DOCUMENT RESUME

ED 279 710 TM 870 122

AUTHOR	Zwick, Rebecca
TITLE	Another Look at Inter-Rater Agreement. Research Report.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	National Inst. of Mental Health (DHHS), Rockville, MD.
REPORT NO	ETS-RR-86-34
PUB DATE	Sep 86
CONTRACT	5-T32-MH-15745
NOTE	27p.; Product of the Psychometrics Research Group.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Error of Measurement; *Interrater Reliability; *Mathematical Models; *Measurement Techniques; Statistical Distributions; Statistical Studies; Vectors (Mathematics)
IDENTIFIERS	*Homogeneity of Variance

ABSTRACT

Most currently used measures of inter-rater agreement for the nominal case incorporate a correction for "chance agreement." The definition of chance agreement is not the same for all coefficients, however. Three chance-corrected coefficients are Cohen's Kappa; Scott's Pi; and the S index of Bennett, Goldstein, and Alpert, which has reappeared in many guises. For all three measures, chance is defined to include independence between raters. Scott's Pi involves a further assumption of homogeneous rater marginals under chance. For the S coefficient, uniform marginals for both raters under chance are assumed. Because of these disparate formulations, Kappa, Pi and S can lead to different conclusions about rater agreement. Consideration of the properties of these measures leads to the recommendation that a test of marginal homogeneity be conducted as a first step in the assessment of rater agreement. Rejection of the hypothesis of homogeneity is sufficient to conclude that agreement is poor. If the homogeneity hypothesis is retained, Pi can be used as an index of agreement. (Author)

RESEARCH REPORT

ANOTHER LOOK AT INTER-RATER AGREEMENT

Rebecca Zwick

ETS

Educational Testing Service
Princeton, New Jersey
September 1986

Another Look at Inter-rater Agreement

Rebecca Zwick

Psychometrics Research Group

Educational Testing Service

Acknowledgements

The author thanks Paul Holland and Skip Livingston for their comments. This research was supported in part by a 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 the author was a postdoctoral fellow.

Abstract

Most currently used measures of inter-rater agreement for the nominal case incorporate a correction for "chance agreement." The definition of chance agreement is not the same for all coefficients, however. Three chance-corrected coefficients are Cohen's κ, Scott's Π, and the S index of Bennett, Goldstein and Alpert, which has reappeared in many guises. For all three measures, chance is defined to include independence between raters. Scott's Π involves a further assumption of homogeneous rater marginals under chance. For the S coefficient, uniform marginals for both raters under chance are assumed. 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 a test of marginal homogeneity be conducted as a first step in the assessment of rater agreement. Rejection of the hypothesis of homogeneity is sufficient to conclude that agreement is poor. If the homogeneity hypothesis is retained, 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, it is often required that the classification task be performed by two raters. For k categories, the results can be tabled in a k x k agreement matrix in which the main diagonal contains the cases for which the raters agree.

A multitude of inter-rater agreement measures have been proposed by researchers in the fields of statistics, biostatistics, psychology, psychiatry, education, and sociology (see Landis & Koch [1975a, 1975b, 1977] for useful reviews). This article focuses on three coefficients that can be expressed in the form

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

where $P_O = \sum_{i=i}^{k} p_{ii}$ is the observed proportion of ageement, p_{ii} is the proportion of cases in the i^{th} 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 non-chance agreement. (Note, however, that this maximum can be achieved 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 inter-rater agreement than the "raw" proportion of agreement, P_O.

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

In the main portion of the article, the properties of S are compared to 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 1a. Some identities between coefficients are given in Table 1b.
In the final section of the paper, some recommendations are made for assessing inter-rater agreement in the nominal case. In particular, the need for examining the marginal distributions of the raters is stressed. Although most of the article focuses on a descriptive approach to the assessment of inter-rater agreement, an inferential procedure for assessing marginal homogeneity is

Table 1

A

Definition of $P_C(A)$ for κ, Π, and S

Coefficient	Definition
κ	$\sum_{i=1}^{k} p_{i+} p_{+i}$
Π	$\sum_{i=1}^{k} \left(\frac{p_{i+} + p_{+i}}{2}\right)^2$
S (G)	$1/k$

B

Identities Between Coefficients*

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

*In addition; the following identities hold by definition: $RE = G$, $C = \kappa_n = S$.

presented, along with a proposed marginal homogeneity index. Throughout the paper, a uniform notation system has been substituted for the notation used in the original presentations.

The S Coefficient of Bennett, Alpert, and Goldstein

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 same general subject matter as the poll. They proposed the following agreement coefficient:

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

The rationale 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" (p. 307).

The RE and G Coefficients for 2 x 2 Tables

Maxwell (1977) proposed an index of inter-rater agreement for 2 x 2 tables, called the RE (random error) coefficient, that has received some favorable attention in the literature (Carey & Gottesman, 1978; Janes, 1979). Maxwell's model for the assignment of subjects to categories can be outlined as follows: We assume that if both raters are "without doubt" in categorizing a subject, the raters must agree; if one or both raters is 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 I and II, 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 2. If we then wish to obtain the quantity $a_1 + a_2$, the proportion of agreement uncontaminated by doubtful cases, we proceed as follows:

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

where p_{ij} is the proportion of cases in the i^{th} row and the j^{th} column and $P_D = p_{12} + p_{21}$ is the proportion of disagreement. Maxwell's RE coefficient is algebraically equivalent to G, a measure of association for 2 x 2 tables proposed by Guilford (1961) and linear transformation to achieve this result:

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

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

It is not difficult to generalize Maxwell's model to the case of $k > 2$. If we let a_i $(i = 1, 2, \ldots k)$ represent the proportion

Table 2

Theoretical Cell Proportions for Maxwell's Model[a]

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

[a] a_1 and a_2 represent the proportions of "true" agreements for categories I and II.

of true agreement for the i^{th} category, then

$$P_0 = \sum_{i=1}^{k} a_i + \frac{1}{k}(1 - \sum_{i=1}^{k} a_i) \quad . \qquad (5)$$

If we let RE_k denote the generalized RE coefficient,

$$RE_k = \sum_{i=1}^{k} a_i$$

$$= (k - 1)[\sum_{i=1}^{k} a_i/(k - 1)]$$

$$= [k\sum_{i=1}^{k} a_i + (1 - \sum_{i=1}^{k} a_i) - 1]/(k - 1)$$

From Equation 5, we can see that this is equal to

$$RE_k = \frac{kP_0 - 1}{k - 1} = \frac{k}{k - 1}(P_0 - \frac{1}{k}) = S \qquad (6)$$

The C and κ_n Coefficients for k x k Tables

Janson and Vegelius (1979) proposed a coefficient, C, which is identical to RE_k. Although C was described as a generalization of the G index, its equivalence to S was not noted. Brennan and Prediger (1981) presented a coefficient, κ_n, which, as they noted (p. 693), is equivalent to S. (No mention was made of C, G, or RE.) For reasons described further below, Brennan and Prediger recommended that κ_n rather than κ, be used in typical inter-rater reliability studies.

Comparison of S, κ, and Π

To simplify the discussion below, RE, G, C, and κ_n are all referred to 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 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 under chance and then summing these products over the k categories. In its most general form, this sum can be expressed as

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

where h_{i+} is the hypothesized marginal proportion of cases assigned to category i by rater 1 under chance and h_{+i} is the corresponding proportion for rater 2. However, the three coefficients incorporate differing assumptions about the marginal distributions of each rater under chance, which, of course, are unobservable.

Let us now consider how each of the three agreement coeffficients defines the proportion of chance agreement. $P_C(S)$ is defined as 1/k. In this case, "chance" is understood to mean that the two raters independently assign cases to categories in a random fashion, each producing a uniform distribution; that is $h_{i+} = h_{+i} = 1/k$, i = 1, 2, ... k. Under these circumstances, 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 $k(1/k^2) = 1/k$. The assumption of random assignment of cases to categories, however, seems unlikely to hold: Even if both raters were ignorant of the rules to be used for assigning cases to categories, their marginal distributions might depart from uniformity because of a knowledge of the base rate (as in the case of diagnosis), a desire to minimize false positives or negatives with respect to a particular category, or a response set, such as a tendency to avoid categories perceived as extreme. If the unobservable marginal distributions departed from uniformity, the term 1/k would be an inappropriate chance correction. Minimization of the expression for $P_C(A)$ in Equation 7, subject to the constraints that $\sum_{i=1}^{k} h_{i+} = \sum_{i=1}^{k} h_{+i} = 1.00$, shows that min $[P_C(A)] = 1/k$. Therefore, 1/k is a lower bound to the proportion of agreement due to chance. It can be shown algebraically that underestimation of $P_C(A)$ leads to inflated values of A.

A less fundamental problem with the use of the S coefficient was noted by Scott (1955): For a fixed value of P_0, the value of S increases as the number of categories, k, increases: "Given a two-category sex dimension and a P_0 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" (Scott, 1955, p. 322).

Scott's (1955) Π coefficient was designed to overcome the defects of S. It does not involve an unrealistic assumption of random allocation under chance and does not become inflated by the inclusion of non-functional categories. $P_C(\Pi)$ is defined as $\sum_{i=1}^{k} (\frac{p_{i+} + p_{+i}}{2})^2$, where p_{i+} and p_{+i} are the observed marginal proportions for raters 1 and 2, respectively. Scott argued that "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" (Scott, 1955, p. 324). In computing Π, then, we assume that under chance, the raters would have identical marginals. We treat the quanitity $\frac{p_{i+} + p_{+i}}{2}$ as the unobservable proportion of cases assigned to category i by both raters under chance. In terms of Equation 7, we let $h_{i+} = h_{+i} = \frac{p_{i+} + p_{+i}}{2}$.

The Π index was criticized by Cohen, who remarked that "one source of disagreement between a pair of judges is precisely their proclivity to distribute their judgments differently over the categories" (Cohen, 1960, p. 41). A similar objection was raised by Fleiss (1975). Cohen (1960) recommended that κ, rather than Π, be used to assess rater agreement. $P_C(\kappa)$ is defined as $\sum_{i=1}^{k} p_{i+}p_{+i}$. Thus, "chance" in this context means independence of raters 1 and 2, given the obtained marginals. In applying κ, we make the assumption that each rater's distribution of cases to categories categories under chance would be the same as his or her observed distribution; that is $h_{i+} = p_{i+}$ and $h_{+i} = p_{+i}$. When raters have the same marginals, $\Pi = \kappa$ (and, for $k = 2$, $\Pi = \kappa = \phi$, the phi correlation). When, in addition, the marginals are uniform, as in Case I, $S = \Pi = \kappa$ (for any k).

To further explore 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 3 shows three cases, all of which have $P_O = .60$. Let us first consider the situation, represented in Cases I and II, in which the two raters have identical marginals. In Case I, $P_C(\kappa) = .25$ and $\kappa = .467$, whereas in Case II, $P_C(\kappa) = .28$ and $\kappa = .444$. κ is larger in Case I because, if both raters have the same marginal distributions, $P_C(\kappa)$ is minimized (and thus κ maximized) when the marginal distributions are uniform. (This property applies to Π as well.) This property of κ and the analogous property of the intraclass correlation in the ordinal case were found objectionable by Whitehurst (1984), who regarded it as a statistical artifact (see also Finn, 1970; Selvage, 1976). It is not clear, however, that the relationship between the shape of the marginal distributions and the size of κ is undesirable: If cases are concentrated into a small number of categories, we cannot determine whether our rating system includes decision criteria that are adequate for discrimination among all k categories. Therefore, it is not unreasonable that the value of an agreement coefficient should be smaller in this situation than in the case of uniform marginals.

Table 3

Values of κ, S, and Π for Three Cases

Rater 2

Categories	A	B	C	D	Total
	Case I: Marginals uniform (κ = .467, S = .467, Π = .467)				
Rater 1					
A	.20	-	-	.05	.25
B	-	.10	.15	-	.25
C	-	.15	.10	-	.25
D	.05	-	-	.20	.25
Total	.25	.25	.25	.25	1.00
	Case II: Marginals equal but not uniform (κ = .444, S = .467, Π = .444)				
Rater 1					
A	.20	.10	.10	-	.40
B	.10	.10	-	-	.20
C	.10	-	.10	-	.20
D	-	-	-	.20	.20
Total	.40	.20	.20	.20	1.00
	Case III: Marginals unequal (κ =.474, S = .467, Π = .460)				
Rater 1					
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

But let us consider another factor that affects the size of κ: the degree to which raters agree in their marginal distributions. In both Cases II and III of Table 3, $P_O = .60$. In Case II, where the raters have identical marginals, $P_C(\kappa) = .28$ and $\kappa = .444$. In Case III, however, where the raters have different marginals, $P_C(\kappa) = .24$ and $\kappa = .474$. Thus the raters in Case II 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 is certainly an undesirable property. Because there are ordinarily no external restrictions on the marginals, there appears to be no justification for treating marginal discrepancies as an obstacle which raters should be credited for overcoming.

Recommendations

It appears that S, Π, and κ all have major drawbacks. S requires the assumption of random assignment of cases to categories under chance, Π fails to take into account the differences between rater's marginals, and κ gives credit, for fixed P_O, to raters who produce different marginals. How, then, should inter-rater agreement be assessed? The answer lies in the examination of the degree of marginal agreement or homogenity per se. Rather than correcting for marginal disagreement, we should be studying it to determine whether we believe it reflects important rater differences or merely random error. The absence of discussion of this issue in the educational and psychological literature on chance-corrected agreement is striking. (Fleiss, 1965, is an exception, but only the dichotomous case is discussed.)

It is proposed 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 Π as a measure of chance-corrected agreement. The rationale for this approach is as follows. If we reject the hypothesis of marginal homogeneity, we need go no further: We have sufficient information to conclude that agreement is unsatisfactory. 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 we assess marginal homogeneity? If we have a fairly large random sample, we can make use of Stuart's (1955) test. The hypothesis of interest is H_0: $\underset{\sim}{\pi}_{i+} = \underset{\sim}{\pi}_{+i}$, where $\underset{\sim}{\pi}_{i+}$ is the k x 1 vector of elements π_{i+}, which represent the marginal probability of being in row i (corresponding to rater 1), and $\underset{\sim}{\pi}_{+i}$ is the corresponding vector of column probabilities (corresponding to rater 2). The test statistic is

$$\chi_S^2 = (\underset{\sim}{p}_{i+} - \underset{\sim}{p}_{+i})' \underset{\sim}{V}^{-1} (\underset{\sim}{p}_{i+} - \underset{\sim}{p}_{+i}) \ , \qquad (8)$$

where $(\underset{\sim}{p}_{i+} - \underset{\sim}{p}_{+i})$ is the (k - 1) x 1 vector of differences $(p_{i+} - p_{+i})$ between the i^{th} row marginal proportion and the i^{th} column marginal proportion for the first k - 1 categories. (The k^{th} difference is determined.) $\underset{\sim}{V}$ is the (k - 1) x (k - 1) variance-covariance matrix of the random vector $(\underset{\sim}{p}_{i+} - \underset{\sim}{p}_{+i})$, defined under H_0, with diagonal elements

$$v_{ii} = \frac{p_{i+} + p_{+i} - 2p_{ii}}{n} \qquad (9)$$

and off-diagonal elements

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

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 test.)

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

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

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

We find that $\chi^2_S = 21.82$ is larger than $\chi^2_{3:95} = 7.81$. Therefore, the null hypothesis of marginal homogeneity is rejected at $\alpha = .05$ and no further investigation is needed in order to conclude that rater agreement is inadequate.

It is also possible to formulate an index of marginal agreement, based on Stuart's 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 III of Table 3), the value of M will be the same, regardless of sample size.

To determine which categories are the source of rater disagreements, the post hoc procedures for Stuart's test, described by Marascuilo and McSweeney (1977) and Zwick, Neuhoff, Marascuilo, and Levin (1982) can be applied. 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 they have been ignored in education and psychology, tests of marginal homogeneity have been applied in this context by biostatisticians, such as Landis and Koch (1977). The test they illustrate, which can be formulated in terms of the GSK (Grizzle, Starmer, & Koch, 1969) approach to the analysis of categorical data, is essentially the same as Stuart's test. (The difference lies in the formulation of $\underset{\sim}{V}$. In Stuart's test, $\underset{\sim}{V}$ is computed under the assumption that H_0 is true. This restriction is not imposed in the GSK approach.)

In Cases I and II, it is obvious that the hypothesis of marginal homogeneity would be retained. We could then use Π as 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 objection to Π -- that it ignores differences in rater marginals -- is no longer an issue if Π is applied only when the marginal homogeneity hypothesis is retained. It is possible to test Π for significance as well, although the standard error provided by Scott (1955) is not correct. One possible approach to hypothesis testing is given by Hubert (1977, pp. 293-294), who uses a matching model to derive the expected value and variance of a statistic equivalent to Π.

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. (1965). Estimating the accuracy of dichotomous judgments. Psychometrika, 30, 469-479.

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.

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). Preparation of item scores for correlation betweenindividuals 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 index and the phi coefficient to nominal scales. Multivariate Behavioral Research, 14, 255-269.

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

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

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. (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.

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.