Psychological Bulletin
1988, Vol. 103, No. 3, 374-378

0033-2909/88/$00.75

Another Look at Interrater Agreement

Rebecca Zwick
Educational Testing Service
Princeton, New Jersey

Most currently used measures of interrater agreement for the nominal case incorporate a correction for chance agreement. The definition of chance agreement, however, is not the same for all coefficients. Three chance-corrected coefficients are Cohen's (1960) κ; Scott's (1955) π; and the S index of Bennett, Alpert, and Goldstein (1954), which has reappeared in many guises. For all three measures, independence between raters is assumed in deriving the proportion of agreement expected by chance. Scott's π involves a further assumption of homogeneous rater marginals, and the S coefficient requires the assumption of uniform marginal distributions for both raters. Because of these disparate formulations, κ, π, and S can lead to different conclusions about rater agreement. Consideration of the properties of these measures leads to the recommendation that marginal homogeneity be assessed as a first step in the analysis of rater agreement. If marginal homogeneity can be assumed, π can be used as an index of agreement.

In educational and psychological research, it is frequently of interest to assign subjects to nominal categories, such as demographic groups, classroom-behavior types, or psychodiagnostic classifications. Because the reproducibility of the ratings is taken to be an indicator of the quality of the category definitions and the raters' ability to apply them, researchers typically require that the classification task be performed by two raters. For k categories, the results can be tabled in a $k \times k$ agreement matrix in which the main diagonal contains the cases for which the raters agree.

First Setting

Researchers in the fields of statistics, biostatistics, psychology, psychiatry, education, and sociology have proposed a multitude of interrater agreement measures. (See Landis & Koch, 1975a, 1975b, 1977, for useful reviews.) In this article, I focus on three coefficients that can be expressed in the form

$$A = \frac{P_O - P_C(A)}{1 - P_C(A)}, \quad (1)$$

where $P_O = \sum_{i=1}^{k} p_{ii}$ is the observed proportion of agreement, p_{ii} is the proportion of cases in the ith diagonal cell of the table, and $P_C(A)$ is the proportion of agreement expected by chance, as defined for coefficient A. These coefficients represent an attempt to correct P_O by subtracting from it the proportion of cases that fall on the diagonal by "chance." The numerator is then divided by $1 - P_C(A)$, the maximum nonchance agreement. (Note, however, that the numerator can reach this maximum only if the two raters have identical marginals. Otherwise, P_O cannot reach 1.00.) The resulting coefficient, A, is assumed to provide a better description of the degree of interrater agreement than the raw proportion of agreement, P_O.

One agreement index that can be expressed in the form of Equation 1 is Bennett, Alpert, and Goldstein's (1954) S coefficient, in which $P_C(S)$ is defined as $1/k$. This measure has reappeared as Janson and Vegelius's (1979) C coefficient; Brennan and Prediger's (1981) κ_n index; and, in the two-category case, Guilford's (1961; Holley & Guilford, 1964) G index; and Maxwell's (1977) random error (RE) coefficient. I point out the equivalence of these five coefficients, which has largely gone unrecognized in the literature, in the first part of this article.

In the main portion of the article, I compare the properties of S with those of two other coefficients that can be expressed in the form of Equation 1: Scott's (1955) π coefficient and Cohen's (1960) κ, currently the most popular index of rater agreement for nominal categories. For convenience, the definitions of $P_C(A)$ associated with each coefficient are listed in Table 1. Some identities between coefficients are given in Table 2. In the final section of the article, I make some recommendations for assessing interrater agreement in the nominal case. In particular, I stress the need for examining the marginal distributions of the raters. Although I focus in most of the article on a descriptive approach to the assessment of interrater agreement, I present an inferential procedure for assessing marginal homogeneity, along with a proposed marginal homogeneity index. Throughout the article, I substitute a uniform notation system for the notation used in the original presentations.

Bennett et al.'s (1954) S Coefficient

Bennett et al. (1954) sought to evaluate the degree of agreement between two methods of obtaining information about interviewees: a printed poll and a lengthy interview covering the general subject matter covered by the poll. They proposed the following agreement coefficient:

This research was supported in part by National Research Service Award No. 5-T32-MH 15745 from the National Institute of Mental Health to the University of North Carolina at Chapel Hill, where I was a postdoctoral fellow.

I thank Paul Holland and Skip Livingston for their comments.

Correspondence concerning this article should be addressed to Rebecca Zwick, Educational Testing Service, Princeton, New Jersey 08541.

Table 1
Definition of $P_C(A)$ for κ, π, and S

Coefficient	Definition
κ (Cohen, 1960)	$\sum_{i=1}^{k} p_{i+}p_{+i}$
π (Scott, 1955)	$\sum_{i=1}^{k} \left(\frac{p_{i+} + p_{+i}}{2}\right)^2$
S (Bennett, Alpert, & Goldstein, 1954)	$1/k$

$$S = \frac{k}{k-1}\left(P_O - \frac{1}{k}\right) \tag{2}$$

The rationale that they offered is as follows: "The proportion $1/k$ represents the best estimate of $[P_O]$ expected on the basis of chance. . . . The S score . . . ranges from zero to unity as $[P_O]$ ranges from the value most probably expected on the basis of chance to unity" (Bennett et al., 1954, p. 307).

RE, *G*, *C*, and κ_n Coefficients

Maxwell (1977) proposed an index of interrater agreement for 2×2 tables, called the *RE* coefficient, that has received some favorable attention in the literature (Carey & Gottesman, 1978; Janes, 1979). Maxwell's (1977) model for the assignment of subjects to categories can be outlined as follows: One assumes that if both raters are without doubt in categorizing a subject, the raters must agree; if one or both raters are in doubt about a case, they may either agree or disagree. Therefore, P_O is spuriously inflated because it includes some doubtful cases. If a_1 and a_2 denote the proportions of "true" agreements (i.e., excluding doubtful cases) for Categories 1 and 2, respectively, the proportion of doubtful cases is $1 - (a_1 + a_2)$. If it is assumed that these cases are allocated randomly to each of the four cells of the table, the cell frequencies will be as shown in Table 3. If one then wishes to obtain the quantity $a_1 + a_2$, the proportion of agreement uncontaminated by doubtful cases, one proceeds as follows:

$$\begin{aligned} a_1 + a_2 &= p_{11} + p_{22} - \tfrac{1}{2}[1 - (a_1 + a_2)] \\ &= (p_{11} + p_{22}) - (p_{12} + p_{21}) \\ &= P_O - P_D = RE, \end{aligned} \tag{3}$$

where p_{ij} is the proportion of cases in the ith row and the jth column and $P_D = p_{12} + p_{21}$ is the proportion of disagreement. As noted by Fleiss (1981), Maxwell's (1977) *RE* coefficient is algebraically equivalent to *G*, a measure of association for 2×2 tables proposed by Guilford (1961) and described by Holley and Guilford (1964). According to Holley and Guilford, McClung (1963) independently proposed a coefficient formulated as in Equation 3 for use in Q factor analysis. Holley and Guilford gave a rationale for the use of *G* that is much simpler than Maxwell's (1977) development of *RE*: P_O, the observed proportion of agreement, ranges from 0 to 1, and one would prefer a measure that ranges from -1 to 1. A linear transformation achieves this result:

$$\begin{aligned} G &= 2P_O - 1 \\ &= P_O + (1 - P_D) - 1 \\ &= P_O - P_D = RE. \end{aligned} \tag{4}$$

Green (1981) developed a post hoc rationale for the *G* coefficient that is very similar to Maxwell's (1977) development of *RE*.

First Setting

It is not difficult to generalize Maxwell's (1977) model to the case of $k > 2$. If one lets a_i $(i = 1, 2, \ldots k)$ represent the proportion of true agreement for the ith category and lets $RE_k = \sum_{i=1}^{k} a_i$ denote the generalized *RE* coefficient, then

$$\begin{aligned} P_O &= RE_k + \frac{1}{k}(1 - RE_k) \\ &= \frac{k-1}{k} RE_k + \frac{1}{k}. \end{aligned} \tag{5}$$

Now, solving for RE_k, one finds that

$$RE_k = \frac{k}{k-1}\left(P_O - \frac{1}{k}\right) = S. \tag{6}$$

Janson and Vegelius (1979) proposed a coefficient, *C*, which is identical to RE_k. Although Janson and Vegelius described *C* as a generalization of the *G* index, they did not note its equivalence to *S*. Brennan and Prediger (1981) presented yet another coefficient, κ_n, which, as they noted (p. 693), is equivalent to *S*.

Comparison of *S*, κ, and π

To simplify the discussion below, I refer to *RE*, *G*, *C*, and κ_n as *S*. As mentioned above, *S*, κ, and π can be expressed in a common form (Equation 1), with the difference among them lying in the definition of $P_C(A)$, the proportion of agreement expected to occur by chance. For each of the three coefficients, the formulation of $P_C(A)$ involves an assumption of independence of raters. That is, $P_C(A)$ is derived by multiplying, for each category, the hypothesized values of the raters' marginal proportions and then summing these products over the k categories. In its most general form, the proportion of chance agreement can be expressed as

$$P_C(A) = \sum_{i=1}^{k} h_{i+}h_{+i}, \tag{7}$$

Table 2
Identities Between Coefficients

Condition	Identity
$p_{i+} = p_{+i}$, $i = 1, 2 \ldots k$	$\pi = \kappa$
$k = 2$, $p_{i+} = p_{+i}$, $i = 1, 2$	$\pi = \kappa = \phi$ (phi correlation)
$p_{i+} = p_{+i} = 1/k$, $i = 1, 2 \ldots k$	$S = \pi = \kappa$
$k = 2$, $p_{i+} = p_{+i} = 1/k$, $i = 1, 2$	$S = \pi = \kappa = G = \phi$

Note. In addition, the following identities hold by definition: *RE* (Maxwell, 1977) = *G* (Holley & Guilford, 1964), and *C* (Janson & Vegelius, 1979) = κ_n (Brennan & Prediger, 1981) = *S*.

Table 3
Theoretical Cell Proportions for Maxwell's (1977) Model

	Rater 2		
Rater 1	Category 1	Category 2	Total
Category 1	$a_1 + \frac{1}{4}[1 - (a_1 + a_2)]$	$\frac{1}{4}[1 - (a_1 + a_2)]$	$a_1 + \frac{1}{2}[1 - (a_1 + a_2)]$
Category 2	$\frac{1}{4}[1 - (a_1 + a_2)]$	$a_2 + \frac{1}{4}[1 - (a_1 + a_2)]$	$a_2 + \frac{1}{2}[1 - (a_1 + a_2)]$
Total	$a_1 + \frac{1}{2}[1 - (a_1 + a_2)]$	$a_2 + \frac{1}{2}[1 - (a_1 + a_2)]$	1.00

Note. a_1 and a_2 represent the proportions of true agreements for Categories 1 and 2, respectively.

where h_{i+} is the hypothesized marginal proportion of cases assigned to category i by Rater 1 and h_{+i} is the corresponding proportion for Rater 2. As detailed below, these hypothetical marginal proportions are defined differently for the three coefficients.

In deriving $P_C(S)$, two assumptions are invoked: (a) homogeneity of rater marginals, $h_{i+} = h_{+i} \equiv h_i$, $i = 1, 2 \ldots k$, and (b) uniformity of rater marginals, $h_i = 1/k$, $i = 1, 2 \ldots k$.

Under these assumptions, each cell in the agreement matrix is expected to contain $1/k^2$ of the cases, and the total proportion of cases expected to fall in the k diagonal cells is $k(1/k^2) = 1/k$. Thus, $P_C(S) = 1/k$. Under some circumstances, however, the incorporation of the second assumption leads to underestimates of $P_C(A)$. Scott (1955) provided a clear example of this phenomenon:

> Given a two-category sex dimension and a P_O of 60 percent, the S . . . would be 0.20. But a whimsical researcher might add two more categories, "hermaphrodite" and "indeterminant," thereby increasing S to 0.47, though the two additional categories are not used at all. (p. 322)

In fact, minimization of the expression for $P_C(A)$ in Equation 7, subject to the first assumption and the constraint that $\sum_{i=1}^{k} h_i = 1.00$, shows that $\min[P_C(A)] = 1/k$. Therefore, under the assumption of homogeneous rater marginals, $1/k$ is a lower bound to the proportion of agreement due to chance. Underestimation of $P_C(A)$ can be shown algebraically to lead to inflated values of A.

Scott's (1955) π coefficient was designed to overcome the defects of S. It does not require the restrictive second assumption and does not become inflated by the inclusion of nonfunctional categories. Scott (1955) argued, "It is convenient to assume that the distribution for the entire set of interviews represents the most probable (and hence 'true' in the long-run probability sense) distribution for any individual coder" (p. 324). In computing π, then, one invokes only the first assumption, implying that $P_C(A) = \sum_{i=1}^{k} h_i^2$. Furthermore, one lets $h_i = \frac{p_{i+} + p_{+i}}{2}$, where p_{i+} and p_{+i} are the observed marginal proportions for Raters 1 and 2, respectively. Therefore, $P_C(\pi) = \sum_{i=1}^{k} \left(\frac{p_{i+} + p_{+i}}{2}\right)^2$.

Cohen (1960), who criticized the π index, remarked, "One source of disagreement between a pair of judges is precisely their proclivity to distribute their judgments differently over the categories" (p. 41). Fleiss (1975) raised a similar objection. Cohen recommended that κ, rather than π, be used to assess rater agreement. The chance correction for κ is based on the observed marginal distributions for each rater; that is, one lets $h_{i+} = p_{i+}$ and $h_{+i} = p_{+i}$. Therefore, $P_C(\kappa) = \sum_{i=1}^{k} p_{i+}p_{+i}$. When raters have the same observed marginals, $\pi = \kappa$ (and, for $k = 2$, $\pi = \kappa = \phi$, the phi correlation). When, in addition, these marginals are uniform, $S = \pi = \kappa$ (for any k).

For purposes of exploring the properties of κ, it is useful to examine, for fixed P_O, the effect of the rater marginals on the size of the coefficients. Table 4 contains three cases, all of which have $P_O = .60$. First consider the situation, represented in Cases 1 and 2, in which the two raters have identical marginals. In Case 1, $P_C(\kappa) = .25$ and $\kappa = .467$, whereas in Case 2, $P_C(\kappa) = .28$ and $\kappa = .444$. κ is larger in Case 1 because, if both raters

Table 4
Values of κ, S, and π for Three Cases

	Rater 2				
Rater 1	A	B	C	D	Total
	Case 1: Marginals uniform ($\kappa = .467, S = .467, \pi = .467$)				
A	.20	—	—	.05	.25
B	—	.10	.15	—	.25
C	—	.15	.10	—	.25
D	.05	—	—	.20	.25
Total	.25	.25	.25	.25	1.00
	Case 2: Marginals equal but not uniform ($\kappa = .444, S = .467, \pi = .444$)				
A	.20	.10	.10	—	.40
B	.10	.10	—	—	.20
C	.10	—	.10	—	.20
D	—	—	—	.20	.20
Total	.40	.20	.20	.20	1.00
	Case 3: Marginals unequal ($\kappa = .474, S = .467, \pi = .460$)				
A	.20	.05	.05	.10	.40
B	—	.10	.05	.05	.20
C	—	.05	.10	.05	.20
D	—	—	—	.20	.20
Total	.20	.20	.20	.40	1.00

have the same marginal distributions, $P_C(\kappa)$ is minimized (and thus κ maximized, for a fixed value of P_O) when the marginal distributions are uniform. (This property applies to π as well.) Whitehurst (1984) regarded this property of κ and the analogous property of the intraclass correlation in the ordinal case as statistical artifacts and therefore objectionable (see also Finn, 1970; Selvage, 1976). It is not clear, however, that the relation between the shape of the marginal distributions and the size of κ is undesirable: If cases are concentrated into a small number of categories, one cannot determine whether the rating system includes decision criteria that are adequate for discrimination among all k categories. Therefore, that the value of an agreement coefficient should be smaller in this situation than in the case of uniform marginals is not unreasonable.

But consider another factor that affects the size of κ: the degree to which raters agree in their marginal distributions. In both Cases 2 and 3 of Table 4, $P_O = .60$. In Case 2, in which the raters have identical marginals, $P_C(\kappa) = .28$ and $\kappa = .444$. In Case 3, however, in which the raters have different marginals, $P_C(\kappa) = .24$ and $\kappa = .474$. Thus, the raters in Case 2 are penalized for producing identical marginals. This phenomenon results from a property of κ pointed out by Brennan and Prediger (1981). In computing $P_C(\kappa)$, the marginal distributions associated with each rater are in a sense regarded as prior, despite the fact that they are in themselves evidence of the degree to which the raters agree. As Brennan and Prediger (1981) stated, "Two judges who independently, and with no a priori knowledge, produce similar marginal distributions must obtain a much higher agreement rate to obtain a given value of kappa, than two judges who produce radically different marginals" (p. 692). This appears to be an undesirable property. Because there are ordinarily no external restrictions on the marginals, there seems to be no justification for treating marginal discrepancies as an obstacle that raters should be credited for overcoming.

Recommendations

S, π, and κ all appear to have drawbacks. In formulating chance corrections, π invokes an assumption of homogeneous rater marginals, and S requires the further assumption of uniform marginals. The chance correction for κ is based on the observed marginals, but this has the effect of giving credit, for fixed P_O, to raters who produce different marginals. How, then, should interrater agreement be assessed? The answer lies, in part, in the examination of the degree of marginal homogeneity per se. Rather than ignoring marginal disagreement or attempting to correct for it, researchers should be studying it to determine whether it reflects important rater differences or merely random error.

I propose here that the assessment of rater agreement should consist of two phases: (a) the investigation of marginal homogeneity and, (b) if marginal homogeneity holds, the computation of Scott's (1955) π as a measure of chance-corrected agreement. The rationale for this approach is as follows. If one rejects the hypothesis of marginal homogeneity, one need go no further: One can express the degree of disagreement between raters in terms of the discrepancies between their marginal distributions. On the other hand, if marginal differences are small, it is reasonable to apply Scott's π, thus averaging out unimportant marginal differences in computing P_C. If marginal differences are small, the value of κ will, in any case, be close to that of π; the choice between them is therefore no longer important.

How can marginal homogeneity be assessed? With a fairly large random sample, Stuart's (1955) test can be used. The hypothesis of interest is H_0: $\pi_{i+} = \pi_{+i}$, where π_{i+} is the $k \times 1$ vector of elements π_{i+}, which represent the marginal probability of being in row i (corresponding to Rater 1), and π_{+i} is the corresponding vector of column probabilities (corresponding to Rater 2). The test statistic is

$$\chi^2_S = (\mathbf{p}_{i+} - \mathbf{p}_{+i})\mathbf{V}^{-1}(\mathbf{p}_{i+} - \mathbf{p}_{+i}), \tag{8}$$

where $(\mathbf{p}_{i+} - \mathbf{p}_{+i})$ is the $(k-1) \times 1$ vector of differences $(p_{i+} - p_{+i})$ between the ith-row marginal proportion and the ith-column marginal proportion for the first $k-1$ categories. (Actually, any $k-1$ of the k elements in the complete vector of marginal differences may be used in the computations, resulting in the same value of χ^2_S. The kth difference is determined.) $\mathbf{V}$ is the $(k-1) \times (k-1)$ variance–covariance matrix of the random vector $(\mathbf{p}_{i+} - \mathbf{p}_{+i})$, defined under H_0, with diagonal elements

$$v_{ii} = \frac{p_{i+} + p_{+i} - 2p_{ii}}{n} \tag{9}$$

and off-diagonal elements

$$v_{ij} = -\left(\frac{p_{ij} + p_{ji}}{n}\right), \tag{10}$$

where n is the sample size. The test statistic is asymptotically distributed as χ^2 with $k-1$ degrees of freedom under H_0. (When there are $k = 2$ categories, Stuart's test reduces to the McNemar, 1947, test.)

As an example, consider Case 3 of Table 4, assuming that $n = 100$. Then

$$(\mathbf{p}_{i+} - \mathbf{p}_{+i})' = [(.4 - .2), (.2 - .2), (.2 - .2)] \text{ and}$$

$$\mathbf{V} = \begin{bmatrix} \frac{.4 + .2 - 2(.2)}{100} & -\frac{.05 + 0}{100} & -\frac{.05 + 0}{100} \\ & \frac{.2 + .2 - 2(.1)}{100} & -\frac{.05 + .05}{100} \\ & & \frac{.2 + .2 - 2(.1)}{100} \end{bmatrix}$$

One finds that $\chi^2_S = 26.67$ is larger than $\chi^2_{3;.95} = 7.81$. Therefore, the null hypothesis of marginal homogeneity is rejected at $\alpha = .05$.

It is also possible to formulate an index of marginal agreement on the basis of Stuart's (1955) test, as follows:

$$M = 1 - \chi^2_S/n. \tag{11}$$

It can be shown that $\max(\chi^2_S) = n$, the sample size. (This maximum occurs when one rater assigns all objects to a single category and the other rater assigns all objects to a different category.) Therefore, the proposed index takes on a value of zero under maximal marginal disagreement and a value of one when the marginals are identical. For the example above,

$$M = 1 - \frac{21.82}{100} = .78.$$

Note that for a given table of observed proportions (e.g., Case 3 of Table 4), the value of M will be the same, regardless of sample size. Therefore, the M index may provide a useful supplement to or substitute for a test of statistical significance.

First Setting

To determine which categories are the source of rater disagreements, one can apply the post hoc procedures for Stuart's (1955) test, described by Marascuilo and McSweeney (1977) and Zwick, Neuhoff, Marascuilo, and Levin (1982). In fact, because these procedures do not involve matrix inversion, the researcher may want to perform only the category-by-category comparisons and bypass the overall tests.

Although Maxwell (1970) and Fleiss and Everitt (1971) also proposed that marginal agreement between raters be tested, researchers in psychology and education have rarely followed this recommendation. Applications appear to be more frequent in the biostatistical literature. For example, Landis and Koch (1977) illustrated a test of the homogeneity of rater marginals, essentially the same as Stuart's (1955) test, which can be formulated in terms of the GSK (Grizzle, Starmer, & Koch, 1969) approach to the analysis of categorical data. (The difference between the tests lies in the formulation of V. In Stuart's test, V is computed under the assumption that H_0 is true. This restriction is not imposed in the GSK approach.)

In Cases 1 and 2, it is obvious that the hypothesis of marginal homogeneity would be retained. One could then use π as a chance-corrected measure of agreement. π is always less than or equal to κ; the equality holds when the rater marginals are identical. For fixed values of $\frac{p_{i+} + p_{+i}}{2}$, π does not give credit, as does κ, for marginal discrepancies between raters. Cohen's (1960) objection to π—that it ignores differences in rater marginals—is no longer an issue if π is applied only when it is reasonable to assume marginal homogeneity. It is possible to test π for significance as well, although the standard error provided by Scott (1955) is not correct. Hubert (1977, pp. 293–294) used a matching model to derive the expected value and variance of a statistic equivalent to π, providing one possible approach to hypothesis testing.

References

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly, 18,* 303–308.

Brennan, R. L., & Prediger, D. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41,* 687–699.

Carey, G., & Gottesman, I. I. (1978). Reliability and validity in binary ratings. *Archives of General Psychiatry, 35,* 1454–1459.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement, 30,* 71–76.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31,* 651–659.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Fleiss, J. L., & Everitt, B. S. (1971). Comparing the marginal totals of square contingency tables. *British Journal of Mathematical and Statistical Psychology, 24,* 117–123.

Green, S. B. (1981). A comparison of three indexes of agreement between observers: Proportion of agreement, *G*-index, and kappa. *Educational and Psychological Measurement, 41,* 1069–1072.

Grizzle, J. E., Starmer, C. F., & Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics, 25,* 489–504.

Guilford, J. P. (1961, November). *Preparation of item scores for correlation between individuals in a Q factor analysis.* Paper presented at the annual convention of the Society of Multivariate Experimental Psychologists.

Holley, W., & Guilford, J. P. (1964). A note on the *G*-index of agreement. *Educational and Psychological Measurement, 24,* 749–753.

Hubert, L. (1977). Kappa revisited. *Psychological Bulletin, 84,* 289–297.

Janes, C. L. (1979). Agreement measurement and the judgment process. *The Journal of Nervous and Mental Disease, 167,* 343–347.

Janson, S., & Vegelius, J. (1979). On generalizations of the G index and the phi coefficient to nominal scales. *Multivariate Behavioral Research, 14,* 255–269.

Landis, J. R., & Koch, G. G. (1975a). A review of statistical methods in the analysis of data arising from observer reliability studies (Part I). *Statistica Neerlandica, 29,* 101–123.

Landis, J. R., & Koch, G. G. (1975b). A review of statistical methods in the analysis of data arising from observer reliability studies (Part II). *Statistica Neerlandica, 29,* 151–161.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences.* Monterey, CA: Brooks/Cole.

Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry, 116,* 651–655.

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry, 130,* 79–83.

McClung, J. (1963). *Dimensional analysis of inventory responses in the establishment of occupational personality types.* Unpublished doctoral dissertation, University of Southern California, Los Angeles.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12,* 153–157.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19,* 321–325.

Selvage, R. (1976). Comments on the analysis of variance strategy for the computation of intraclass reliability. *Educational and Psychological Measurement, 36,* 605–609.

Stuart, A. (1955). A test of homogeneity of marginal distributions in a two-way classification. *Biometrika, 42,* 412–416.

Whitehurst, G. J. (1984). Interrater agreement for journal manuscript reviews. *American Psychologist, 39,* 22–28.

Zwick, R., Neuhoff, V., Marascuilo, L. A., & Levin, J. R. (1982). Statistical tests for correlated proportions. *Psychological Bulletin, 92,* 258–271.

Received April 29, 1986
Revision received July 27, 1987
Accepted August 5, 1987 ■