OF JUDGMENTS BY TWO OBSERVERS ON A SAMPLE OF n SUBJECTS

	Observer 2		
Observer 1	Positive	Negative	Total
Positive	a	b	p_1
Negative	c	d	q_1
Total	p_2	q_2	1

Goodman and Kruskal ([1954] p. 758) contend that the only reasonable indexes of agreement on a present-absent trait are those based on $a + d$. Nevertheless, other reasonable indexes have been proposed which are not functions of $a + d$ because they do not treat presence and absence symmetrically.

Suppose that negative judgments are more frequent than positive judgments, i.e., that

$$\bar{q} > \bar{p}, \quad \text{where} \quad \bar{q} = \tfrac{1}{2}(q_1 + q_2)$$

and

$$\bar{p} = \tfrac{1}{2}(p_1 + p_2) = 1 - \bar{q}.$$

One may then wish to measure agreement only on those subjects judged positive by at least one of the observers, and thus to ignore the proportion d. Indexes which ignore d are especially popular in numerical taxonomy (Sneath and Sokal [1973] pp. 131–132). One due to Dice [1945] seems the most defensible as an index of agreement. It is

$$S_D = \frac{a}{\bar{p}}. \tag{1}$$

S_D may be interpreted as a conditional probability. Let one of the two observers be selected at random, and let attention be focused on the subjects judged positive by him. S_D is the conditional probability that the second observer judges positive, given that the randomly selected first observer judged positive.

If instead of ignoring d one chooses to ignore a, one would calculate the corresponding index

$$S_D' = \frac{d}{\bar{q}}. \tag{2}$$

Rogot and Goldberg [1966] proposed simply taking the mean of S_D and S_D' as an index of agreement. They denote the resulting index A_2 :

$$A_2 = \frac{a}{p_1 + p_2} + \frac{d}{q_1 + q_2}. \tag{3}$$

$A_2 = 0$ when there is complete disagreement and $A_2 = +1$ when there is complete agreement.

Rogot and Goldberg proposed another index based on the four conditional probabilities a/p_1 , a/p_2 , d/q_1 and d/q_2 . The first two are conditional probabilities that a specified observer judges positive given that the other has judged positive, and the last two are conditional probabilities that a specified observer judges negative given that the other has judged negative. The proposed index is simply the mean of the four conditional probabilities,

$$A_1 = \frac{1}{4}\left(\frac{a}{p_1} + \frac{a}{p_2} + \frac{d}{q_1} + \frac{d}{q_2}\right). \tag{4}$$

$A_1 = 0$ when there is complete disagreement and $A_1 = +1$ when there is complete agreement. A_1 has an important additional property. When the degree of agreement is exactly that predicted by chance (i.e. when $a = p_1p_2$ and $d = q_1q_2$), then $A_1 = \frac{1}{2}$.

Armitage *et al.* [1966] proposed as another index of agreement the standard deviation of the subjects' total scores, where a subject scores 2 if both observers judged him positive, scores 1 if one observer judged him positive and the other negative, and scores 0 if both observers judged him negative. The "standard deviation agreement index" is given by the square root of

$$SDAI^2 = \frac{n}{n-1}(a + d - (a - d)^2). \tag{5}$$

The $SDAI$ assumes the value 0 when there is complete disagreement, but assumes its maximum value of $\sqrt{n/(n-1)}$ only when there is complete agreement *and* when $p_1 = p_2 = \frac{1}{2}$. This latter feature would appear to render the $SDAI$, as it stands, inadequate as an index of agreement, but the following rescaling results in an index which does vary from 0 for complete disagreement to 1 for complete agreement. Complete agreement exists only when $a = p_1 = p_2$ (and thus equal $\bar{p}$) and $d = q_1 = q_2 = \bar{q}$. When these equalities hold, the value of $SDAI^2$ becomes $n(1 - (\bar{p} - \bar{q})^2)/(n - 1)$. The rescaled index, say

$$RSD^2 = \frac{a + d - (a - d)^2}{1 - (\bar{p} - \bar{q})^2}, \tag{6}$$

has the desired range of variation.

Suppose that negative judgments are more frequent than positive judgments. Goodman and Kruskal [1954] proposed

$$\lambda_r = \frac{(a + d) - \bar{q}}{\bar{p}} = \frac{2a - (b + c)}{2a + (b + c)} \tag{7}$$

as an index of agreement. λ_r is motivated less by notions of agreement than by a consideration of the frequencies of correct predictions of a subject's status when predictions are made with and without knowledge of the joint judgments. λ_r assumes its maximum value of $+1$ when there is complete agreement, but assumes its minimum value of -1 whenever $a = 0$, irrespective of the value of d (not, as Goodman and Kruskal ([1954] p. 758) imply, only when $a + d = 0$). A noteworthy identity is $\lambda_r = 2S_D - 1$, where S_D is defined in (1). Thus, λ_r is appropriate as an index of agreement under the same conditions that S_D is, namely whenever the investigator deems it appropriate to restrict his attention to subjects judged positive by at least one of the observers.

3. CORRECTING FOR AGREEMENT EXPECTED BY CHANCE

Except in the most extreme circumstances (either $p_1 = q_2 = 0$ or $p_2 = q_1 = 0$), some degree of agreement is to be expected by chance alone (see Table 2). For example, if observer 1 employs one set of criteria for distinguishing between the presence and absence of the trait, and if observer 2 employs an entirely different and independent set of criteria, then *all* of the observed agreement is explainable by chance. If there is some overlap between their sets of criteria, then the observed agreement will tend to exceed the chance-expected agreement; the greater the overlap, the greater the excess. In the other direction, if some of observer 1's criteria are used in the opposite way by observer 2, the observed agreement will tend to be less than the chance-expected agreement.

TABLE 2

CHANCE-EXPECTED PROPORTIONS OF JOINT JUDGMENTS BY TWO OBSERVERS

	Observer 2		
Observer 1	Positive	Negative	Total
Positive	p_1p_2	p_1q_2	p_1
Negative	q_1p_2	q_1q_2	q_1
Total	p_2	q_2	1

Different opinions have been stated on the need to incorporate chance agreement into the assessment of interobserver reliability. Rogot and Goldberg [1966], for example, emphasize the importance of contrasting observed with expected agreement when comparisons are to be made between different pairs of observers or different kinds of subjects. In fact, they cite the result that the chance-expected value of A_1 is always $\frac{1}{2}$ whereas that of A_2 depends on the marginal proportions as sufficient reason for A_1 to be preferred to A_2.

Goodman and Kruskal ([1954] p. 758), on the other hand, contend that chance agreement need not cause much concern, that the observed degree of agreement may usually be assumed to be in excess of chance. Even granting this assumption, one must nevertheless check whether the excess is trivially small or substantively large.

Armitage *et al.* ([1966] p. 102) occupy a position between that of Rogot and Goldberg and that of Goodman and Kruskal. They appreciate the necessity for introducing chance agreement whenever different sets of data are compared, but claim that too much uncertainty exists as to how the correction for chance is to be incorporated into the measure of agreement. They conclude that no corrections for chance be made, and suggest that comparisons be restricted to sets of data with equal overall proportions of positive judgments.

There does exist, however, a natural means for correcting for chance. Consider any index which assumes the value 1 when there is complete agreement. Let I_o denote the

observed value of the index (calculated from the proportions in Table 1) and let I_e denote the value expected on the basis of chance alone (calculated from the proportions in Table 2).

The obtained excess beyond chance is $I_o - I_e$, whereas the maximum possible excess is $1 - I_e$. The ratio of these two differences,

$$M(I) = \frac{I_o - I_e}{1 - I_e}, \tag{8}$$

is a measure of agreement with desirable properties. If there is complete agreement, $M = +1$. If observed agreement is greater than or equal to chance agreement, $M \geqq 0$, and if observed agreement is less than or equal to chance agreement, $M \leqq 0$. The minimum value of M usually depends on the marginal proportions. If they are such that $I_e = \frac{1}{2}$, then the minimum equals -1.

Scott [1955] and Cohen [1960] seem to have been the first to propose measures like M. Both took I_o to be the crude index of agreement, $a + d$, but defined I_e differently. Scott [1955] assumed that the two judges' underlying base rates were the same, and took $I_e = \bar{p}^2 + \bar{q}^2$. His measure, denoted π, becomes

$$\pi = \frac{4(ad - bc) - (b - c)^2}{(p_1 + p_2)(q_1 + q_2)}. \tag{9}$$

Fleiss [1965] derived the same measure (denoted r^* by him) as an intraclass correlation coefficient (see equation 18 below), but suggested that its use be restricted to the case where the marginal distributions were similar ([1965] p. 473).

Cohen [1960] made no assumption concerning equality of the marginal distributions, and took $I_e = p_1p_2 + q_1q_2$. His index, denoted κ, becomes

$$\kappa = M(a + d) = \frac{2(ad - bc)}{p_1q_2 + p_2q_1}. \tag{10}$$

Because Cohen's approach does not make what might be an unwarranted assumption about the marginal proportions, it appears to be preferable to Scott's.

Application of equation (8) to the indexes introduced in section 2 succeeds in unifying most of them. The first index introduced which was not a function of $a + d$ was S_D (1). Its chance-expected value is estimated as $E(S_D) = p_1p_2/\bar{p}$. $M(S_D)$ is then

$$M(S_D) = \frac{S_D - E(S_D)}{1 - E(S_D)} = \frac{2(ad - bc)}{p_1q_2 + p_2q_1}, \tag{11}$$

identically equal to κ. Because of the identity $\lambda_r = 2S_D - 1$, it is obvious that, in addition, $M(\lambda_r) = \kappa$.

The second index introduced was A_2 (3). As Rogot and Goldberg [1966] pointed out, the chance-expected value of A_2 may be estimated as

$$E(A_2) = \frac{p_1p_2}{p_1 + p_2} + \frac{q_1q_2}{q_1 + q_2}. \tag{12}$$

$M(A_2)$ becomes

$$M(A_2) = \frac{A_2 - E(A_2)}{1 - E(A_2)} = \frac{2(ad - bc)}{p_1q_2 + p_2q_1}, \tag{13}$$

again identical to κ.

Armitage, Blendis and Smyllie's [1966] standard deviation agreement index (5) does not necessarily assume the value 1 in the case of complete agreement, but the rescaled index, RSD^2 (6), does. Because the numerator of RSD^2 involves squares and products of proportions, its expected value will contain terms of order $1/n$. These terms may be ignored if n is large, and the chance-expected value of RSD^2 is then estimated as

$$E(RSD^2) = \frac{p_1q_1 + p_2q_2}{1 - (\bar{p} - \bar{q})^2}. \tag{14}$$

It is easily checked that $M(RSD^2)$ is also equal to κ.

Thus, five of the indexes introduced earlier yield κ as the chance-corrected measure of agreement. A sixth, A_1 (4), does not. The chance-expected value of A_1 was shown by Rogot and Goldberg [1966] to equal $\frac{1}{2}$. The measure $M(A_1)$ becomes

$$M(A_1) = \frac{A_1 - 1/2}{1/2} = \frac{(ad - bc)(p_1q_1 + p_2q_2)}{2p_1q_1p_2q_2}$$

$$= \frac{ad - bc}{HM}, \tag{15}$$

where HM is the *harmonic mean* of the two marginal variances.

$M(A_1)$ has the important property that it assumes its minimum value of -1 whenever there is complete disagreement; κ does so only when, in addition, $p_1 = p_2 = \frac{1}{2}$. Other measures besides $M(A_1)$ also have the property that they equal $+1$ if and only if there is complete agreement, 0 if and only if observed agreement equals chance agreement, and -1 if and only if there is complete disagreement. One such measure is the phi coefficient,

$$\phi = \frac{ad - bc}{\sqrt{p_1q_1p_2q_2}} = \frac{ad - bc}{GM}. \tag{16}$$

where GM is the *geometric mean* of the two variances. Another is a measure proposed by Maxwell and Pilliner [1968]. Their measure, which they derived from the psychometric theory of internal consistency, is

$$r_{11} = \frac{2(ad - bc)}{p_1q_1 + p_2q_2} = \frac{ad - bc}{AM}, \tag{17}$$

where AM is the *arithmetic mean* of the two variances.

The three measures given by (15) to (17) are distinguished only by which average variance is taken in the denominator, so that none stands out as superior to the others as a chance-corrected measure of agreement. It will be shown in the next section, however, that only the last of them, r_{11} (17), is also interpretable as an intraclass correlation coefficient.

4. INTRACLASS CORRELATION COEFFICIENTS OF AGREEMENT

When each of a sample of subjects is rated on a quantitative scale by two or more observers, agreement is usually measured by an intraclass correlation coefficient (Bartko [1966], Ebel [1951]). The application of the algebra of intraclass correlation to judgments on a present-absent trait reveals new insights into some of the measures introduced in the preceding section.

Table 3 presents the sums of squares in the analysis of variance table appropriate to quantitative data (where X_{ij} is the score given by observer i to subject j), and the values

TABLE 3

SUMS OF SQUARES IN ANALYSIS OF VARIANCE TABLE FOR RATINGS BY TWO OBSERVERS ON n SUBJECTS

Source	df	Sum of Squares: Quantitative Data	Sum of Squares: Present-Absent Data
Observers	1	$\frac{n}{2}(\bar{X}_{1\cdot} - \bar{X}_{2\cdot})^2$	$\frac{n}{2}(b - c)^2$ $(= O)$
Subjects	$n - 1$	$2\sum_j(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2$	$\frac{n}{2}(a + d - (a - d)^2)$ $(= S)$
Error	$n - 1$	$\frac{1}{2}\sum_j([X_{1j} - X_{2j}] - [\bar{X}_{1\cdot} - \bar{X}_{2\cdot}])^2$	$\frac{n}{2}(b + c - (b - c)^2)$ $(= E)$

of these sums of squares when X_{ij} may assume only the values 0 (when the judgment is "negative") or 1 (when the judgment is "positive").

If one ignores the fact that the same two observers made all the judgments, one would have the simplest intraclass model: variation between subjects (with sum of squares equal to S) and variation within subjects (with sum of squares equal to $O + E$). Suppose that n is large enough so that the fraction $n/(n - 1)$ is effectively equal to 1. The appropriate intraclass correlation coefficient would then be

$$R_1 = \frac{S - (O + E)}{S + (O + E)} = \frac{4(ad - bc) - (b - c)^2}{(p_1 + p_2)(q_1 + q_2)}, \tag{18}$$

identical to Scott's π and Fleiss' r^* (see equation 9). Because it seems inappropriate to fail to partition the within subject sum of squares into its components O and E, R_1 (and therefore π and r^*) seems inappropriate as a measure of agreement.

Suppose that the reliability study is carried out prior to a substantive study in which each observer will evaluate a different sample of subjects, and in which the judgments by the two observers will be analyzed separately. In such a case, one would not include observer differences in the estimate of overall variability (see Bartko [1966]), and would take as the intraclass correlation coefficient

$$R_2 = \frac{S - E}{S + E} = \frac{2(ad - bc)}{p_1q_1 + p_2q_2}, \tag{19}$$

equal to Maxwell and Pilliner's r_{11} (17).

Suppose, now, that the plan of the substantive study calls for the separate samples of subjects evaluated by the two observers to be combined into a single sample, or that the reliability study is carried out for the more general purpose of determining the degree of agreement attainable in making a certain kind of judgment. In either case, interest is in the extent to which the judgments by different observers are interchangeable.

Systematic differences between the observers then comprise a part of overall variability, and the appropriate intraclass correlation coefficient is

$$R_3 = \frac{S - E}{S + E + 2O} \tag{20}$$

(see Bartko [1966] for a derivation of this equation). Krippendorff [1970] and Fleiss and Cohen [1973] have shown that R_3 is identically equal to κ (10). The existence of different base rates (i.e. inequality of p_1 and p_2) reduces the absolute value of R_3 below that of R_2, so that the absolute value of κ is always less than or equal to that of r_{11}. Because different base rates imply that the judgments are not completely interchangeable, this inequality is reasonable.

5. DISCUSSION

This survey has been concerned only with measuring the agreement between a pair of observers on the presence or absence of a trait. Measures of chance-corrected agreement for the case of more than a pair of judges, and for the case of traits defined by more than two states, have been proposed by Cohen [1960], Cronholm [1963], Fleiss [1971], Light [1971], and Maxwell and Pilliner [1968]. The measures given by Fleiss [1971] correct for chance the "mean pair agreement index" of Armitage *et al.* [1966]. Measures appropriate to ordinal scales, or to the case where the relative seriousness of each kind of disagreement can be quantified, have been proposed by Cicchetti and Allison [1971] and by Cohen [1968]. Large sample expected values and standard errors for many of these measures have been derived by Bennett [1972], Everitt [1968], Fleiss [1971] and Fleiss *et al.* [1969].

It should be obvious that no index of agreement is informative by itself, that it should be expressed as a relative excess (or deficit) over the degree of agreement expected by chance alone. Of the many measures of agreement which have been proposed in the literature, only Cohen's κ (10) and Maxwell and Pilliner's r_{11} (17) are defensible both as chance-corrected measures and as intraclass correlation coefficients. Instead of having to choose one from a dozen or more available indexes of agreement with little to base the choice on, the biometrician may choose one of only a pair of measures on the basis of its interpretation as a reliability coefficient.

ACKNOWLEDGMENTS

This work was supported in part by grant MH 23964 from the National Institute of Mental Health. I wish to thank the anonymous referees for a number of helpful criticisms, especially for pointing out the identity involving λ_r and S_D.

MESURE DE L'ACCORD DE DEUX JUGES CONCERNANT LA PRÉSENCE OU L'ABSENCE D'UN TRAIT

RESUME

Au moins une douzaine d'indexes ont été proposés pour mesurer l'accord de deux juges concernant une échelle de catégories. A l'aide du cas binaire (positif-négatif) pris comme modèle, cet article présente et évalue d'une manière critique quelques unes de ces mesures. On insiste sur l'importance de la correction de l'accord du au hasard et on souligne des identités avec les coefficients de corrélation intraclasse.

REFERENCES

Armitage, P., Blendis, L. M. and Smyllie, H. C. [1966]. The measurement of observer disagreement in the recording of signs. *J. Royal Statist. Soc., Series A 129*, 98–109.

Bartko, J. J. [1966]. The intraclass correlation coefficient as a measure of reliability. *Psychol. Rep. 19*, 3–11.
Bennett, B. M. [1972]. Measures for clinicians' disagreements over signs. *Biometrics 28*, 607–12.
Cicchetti, D. V. and Allison, T. [1971]. A new procedure for assessing reliability of scoring EEG sleep recordings. *Amer. J. EEG Technology 11*, 101–9.
Cohen, J. [1960]. A coefficient of agreement for nominal scales. *Educ. Psychol. Measmt. 20*, 37–46.
Cohen, J. [1968]. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull. 70*, 213–20.
Cronholm, J. N. [1963]. A pair of nonparametric indices of agreement and disagreement. Report Number 592 from the Psychology Division of the U. S. Army Medical Research Laboratory, Fort Knox, Kentucky.
Dice, L. R. [1945]. Measures of the amount of ecologic association between species. *Ecology 26*, 297–302.
Ebel, R. L. [1951]. Estimation of the reliability of ratings. *Psychometrika 16*, 407–24.
Everitt, B. S. [1968]. Moments of the statistics kappa and weighted kappa. *Brit. J. Math. Statist. Psychol. 21*, 97–103.
Fleiss, J. L. [1965]. Estimating the accuracy of dichotomous judgments. *Psychometrika 30*, 469–79.
Fleiss, J. L. [1971]. Measuring nominal scale agreement among many raters. *Psychol. Bull. 76*, 378–82.
Fleiss, J. L. [1973]. *Statistical Methods for Rates and Proportions*. Wiley, New York.
Fleiss, J. L. and Cohen, J. [1973]. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Measmt. 33*, 613–9.
Fleiss, J. L., Cohen, J. and Everitt, B. S. [1969]. Large sample standard errors of kappa and weighted kappa. *Psychol. Bull. 72*, 323–7.
Goodman, L. A. and Kruskal, W. H. [1954]. Measures of association for cross classifications. *J. Amer. Statist. Assoc. 49*, 732–64.
Krippendorff, K. [1970]. Bivariate agreement coefficients for reliability of data, in E. F. Borgatta, ed., *Sociological Methodology, 1970*. Jossey-Bass, San Francisco.
Light, R. J. [1971]. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychol. Bull. 76*, 365–77.
Maxwell, A. E. and Pilliner, A. E. G. [1968]. Deriving coefficients of reliability and agreement for ratings. *Brit. J. Math. Statist. Psychol. 21*, 105–16.
Rogot, E. and Goldberg, I. D. [1966]. A proposed index for measuring agreement in test-retest studies. *J. Chron. Dis. 19*, 991–1006.
Scott, W. A. [1955]. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quart. 19*, 321–5.
Sneath, P. H. A. and Sokal, R. R. [1973]. *Numerical Taxonomy*. W. H. Freeman, San Francisco.

Received November 1973, *Revised July* 1974

Key Words: Dichotomous data; Inter-rater reliability; Coefficients of agreement.