EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XX, No. 1, 1960

A COEFFICIENT OF AGREEMENT FOR NOMINAL SCALES[1]

JACOB COHEN
New York University

CONSIDER Table 1. It represents in its formal characteristics a situation which arises in the clinical-social-personality areas of psychology, where it frequently occurs that the only useful level of measurement obtainable is nominal scaling (Stevens, 1951, pp. 25–36), i.e. placement in a set of k unordered categories. Because the categorizing of the units is a consequence of some complex judgment process performed by a "two-legged meter" (Stevens, 1958), it becomes important to determine the extent to which these judgments are reproducible, i.e., reliable. The procedure which suggests itself is that of having two (or more) judges independently categorize a sample of units and determine the degree, significance, and

TABLE 1

An Agreement Matrix of Proportions

		Judge A			
	Category	1	2	3	p_{iB}
Judge B	1	.25 (.20)*	.13 (.15)	.12 (.15)	.50
	2	.12 (.12)	.02 (.09)	.16 (.09)	.30
	3	.03 (.08)	.15 (.06)	.02 (.06)	.20
	p_{iA}	.40	.30	.30	$\sum p_i = 1.00$

$$p_o = .25 + .02 + .02 = .29$$
$$p_c = .20 + .09 + .06 = .35$$

* Parenthetical values are proportions expected on the hypothesis of chance association, the joint probabilities of the marginal proportions.

[1] From the Psychiatric Evaluation Project of the Psychology Service, Veterans Administration Hospital, Montrose, New York. Acknowledgements are due the staff, particularly H. Spohn, L. Solomon, and A. Steinman, whose discussions with the author led to this article, and to Catherine S. Henderson, who typed the manuscript.

sampling stability of their agreement. This quite parallels in its logic the concept of the coefficient of equivalence used with tests—the judges are analogous to alternate forms, and the nominal data are analogous to scores.

Thus, the judges may be clinical psychologists, the categories schizophrenic, neurotic, and brain-damaged, and the units psychological test protocols; or the judges may be social psychologists, the categories various types of leadership, and the units small groups, etc. These have in common the following conditions, which may be taken as assumptions of the coefficient of agreement to be proposed:

1. The units are independent.
2. The categories of the nominal scale are independent, mutually exclusive, and exhaustive.
3. The judges operate independently.

In the typical situation, there is no criterion for the "correctness" of judgments, and the judges are a priori deemed equally competent to make judgments. Also, there is no restriction placed on the distribution of judgments over categories for either judge.

An implication of the lack of order of the categories needs to be pointed up. Unlike stronger measurement situations, discrepancies between paired judgments are treated as equal to each other, e.g., a psychoneurotic-schizophrenic discrepancy counts equally with a psychoneurotic-brain damaged discrepancy.

In the literature, situations of the type herein considered have been variously handled. The most primitive approach has been to simply count up the proportion of cases in which the judges agreed, p_o, and let the issue rest there. Thus, for Table 1, there is .29 agreement.

It takes relatively little in the way of sophistication to appreciate the inadequacy of this solution. A certain amount of agreement is to be expected by chance, which is readily determined by finding the joint probabilities of the marginals; e.g., in Table 1 Judge A has placed .40 of his units in Category 1, while Judge B has placed .50 of his units in this category, leading us to expect chance agreement for Category 1 to be (.40) (.50) = .20. (These values are the parenthetical entries in Table 1.) This having been done, many investigators have computed χ^2 over the table for use as a test of the hypothesis of chance agreement, and some have gone on to compute the contingency coefficient (C) as a measure of degree of agreement (Guilford, 1950, pp. 343–345).

For Table 1, if we assume an N of 200, χ^2 is found to equal 64.59 (4 degrees of freedom), a highly significant result.[2] At this point some investigators would rest content that agreement is adequate, and others would go on to find C, which equals .49.

It is readily demonstrable that the use of χ^2 (and therefore the C which is based on it) for the evaluation of agreement is indefensible. When applied to a contingency table, χ^2 tests the null hypothesis with regard to *association*, not agreement. In Table 1, the largest contribution to the χ^2 comes from cell A2-B3, where the cell contribution is $\frac{(.15 - .06)^2}{.06}(200) = 27.00$. This large value reflects the fact that the judges *disagreed* to an extent significantly greater than chance. Thus, χ^2 and C are here improperly used for measuring agreement, since they will be inflated quite impartially by any departure from chance association, either disagreement or agreement.

That the judges in Table 1 do not agree adequately has no doubt already been noted by the reader. The proportion of observed agreement of .29 is less than the proportion of agreement to be expected by chance (p_c) of .35, found by simply adding the parenthetical (chance) values in the agreement diagonal. The significant χ^2 simply means that the judgments are associated, but unfortunately not in the direction of agreement.

The purpose of this article is to present a coefficient to measure the degree of agreement in nominal scales, and to provide means of testing hypotheses and setting confidence limits for this coefficient.

A Coefficient of Agreement

The discussion thus far suggests that, for any problem in nominal scale agreement between two judges, there are only two relevant quantities:

p_o = the proportion of units in which the judges agreed

p_c = the proportion of units for which agreement is expected by chance.

The test of agreement comes then with regard to the $1 - p_c$ of the units for which the hypothesis of no association would predict disagreement between the judges. This term will serve as the denominator.

[2] For a table of proportions, χ^2 is N times the value obtained by performing the usual operations on the proportions rather than the frequencies.

To the extent to which nonchance factors are operating in the direction of agreement, p_o will exceed p_c; their difference, $p_o - p_c$, represents the proportion of the cases in which beyond-chance agreement occurred and is the numerator of the coefficient.

The coefficient κ is simply the proportion of chance-expected disagreements which do not occur, or alternatively, it is the proportion of agreement *after* chance agreement is removed from consideration:

$$\kappa = \frac{p_o - p_c}{1 - p_c} \tag{1}$$

Expressed in frequencies to facilitate computation,

$$\kappa = \frac{f_o - f_c}{N - f_c} \tag{2}$$

There are approaches to this problem in the literature which resemble κ. In 1941, Guttman (1941, pp. 258–263) proposed a generalized measure of association usable for qualitative variates, which when so used was named λ by Goodman and Kruskal (1954, p. 758). λ is of identical form with κ, but Guttman's equivalent of p_c is quite differently defined. Proceeding from the point of view of prediction, Guttman (1941, p. 262) contrasts the "predictability of A before and after knowledge of B." However, before knowledge of B, all A is predicted as falling in the modal (largest) category of A ("optimal prediction"). When applied to reliability, Goodman and Kruskal (1954, p. 757) define the equivalent of p_c to be the mean of the two judges' modal categories (e.g., for Table 1 the value would be the mean of .40 and .50 = .45). Although a good case can be made for defining p_c in terms of the modal category for prediction, its use for the reliability problem is questionable—each judge in fact distributes his judgments over the k categories, he does not simply lump them in one. The determination of p_c for κ follows from the logic of the reliability situation, accords with the familiar approach to contingency tables, and results in a coefficient which is simply and directly interpreted.

A coefficient recently proposed by Scott (1955) is also quite similar to κ. Here, too, p_c is differently defined. Working in the area of content analysis in survey research, he assumes for his "coefficient of intercoder agreement," π, that the distribution of proportions over the categories for the population is known and is taken to be equal for the judges. The former assumption is reasonable in survey research, but the latter may be questioned in more general applications, since one source of disagreement between a pair of judges is precisely their proclivity to distribute their judgments differently over the categories (as in Table 1: compare the p_{iA} and the p_{iB} values).

Characteristics of κ

Limits

When obtained agreement equals chance agreement, $\kappa = 0$. Greater than chance agreement leads to positive values of κ, less than chance agreement leads to negative values. The upper limit of κ is +1.00, occurring when (and only when) there is perfect agreement between the judges. For perfect agreement there is a necessary condition that the p_{iA}'s equal the p_{iB}'s (i varies from 1 to k). This makes good sense, since inequalities here automatically force disagreement (see *Maximum* κ below).

The lower limit of κ is more complicated, since it depends on the marginal distributions. Let us define r_m as the product-moment correlation coefficient between the p_{iA}'s and the p_{iB}'s. It is readily demonstrable that when $r_m = 0$, $p_c = 1/k$. Substituting this value and $p_o = 0$ in Equation 1:

$$\kappa_{1_o} = -\frac{1}{k - 1} \tag{3}$$

where κ_{1_o} = lower limit of κ when $r_m = 0$.

When one (or both) of the marginal distributions is rectangular ($p_{iA} = p_{2A} = \ldots = p_{kA}$), r_m will perforce equal zero, although r_m can, of course, be zero under other circumstances.

When r_m is negative (unlikely in practice), $p_c < 1/k$. With $p_o = 0$, from Equation 1:

$$\kappa_{1_-} > -\frac{1}{k - 1} \tag{4}$$

where κ_{1_-} = lower limit of κ when $r_m < 0$.

In the limit, as k increases, r_m approaches -1, and the variances of the marginal proportions increase, κ_{1_-} approaches 0.

When r_m is positive (the usual situation encountered in practice), $p_c > 1/k$. With $p_o = 0$, from Equation 1:

$$\kappa_{1_+} < -\frac{1}{k - 1} \tag{5}$$

where κ_{1_+} = lower limit of κ when $r_m > 0$.

In the limit, as k decreases to 2, r_m approaches +1, and the variances of the marginal proportions increase, κ_{1+} approaches −1. Note that when $k = 2$ and $r_m > 0$, it is not possible for p_o to equal zero; it must take some (positive) value, therefore $-1 < \kappa_{1+} < 0$ under this circumstance.

Since κ is used as a measure of *agreement*, the complexities of its lower limit are of primarily academic interest. It is of importance that its upper limit be 1.00. If it is less than zero (i.e., if the observed agreement is less than expected by chance), it is likely to be of no further practical interest.

Maximum κ

It has already been noted that κ can only reach +1.00 when the off-diagonal (disagreement) cells are zero. This in turn demands that $p_{iA} = p_{iB}$, i.e., the marginals must be identical. This is as it should be, since discrepancies between judges in their distribution of units into categories by its very nature constitutes disagreement. It is of some interest to determine in any reliability study the maximum value of κ set by the marginal distributions. It is

$$\kappa_M = \frac{p_{oM} - p_e}{1 - p_e} \tag{6}$$

where p_{oM} is found by pairing the p_{iA} and p_{iB} values, selecting the smaller of each pair, and summing the k values. (For Table 1, p_{oM} $= .40 + .30 + .20 = .90$, therefore $\kappa_M = .85$).

Thus, κ_M is the maximum value of κ permitted by the marginals and $1 - \kappa_M$ represents the proportion of the possibilities of agreement (chance excluded) which cannot be achieved as a consequence of differing marginals. The latter quantity can serve as an indicator to the investigator of the fuzziness of his category boundaries, and may be reduced and his boundaries sharpened by further training of his judges.

It may occasionally occur that the setting of the interjudge agreement may impose some inherent difference in category widths between judges. For example, Judge A may be a clinical psychologist working with projective techniques who may make the diagnosis of schizophrenia more frequently than Judge B, a psychiatrist working with screening interview material. Under such special circumstances, it is meaningful to raise the question, "How much

the marginally permitted agreement is present?" This question may be answered by computing the ratio κ/κ_M. In most applications, however, the question and its answer are of sharply limited relevance, since disagreement which is forced by marginal disagreement has the same negative consequences as disagreement not so forced—in short, it is disagreement.

κ and ϕ

The reader will have noted a similarity in the preceding section with ϕ, the product-moment correlation for the dichotomous case. It, too, has a maximum value determined in much the same way by unequal marginals, and it, too, can be "corrected" by division by its maximum under similarly restricted situations.

A further relationship between κ applied to a dichotomy and ϕ is of interest. It is fairly simply proven that when $p_{1A} = p_{1B}$ (i.e., when the marginals are the same for the judges), $\kappa = \phi$. When $p_{1A} \cong p_{1B}$ within .10 — .20, κ will be within a few hundredths of ϕ, closer as κ increases. More generally, the closeness of the approximation of κ to ϕ is a direct function of the discrepancy between the arithmetic and geometric means of p_{1A} and p_{2A}, which is at a minimum at $p = .50$. Finally, when $p_{1A} \neq p_{1B}$, $\kappa < \phi$.

The identity of the two coefficients under the condition of equal marginals tells us something about ϕ, namely, that under the conditions stated, it is interpretable as the proportion of agreement after allowance for chance.

Sampling Characteristics

An approximation to the standard error of κ is given by

$$\sigma_\kappa = \sqrt{\frac{p_o(1 - p_o)}{N(1 - p_e)^2}}, \tag{7}$$

or, in terms of frequencies,

$$\sigma_\kappa = \sqrt{\frac{f_o(N - f_o)}{N(N - f_e)^2}} = \frac{\sqrt{f_o(1 - f_o/N)}}{N - f_e} \tag{8}$$

The formula is an approximation since it treats p_e as a constant and treats p_o as if it were the population value. It should be adequate, nevertheless, since ordinarily p_e will not vary greatly relative to κ, particularly with N large (i.e., $\geqq 100$). With N large, too,

the sampling distribution of κ will approximate normality so that confidence limits can be set in the usual way:

$$95\% \text{ confidence limits} = \kappa \pm 1.96\,\sigma_\kappa$$

$$99\% \text{ confidence limits} = \kappa \pm 2.58\,\sigma_\kappa$$

Tests of the significance of the difference between two *independent* κ's can be performed by evaluating the normal curve deviate:

$$z = \frac{\kappa_1 - \kappa_2}{\sqrt{\sigma^2_{\kappa_1} + \sigma^2_{\kappa_2}}} \tag{9}$$

To test an obtained κ for significance, i.e., to test the null hypothesis that it arose in sampling from a population of units for which $\kappa_p = 0$, we substitute $p_c = p_o$ in Equation 7 and get:

$$\sigma_{\kappa_0} = \sqrt{\frac{p_c}{N(1 - p_c)}}, \tag{10}$$

or in terms of frequencies,

$$\sigma_{\kappa_0} = \sqrt{\frac{f_c}{N(N - f_c)}}. \tag{11}$$

The significance is determined by dividing κ by σ_{κ_0} and referring the resulting critical ratio to the normal curve. It needs pointing out that it is generally of as little value to test κ for significance as it is for any other reliability coefficient—to know merely that κ is beyond chance is trivial since one usually expects much more than this in the way of reliability in psychological measurement. It may, however, serve as a minimum demand in some applications.

Illustrative Example

Table 2 presents an agreement matrix in terms of both proportions and frequencies for the purpose of illustration. Chance expectancies are given only for the cells in the agreement diagonal since the other values are immaterial.

With $\kappa = .492$, we see that just under half of the joint judgments are agreements (with chance excluded). The marginals are such that κ_M is only .831, therefore a substantial part of the disagreements is a consequence of marginal discrepancies. It is estimated that the chances are 95% that the population value of κ falls between .384 and .600. Finally, the obtained κ value is highly si

Illustrative Agreement Matrix

a) Proportions

		Judge A			
	Category	1	2	3	p_{iB}
Judge B	1	.44(.30)*	.07	.09	.60
	2	.05	.20(.09)	.05	.30
	3	.01	.03	.06(.02)	.10
	p_{iA}	.50	.30	.20	$\sum p_i = 1.00$

$p_o = .44 + .20 + .06 = .70$

$p_c = .30 + .09 + .02 = .41$

$\kappa = \dfrac{.70 - .41}{1 - .41} = .492$ (Eq. 1)

$\kappa_M = \dfrac{(.50 + .30 + .10) - .41}{1 - .41} = .831$ (Eq. 6)

$\sigma_\kappa = \sqrt{\dfrac{.70\,(1 - .70)}{200\,(1 - .41)^2}} = .055$ (Eq. 7)

95% confidence limits = $.492 \pm 1.96\,(.055) = .384 \leftrightarrow .600$

$\sigma_{\kappa_0} = \sqrt{\dfrac{.41}{200\,(1 - .41)}} = .059$ (Eq. 10)

$z = \dfrac{.492}{.059} = 8.34$; κ significant at $P < .001$

b) Frequencies

		Judge A			
	Category	1	2	3	f_{iB}
Judge B	1	88(60)*	14	18	120
	2	10	40(18)	10	60
	3	2	6	12(4)	20
	f_{iA}	100	60	40	$N = 200$

$f_o = 88 + 40 + 12 = 140$

$f_c = 60 + 18 + 4 = 82$

$\kappa = \dfrac{140 - 82}{200 - 82} = .492$ (Eq. 2)

$\sigma_\kappa = \sqrt{\dfrac{140\,(200 - 140)}{200\,(200 - 82)^2}} = .055$ (Eq. 8)

$\sigma_{\kappa_0} = \sqrt{\dfrac{82}{200\,(200 - 82)}} = .059$ (Eq. 11)

* Chance expectancy

nificant, the chances being far less than 1 in 1000 that the population value of κ is zero.

Summary

A coefficient of interjudge agreement for nominal scales, $\kappa = \frac{p_o - p_c}{1 - p_c}$, is presented. It is directly interpretable as the proportion of joint judgments in which there is agreement, after chance agreement is excluded. Its upper limit is +1.00, and its lower limit falls between zero and −1.00, depending on the distribution of judgments by the two judges.

The maximum value which κ can take for any given problem is given, and the implications of this value to the question of agreement discussed. An interesting characteristic of κ is its identity with ϕ in the dichotomous case when the judges give the same marginal distributions.

Finally, its standard error and techniques for estimation and hypothesis testing are presented.

REFERENCES

Goodman, L. A. and Kruskal, W. H. "Measures of Association for Cross Classifications." *Journal of the American Statistical Association,* XLIX (1954), 732–764.

Guilford, J. P. *Fundamental Statistics in Psychology and Education.* (Second Edition) New York: McGraw-Hill, 1950.

Guttman, L. "An Outline of the Statistical Theory of Prediction." In P. Horst (Ed.), *The Prediction of Personal Adjustment.* New York: Social Science Research Council, 1941.

Scott, W. A. "Reliability of Content Analysis: The Case of Nominal Scale Coding." *Public Opinion Quarterly,* XIX (1955), 321–325.

Stevens, S. S. "Mathematics, Measurement and Psychophysics." In S. S. Stevens (Ed.), *Handbook of Experimental Psychology.* New York: John Wiley & Sons, 1951.

Stevens, S. S. "Problems and Methods of Psychophysics." *Psychological Bulletin,* LV (1958), 177–196.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XX, No. 1, 1960

SOCIAL CLASS, SEX, AND RESPONSE TO A FIVE-PART PERSONALITY INVENTORY

S. B. G. EYSENCK
Institute of Psychiatry, University of London[1]

PERSONALITY inventories have usually been administered to one of three main groups of subjects, namely college students, psychiatric patients, and working class applicants for certain types of jobs in which questionnaires were being used as selection tests. Relatively little is known about the relationship between responses to personality inventory questions and social class when nation-wide quota samples are under investigation. Similarly, and less excusably, little is known about the relationship between such responses and the sex of the subject. The work of Eysenck (1958) has suggested that according to their questionnaire responses women are somewhat more neurotic than men, that men are somewhat more extraverted than women, and that working class subjects are slightly more neurotic than middle class subjects.

These results were reported in connection with standardization studies of the short and long scales, respectively, of the Maudsley Personality Inventory, a measure designed to provide a numerical system of a subject's neuroticism and extraversion (Eysenck, 1956; Eysenck, 1958–59; Eysenck, in press). The relationships with sex and class were established on the basis of analysis of variance and comparisons of mean scores of different groups. The present study, while also making use of the short scale of the Maudsley Personality Inventory, is in addition using three further short scales relating to the personality traits of rigidity, emotionality, and nervousness; the items were selected and adapted from the scales published by Guilford (1939), Guilford (1939), and Evans and McConnell (1941). The actual items used, the traits they are supposed to meas-

[1] The writer is indebted to the Research Fund Committee of the Bethlem Royal and Maudsley Hospitals for financial support.