VOL. 70, NO. 4 OCTOBER 1968

Psychological Bulletin

WEIGHTED KAPPA:

NOMINAL SCALE AGREEMENT WITH PROVISION FOR SCALED DISAGREEMENT OR PARTIAL CREDIT

JACOB COHEN[1]
New York University

A previously described coefficient of agreement for nominal scales, kappa, treats all disagreements equally. A generalization to weighted kappa (κ_w) is presented. The κ_w provides for the incorporation of ratio-scaled degrees of disagreement (or agreement) to each of the cells of the $k \times k$ table of joint nominal scale assignments such that disagreements of varying gravity (or agreements of varying degree) are weighted accordingly. Although providing for partial credit, κ_w is fully chance corrected. Its sampling characteristics and procedures for hypothesis testing and setting confidence limits are given. Under certain conditions, κ_w equals product-moment r. Although developed originally as a measure of reliability, the use of unequal weights for symmetrical cells makes κ_w suitable as a measure of validity.

A previous article (Cohen, 1960) described κ (kappa), an index of agreement for use with nominal scales, as analogous to an alternate form reliability coefficient for magnitude-scaled data. Reliability has long been the keystone of psychometric theory (Gulliksen, 1950; Rozeboom, 1966), but the basic models have been developed for one-dimensional equal interval scales. Reliability plays the same crucial role in nominal scales as it does in magnitude scales (e.g., setting an upper bound for empirical validity), yet the relevant methodological literature is impoverished. The need to assess nominal scale reliability arises in fields as diverse as psychiatric diagnosis (Spitzer, Cohen, Fleiss, & Endicott, 1967) and survey interview coding (Scott, 1955).

Past approaches to the problem were deficient both in the indices used to measure degree of agreement and in their statistical treatment. The most frequently used index has been percentage or proportion of agreement (p_o—in Table 1), which suffers in that it includes agreement which can be accounted for by chance. Occasionally, the $k \times k$ table of joint categorical assignment frequencies (where each "judge" has made assignments to the same k-level nominal scale) has been treated as a contingency table, and the contingency coefficient, C, based on chi-square, χ^2, (McNemar, 1962) has been used as a measure of agreement. The defect of χ^2 in this context, and therefore of C, is that it indexes *association* and not necessarily *agreement*, which is the special kind of association of interest in reliability. In an agreement matrix, high reliability dictates that the values observed in the k cells of the leading or agreement diagonal be higher than the chance expectation dictated by the marginal values, and that, conversely, the off-diagonal cells representing disagreement have observed values which are smaller than those expected by chance. The χ^2 and hence C increase monotonically with increases in the absolute discrepancies between observed and chance-expected values in each of the cells, whether these discrepancies are in the direction of agreement or disagreement, quite impartially.

[1] The author is greatly indebted to Joseph L. Fleiss of Columbia University School of Public Health for developing the asymptotic standard error of an observed κ_w. Acknowledgments are also due Robert L. Spitzer of Biometrics Research, whose research in computer-based psychiatric diagnosis stimulated the work reported here, and Patricia Waly, for a critical reading of the manuscript.

Cohen (1960) proposed as a coefficient of agreement for nominal scales, the proportion of agreement corrected for chance

$$\kappa = \frac{p_o - p_c}{1 - p_c} \qquad [1]$$

where p_o is the observed proportion of agreement, and p_c is the proportion of agreement expected by chance. This is found by summing over the agreement diagonals, the product of the proportions for the row and column of the cell, as illustrated by the parenthetical values in each cell of Table 1. Cohen also presented large sample formulae for the standard error of an observed κ

$$\sigma_\kappa \cong \sqrt{\frac{p_o(1 - p_o)}{N(1 - p_c)^2}} \qquad [2]$$

used for setting confidence limits and performing two-sample hypothesis tests, and the standard error of κ when the population $\kappa = 0$

$$\sigma_{\kappa_0} = \sqrt{\frac{p_c}{N(1 - p_c)}} \qquad [3]$$

used for one-sample significance tests of κ (1960, Formulas 7 and 10). For large samples, the sampling distribution of κ is approximately normal, and statistical tests using normal curve deviates take the familiar classical form (Cohen, 1960).

Thus, κ provides a conceptually simple measure of reliability for nominal scales: the proportion of agreement after agreement which can be attributed to chance has been removed both from the base and from the numerator, as

TABLE 1

AN AGREEMENT MATRIX OF PROPORTIONS WITH ILLUSTRATIVE COMPUTATIONS OF κ, κ_w AND RELEVANT STATISTICS

		Judge A: Personality disorder	Judge A: Neurosis	Judge A: Psychosis	$p_{i.}$
Judge B	Personality disorder	0[a] (.30)[b] .44[c]	1 (.18) .07	3 (.12) .09	.60
	Neurosis	1 (.15) .05	0 (.09) .20	6 (.06) .05	.30
	Psychosis	3 (.05) .01	6 (.03) .03	0 (.02) .06	.10
	$p_{.j}$	.50	.30	.20	1.00

Note.—$N = 200$

$$q_o = 1 - p_o = 1 - (.44 + .20 + .06) = .30$$
$$q_e = 1 - p_e = 1 - (.30 + .09 + .02) = .59$$
$$\kappa = 1 - \frac{.30}{.59} = .492 \qquad [4]$$

$$\Sigma v_{ij} p_{oij} = 0(.44) + 1(.07) + 3(.09) + 1(.05) + \cdots + 0(.06) = .90$$
$$\Sigma v_{ij} p_{eij} = 0(.30) + 1(.18) + 3(.12) + 1(.15) + \cdots + 0(.02) = 1.38$$
$$\kappa_w = 1 - \frac{.90}{1.38} = .348 \qquad [8]$$

$$\Sigma v_{ij}^2 p_{oij} = 0^2(.44) + 1^2(.07) + 3^2(.09) + 1^2(.05) + \cdots + 0^2(.06) = 3.90$$
$$\Sigma v_{ij}^2 p_{eij} = 0^2(.30) + 1^2(.18) + 3^2(.12) + 1^2(.15) + \cdots + 0^2(.02) = 5.10$$
$$\sigma_{\kappa_w} \cong \sqrt{\frac{3.90 - .90^2}{200(1.38^2)}} = .0901 \qquad [10]$$

$$\sigma_{\kappa_{w0}} = \sqrt{\frac{5.10 - 1.38^2}{200(1.38^2)}} = .0916 \qquad [13]$$

95% Confidence Limits on κ_w: $\kappa_w \pm 1.96\sigma_{\kappa_w} =$
$.348 \pm 1.96(.0901) =$
$.171 \leq \kappa_{w_{pop}} \leq .525$

[a] Disagreement weight v_{ij}
[b] Chance-expected cell proportion, $p_{oij} = p_{i.}p_{.j}$
[c] Observed cell proportion, p_{oij}

Formula 1 directly sets forth. It quite reasonably yields negative values when there is less observed agreement than is expected by chance, zero when observed agreement can be (exactly) accounted for by chance, and unity when there is complete agreement.

The further development to κ_w (*weighted* kappa) is motivated by studies in which it is the sense of the investigator that some disagreements in assignments, that is, some off-diagonal cells in the $k \times k$ matrix, are of greater gravity than others. For example, in an assessment of the reliability of psychiatric diagnosis in the categories: (*a*) personality disorder (D), (*b*) neurosis (N), and (*c*) psychosis (P), a clinician would likely consider a diagnostic disagreement between neurosis and psychosis to be more serious than between neurosis and personality disorder (see Table 1). The κ makes no such distinction, implicitly treating all disagreement cells equally. This article describes the development of κ_w, the proportion of *weighted* agreement corrected for chance, to be used when different kinds of disagreement are to be differentially weighted in the agreement index.

Development of Weighted Kappa

The desired weighting is accomplished by an a priori assignment of weights to the k^2 cells of the $k \times k$ matrix, that is, a ratio scaling of the cells. Either degree of agreement or degree of disagreement may be scaled, depending on what seems more natural in the given context. The development here will be in terms of disagreement ratio scaling, for example, 6 represents twice as much disagreement as 3. This will be supplemented later with formulae for use with agreement scaling. Note that in either case, the result is κ_w, a chance-corrected proportion of weighted *agreement*.

We begin with the basic formula for κ (Equation 1). If we define $q = 1 - p$ as the proportion of disagreement, $p = 1 - q$. Substituting $p_o = 1 - q_o$ and $p_c = 1 - q_c$ into Equation 1 and simplifying yields

$$\kappa = \frac{q_c - q_o}{q_c} = 1 - \frac{q_o}{q_c} \qquad [4]$$

an equation for κ expressed in terms of observed and chance disagreement. κ_w simply replaces q_o and q_c by proportions of weighted disagreement, q'_o and q'_c. To find the latter, each of the k^2 cells must have a disagreement weight, v_{ij}, where the ij subscript indexes the cell ($i,j = 1 \cdots k$). These (positive) weights can be assigned by means of any judgment procedure set up to yield a ratio scale (Torgersen, 1958) including the simple direct scaling advocated by Stevens (1958). In many instances, they may be the result of a consensus of a committee of substantive experts, or even, conceivably, the investigator's own judgment. They are, in any case, to be *ratio* weights. It is convenient (but not necessary) to assign zero to the "perfect" agreement diagonal ($i = j$), that is, *no* disagreement. A weight which represents maximum disagreement (v_{max}) is assigned at the convenience of the investigator (for Table 1, it is 6). For any set of v_{ij}, κ_w is invariant over any positive multiplicative transformation, that is, κ_w will not change if its v_{ij} are multiplied by any value greater than zero.

The brief attention to the setting of these weights should not mislead the reader as to their importance. The weights assigned are an integral part of how agreement is defined and therefore how it is measured with κ_w. Moreover, its standard error is also a function of the v_{ij} (or for agreement weighting, the w_{ij}), so that the results of significance tests are also dependent upon the weights. Another way of stating this is that the weights are part of any hypothesis being investigated. An obvious consequence of this is that the weights, however determined, must be set prior to the collection of the data.

Proportions of weighted disagreement, observed and chance, are simply weighted functions over the k^2 cells of the p_{oij} and p_{cij}, respectively, namely

$$q'_o = \frac{\sum v_{ij} p_{oij}}{v_{max}} \qquad [5]$$

$$q'_c = \frac{\sum v_{ij} p_{cij}}{v_{max}} \qquad [6]$$

where the p_{oij} is the proportion of the joint judgments (N in number) observed in the ij cell, and the p_{cij} the proportion in the cell expected by chance, as illustrated in Table 1.

(The summation throughout is over all k^2 cells.) Weighted kappa is then given by

$$\kappa_w = 1 - \frac{q'_o}{q'_c} \qquad [7]$$

When Formulas 5 and 6 are substituted in Formula 7, the v_{max} term drops out, and it simplifies to

$$\kappa_w = 1 - \frac{\sum v_{ij} p_{oij}}{\sum v_{ij} p_{cij}} \qquad [8]$$

Table 1 illustrates the computation of both κ and κ_w, using unweighted and weighted disagreement proportions, respectively. The matrix of proportions of joint assignments in Table 1 is obtained from the usual $k \times k$ table of joint frequencies or paired assignments in which cell, marginal and total (N) observed frequencies have been divided by the latter. To find the p_{cij}, the marginal proportions for the ij cell are multiplied, for example, for the upper left cell, $p_{c11} = p_{1.}p_{.1} = p_{1.}p_{.1} = (.60)(.50) = .30$, given in parentheses in that cell.

For κ, only the p_{oij} and p_{cij} values in the agreement diagonal ($i = j$) are needed, and κ is found from Formula 4 to equal .492, that is, after chance agreement is excluded, about half the judgments are in agreement, all disagreements being counted equally.

For κ_w, the weighted sums over all cells (numerators of Formulas 5 and 6) are substituted in Formula 8.[2] For the v_{ij} in Table 1, $\kappa_w = .348$.

The values of Table 1 were selected in order to emphasize a point which might otherwise go unappreciated: like κ, κ_w is *fully* chance corrected. One might suppose, since the cells are scaled for degrees of disagreement, that this is like not giving some cells *full* disagreement credit (i.e., the obverse of giving partial agreement credit), and that therefore κ_w relative to κ is biased in an upward direction, that is, it overstates agreement. The premise is correct, but the consequent is not. The same weights which generate q'_o also generate q'_c; and κ_w may well be smaller than κ for the same data, as is the case in Table 1. This will occur when the algebraically smaller values of $p_{cij} - p_{oij}$ occur in cells which have large v_{ij} values. It occurs in Table 1 because the $N - P(2,3)$ and $P - N(3,2)$ disagreement cells, which have the largest $v_{ij} = 6$ show $p_{cij} - p_{oij}$ discrepancies of only .01 and .00, while the less serious $D - N$ and $N - D$ disagreement cells ($v_{ij} = 1$) show discrepancies of .12 and .10. This means that these judges disagree much less than chance expectation where it doesn't count very much and disagree at about the chance level where it counts greatly. The result is κ_w smaller than κ. If the v_{ij}'s of 6 and 1 are interchanged in the table, κ_w becomes .574, a value greater than κ.

For a computing formula using frequencies rather than proportions, one simply substitutes f for p values in Formula 8

$$\kappa_w = 1 - \frac{\sum v_{ij} f_{oij}}{\sum v_{ij} f_{cij}} \qquad [9]$$

where f_{oij} is the observed frequency in cell ij, and f_{cij} is the chance-expected frequency in cell ij, computed as for a χ^2 contingency table.

Sampling Characteristics

The asymptotic (large sample approximation) standard error of κ_w is

$$\sigma_{\kappa_w} \cong \sqrt{\frac{\sum v^2_{ij} p_{oij} - (\sum v_{ij} p_{oij})^2}{N(\sum v_{ij} p_{cij})^2}} \qquad [10]$$

or, in terms of cell frequencies

$$\sigma_{\kappa_w} \cong \sqrt{\frac{N \sum v^2_{ij} f_{oij} - (\sum v_{ij} f_{oij})^2}{N(\sum v_{ij} f_{cij})^2}} \qquad [11]$$

The use of Formula 10 is illustrated in Table 1. (Note that it requires, in addition to the terms required by Formula 8, the determination of $\sum v^2_{ij} p_{oij}$.) Since the sampling distribution of κ_w is approximately normal for large samples, κ_w can be used for setting confidence limits on a sample κ_w together with the appropriate unit normal curve deviate (illustrated in Table 1 for 95% limits, where $z = \pm 1.96$), and also for a normal curve test of the significance of the difference between two independent κ_ws

$$z = \frac{\kappa_{w_1} - \kappa_{w_2}}{\sqrt{\sigma^2_{\kappa_{w_1}} + \sigma^2_{\kappa_{w_2}}}} \qquad [12]$$

[2] When the v_{ij} in the diagonal cells are set at zero, they contribute to neither Equation 5 nor 6, so that in practice, the summations are actually over only $k^2 - k$ cells.

To test an obtained κ_w for significance, the standard error of κ_w when the population κ_w equals zero is needed. It is obtained by substituting chance for observed cell proportions where the latter appear in Formula 10

$$\sigma_{\kappa w_0} = \sqrt{\frac{\sum v^2_{ij}p_{ci}p_{cj} - (\sum v_{ij}p_{ci}p_{cj})^2}{N(\sum v_{ij}p_{ci}p_{cj})^2}} \quad [13]$$

the computation of which is illustrated in Table 1. In terms of frequencies

$$\sigma_{\kappa w_0} = \sqrt{\frac{N\sum v^2_{ij}f_{ci}f_{cj} - (\sum v_{ij}f_{ci}f_{cj})^2}{N(\sum v_{ij}f_{ci}f_{cj})^2}} \quad [14]$$

A significance test of κ_w, that is, a test of H_0: Population $\kappa_w = 0$, is accomplished by evaluating the normal curve deviate

$$z = \frac{\kappa_w}{\sigma_{\kappa w_0}} \quad [15]$$

For the data in Table 1, $\kappa_w = .348$, $\sigma_{\kappa w_0} = .0901$, and

$$z = \frac{.348}{.0916} = 3.80, \text{ significant at } p < .001.$$

It should be noted that the demonstration that a population κ_w is greater than zero, at whatever significance level, is, in general, hardly impressive. The κ_w here is a reliability coefficient, and one normally wishes evidence that the population κ_w is some relatively large value, rather than merely that it highly probably exceeds zero. Thus, in most instances, a substantial value for the lower confidence limit (at, say, 95%) rather than zero as implied by the null hypothesis, is a more meaningful criterion for the adequancy of nominal scale reliability.

Weighted Kappa through Agreement Scaling

The κ_w can be developed with cell weights which reflect agreement (w_{ij}) rather than disagreement (v_{ij}). When the concept of "full" credit for complete agreement and varying amounts of "partial" credit (possibly including no credit) for different off-diagonal ($i \neq j$) cells seems natural in a given context, agreement is scaled so as to yield a ratio scale of positive agreement weights, w_{ij}, ranging down from some convenient maximum value assigned to the diagonal ($i = j$) cells representing complete agreement (full credit).[3] The use of zero as the minimum w_{ij} ("no" credit) is convenient, but not necessary. As with the v_{ij}, κ_w is invariant over multiplication of the w_{ij} by any value greater than zero. The stress on the importance of the v_{ij} when they were discussed in the preceding section extend, of course, to the w_{ij}.

Parallel to the above development, we define weighted proportions of observed and chance agreement

$$p'_o = \frac{\sum w_{ij}p_{oij}}{w_{max}} \quad [16]$$

$$p'_c = \frac{\sum w_{ij}p_{cij}}{w_{max}} \quad [17]$$

By replacing weighted for unweighted proportions of agreement in the basic formula for κ (Formula 1), we obtain

$$\kappa_w = \frac{p'_o - p'_c}{1 - p'_c} \quad [18]$$

Substituting the values of Equations 16 and 17 and simplifying yields

$$\kappa_w = \frac{\sum w_{ij}p_{oij} - \sum w_{ij}p_{cij}}{w_{max} - \sum w_{ij}p_{cij}} \quad [19]$$

In terms of frequencies

$$\kappa_w = \frac{\sum w_{ij}f_{oij} - \sum w_{ij}f_{cij}}{w_{max}N - \sum w_{ij}f_{cij}} \quad [20]$$

Also

$$\sigma_{\kappa_w} \cong \sqrt{\frac{\sum w^2_{ij}p_{oij} - (\sum w_{ij}p_{oij})^2}{N(w_{max} - \sum w_{ij}p_{cij})^2}} \quad [21]$$

and, in terms of frequencies

$$\sigma_{\kappa_w} \cong \sqrt{\frac{N\sum w^2_{ij}f_{oij} - (\sum w_{ij}f_{oij})^2}{N(w_{max}N - \sum w_{ij}f_{cij})^2}} \quad [22]$$

Finally, $\sigma_{\kappa w_0}$ is given, for proportions and frequencies, respectively, by replacing observed by chance values wherever the former appear in Formulas 21 and 22.

[3] Because of the ratio property of both the w_{ij} and v_{ij}, the relationship between the two kinds of weights is complementary when they are expressed as proportions of their respective maximum values. Specifically, κ_w remains constant when $(w_{ij}/w_{max}) = 1 - (v_{ij}/v_{max})$ which yields $w_{ij} = (w_{max}/v_{max})(v_{max} - v_{ij})$.

Statistical manipulations, such as setting confidence limits and performing the significance tests of Formulas 12 and 15 are, of course, performed in exactly the same way as when their components were found through disagreement weights.

Weighted Kappa and Kappa

A perspective on κ_w is afforded by considering its relationship to κ. The κ is simply proportion of agreement (p_o) corrected for chance, and κ_w can readily be thought of as a generalization of κ, proportion of *weighted* agreement (as in Formula 16) corrected for chance. The relationship may be more clearly understood if it is inverted: κ is a special case of κ_w. In κ_w we may differentially weight, either by v_{ij} or w_{ij}, the off-diagonal ($i \neq j$) cells, because we mean to consider the various *kinds* of disagreement as representing differing *amounts* of disagreement (or, equivalently, differing amounts of agreement). For κ, the $k(k-1)$ off-diagonal cells representing disagreement are simply treated as if they all represented the same amount of disagreement. It is a matter of simple algebra to show that if all $v_{ij}(i \neq j)$ are given the same weight, unity or any other value greater than the agreement diagonal weights ($= 0$ in the example), κ_w becomes κ. That is, with v_{ij} $(i = j) = 0$ and all v_{ij} $(i \neq j)$ = a constant, the constant cancels out in Formula 8 for κ_w, leaving κ of Formula 4. Similarly, a constant value for w_{ij} $(i \neq j)$, smaller than another constant w_{ij} $(i = j)$ $= w_{max}$, reduces Formula 19 for κ_w to Formula 1 for κ. Thus, κ is the special case of κ_w where all disagreements are given the same weight. Furthermore, under these conditions, the standard error formulas for κ_w simplify to those for κ (Formulas 2 and 3).

Weighted Kappa and Product-Moment Correlation

It is a frequent experience for the methodologist exploring an area apparently remote from product-moment correlation (r) to turn a corner and find it confronting him (for a recent example see Glass, 1966). A discovery of this kind may be of greater importance in its illumination of r than of the area being explored. Such a discovery was made in the case of κ_w. Under certain simple conditions, $\kappa_w = r$. The conditions are these: (*a*) The marginal distributions are the same, that is, $p_{i.} = p_{.j}$ for $i = j$. (*b*) Disagreement weights (v_{ij}) are assigned according to the following pattern: The k cells of the agreement diagonal ($i = j$) have $v_{ij} = 0$. The $k - 1$ cells in each of the two adjacent diagonals have v_{ij} of 1, the $k - 2$ cells in each of the next diagonals out on either side have $v_{ij} = 2^2 = 4$, the $k - 3$ cells in each of the next diagonals have $v_{ij} = 3^2 = 9$, and so on until one reaches the last cell in the upper right and lower left corner whose weights are $(k-1)^2$. For example, for $k = 5$, the pattern of v_{ij} is

0	1	4	9	16
1	0	1	4	9
4	1	0	1	4
9	4	1	0	1
16	9	4	1	0

Using these weights, one can compute κ_w with Formulas 8 or 9.

Now, give the nominal categories *scores* equal to their index numbers, that is, the first category is scored 1, the second is scored 2, etc. If the product-moment r is computed from the observed frequencies or proportions using these scores (or linear transformations thereof), r is found identical to the κ_w above.[4]

This identity did not come as a complete surprise. In the article presenting κ (Cohen, 1960), it was shown that for the 2×2 table under the equal marginal condition, $\kappa = \phi$, the fourfold point correlation coefficient (phi coefficient). For the 2×2 table with symmetrical assignment of weights ($v_{12} = v_{21}$, or $w_{12} = w_{01})\kappa_w$ perforce equals κ. On the other hand, ϕ is simply a product-moment r for dichotomous data. Thus, the previous finding of $\kappa = \phi$ is a special case of the present more general finding that $\kappa_w \equiv r$ under the stated conditions.

[4] The proof proceeds by writing the "difference" formula for r, then letting the means and standard deviations of the two distributions be equal (as in Cohen, 1957, Formula 5), the latter being a consequence of the first condition. In this form, the formula for r has the same structure as that of κ_w (Formulas 8 and 9). If one then notes that the v_{ij} of the required pattern are, in fact, equal to $D_{ij}^2 = (X_i - Y_j)^2$ for $i,j = 1,2 \cdots k$, one can see how the proof proceeds.

Note that the two conditions, equal marginals and the prescribed pattern of weights, are not of equal importance. Relaxation of the first to the degree of inequality of marginals normally found in real data reduces the equality of κ_w and r to a close approximation, with $\kappa_w < r$, but by no more than a few hundredths.

The equality of κ_w and r under the stated conditions is primarily of interest for the new conception it offers for r. It means that r can be conceived as a (suitably chance-corrected) proportion of agreement, where the disagreements are weighted so as to be proportional to the square of the distance between the pair of measures, or, equivalently, from the $X = Y$ line. (In the general case, where X and Y are not expressed in the same units, they must be conceived as first being transformed to a common standard score.) Perhaps this should not be surprising, given the role of least squares in the definition of r. The tendency of statistical neophytes to interpret r as a proportion may be more constructively dealt with pedagogically by showing them by this route (there are others) exactly what kind of a proportion it is.

Weighted Kappa as a Validity Measure

The examples and discussion to this point have implicitly assigned equal weights to symmetric cells, that is, $v_{ij} = v_{ji}$ (or $w_{ij} = w_{ji}$). An error which comes about from Judge A assigning "Neurotic" where Judge B assigns "Psychotic" is of the same gravity, or gets the same partial credit as one in which their assignments are reversed. This is appropriate to the frame of reference of reliability, where the two sources of data are conceived as being of equal status, that is, as alternate forms. Some reflection suggests that the formal difference between reliability and validity lies in the contrast between the equal status of the sources in the former and their differing status in validity, where one is a predictor and the other a criterion. When validity is being assessed, it may (but need not) be eminently reasonable for $v_{ij} \neq v_{ji}$ (or $w_{ij} \neq w_{ji}$). It is this conception which is operative in the different costs attached to false positives and false negatives in dichotomous diagnosis, and in the different values given producer's risk and consumer's risk in statistical quality control. Since there is nothing in the conception or statistical manipulation of κ_w which demands weight symmetry, it can be used for $k \times k$ tables constructed for assessing nominal (and indeed, stronger than merely nominal) scale validity.

For example, reinterpret the situation in Table 1 as follows: Consider Judge A to be the consensus diagnosis of a panel of distinguished diagnosticians—the criterion; and Judge B the diagnosis made by a computer—the predictor, or variable being tested (Spitzer et al., 1967). Given the way the computer diagnosis is to be used, it may well be considered, for example, that a computer error in making a diagnosis of Neurosis when the panel consensus is Psychosis is more serious than a computer diagnosis of Psychosis when the panel consensus is Neurosis. This is realized in the definition of agreement by assigning different weights to these symmetrical cells. For this use of κ_w, the pattern of v_{ij} which are finally assigned may look like this:

		Panel		
		D	N	P
Computer	D	0	1	4
	N	1	0	6
	P	2	2	0

Such a pattern implies a greater concern about failing to identify psychotics (more so by calling them neurotics) than for mistakenly identifying them, and less (and symmetrical) concern for the Neurosis-Personality Disorder confusion, whichever way the error is made.

Assuming the proportions of Table 1 with these new weights, it is found that $\sum v_{ij}p_{oij} = .86$, $\sum v_{ij}p_{eij} = 1.33$, and therefore $\kappa_w = .353$ (Formula 8). With these new weights, $\sigma\kappa_w = .0887$ (Formula 10) and $\sigma\kappa_{w_0} = .0915$ (Formula 13). The κ_w remains the chance-corrected proportion of weighted agreement, but now the weights reflect the "costs" or "utilities" perceived in this situation and their structure is appropriate to what is intended by the "validity" of computer diagnosis.

REFERENCES

Cohen, J. An aid in the computation of correlations based on Q sorts. *Psychological Bulletin*, 1957, 54, 138–139.

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

GLASS, G. V. Note on rank biserial correlation. *Educational and Psychological Measurement*, 1966, **26**, 623–631.

GULLIKSEN, H. *Theory of mental tests*. New York: Wiley, 1950.

MCNEMAR, Q. *Psychological statistics*. (3rd ed.) New York: Wiley, 1962.

ROZEBOOM, W. W. *Foundations of the theory of prediction*. Homewood, Ill.: Dorsey, 1966.

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

SPITZER, R. L., COHEN, J., FLEISS, J. L., & ENDICOTT, J. Quantification of agreement in psychiatric diagnosis. A new approach. *Archives of General Psychiatry*, 1967, **17**, 83–87.

STEVENS, S. S. Problems and methods of psychophysics. *Psychological Bulletin*, 1958, **55**, 177–196.

TORGERSEN, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

(Received October 19, 1967)