Journal of Behavioral Assessment, Vol. 3, No. 1, 1981

Measures of Interobserver Agreement: Calculation Formulas and Distribution Effects

Alvin Enis House,[1,2] Betty J. House,[1] and Martha B. Campbell[1]

Accepted September 17, 1980

Seventeen measures of association for observer reliability (interobserver agreement) are reviewed and computational formulas are given in a common notational system. An empirical comparison of 10 of these measures is made over a range of potential reliability check results. The effects on percentage and correlational measures of occurrence frequency, error frequency, and error distribution are examined. The question of which is the "best" measure of interobserver agreement is discussed in terms of critical issues to be considered

KEY WORDS: interobserver agreement; observer reliability; measures of association; naturalistic observation; interval-by-interval coding systems.

INTRODUCTION

Comparisons of reliability measures are difficult for the behavioral investigator due to two factors. First, most original articles use notational systems unique to the author, which makes direct comparisons difficult. For example, Kendall and Stuart (1961, p. 539) discuss a coefficient of association, Q, defined by the equation

$$Q = \frac{ad - bc}{ad + bc} \qquad \frac{nD}{ad + bc},$$

[1]Psychology Department, Illinois State University, Normal, Illinois 61761.
[2]To whom correspondence should be addressed.

0164-0305/81/0300-0037$03.00/0

whereas Fleiss (1973, p. 45) reviews a measure of association, Q, defined as

$$Q = \frac{\Omega A - \Omega \bar{A}}{\Omega A + \Omega \bar{A}}.$$

The two formulas are exactly equivalent, as might be inferred from the reference of both chapters to Yule (1900), but it would require a translation from one notation system to the other for the reader to convince himself that this is so. A more extreme illustration of this potential difficulty occurred recently. Yelton *et al.* (1977) published an article proposing a relatively complex formula that yielded a probability estimate of the observed number of agreements (or more) occurring between two observers by chance. Hartman (1979) pointed out that, when transformed, this measure turns out to be equivalent to the Fisher Exact Probability Test, an equivalence obviously missed by the original reviewers. A final example of this same difficulty: Sloat (1978) demonstrated that the "weighted-means" formula proposed by Farkas (1978) could be simplified to the weighted total percentage agreement formula (see below).

Determining the equivalence of complex formulas through transformation may be difficult enough when they are represented in the same notation system. When the reader must cope with different symbols, often different breakdowns of the raw concordance data, and different transformation of formulas, the nonmathematician is tempted to give up. Unfortunately, few general review papers are available and none within the behavioral literature. Hartmann (1977) considers only kappa, phi, agreement percentage, and occurrence percentage among the available interval (or "trial") measures. The same is true of Kent and Foster (1977). Hawkins and Dotson (1975) consider four different percentage formulas (agreement, occurrence, nonoccurrence, and averaged occurrence/nonoccurrence) but no correlational measures. Therefore the interested party must obtain original articles and then translate formulas into a uniform base for comparison.

The second difficulty is that even though the computational formula completely states and determines the resulting numerical measures, for the mathematically unsophisticated the practical consequences of different interpretations are not always apparent. As an example, the reader might consider the formulas for kappa, phi, and lambda in Fig. 2. The computational formulas of kappa and phi yield highly similar results, identical under a wide range of specificable conditions ($B = C$). Both differ from lambda, which computationally yields a dissimilar result, identical with kappa and phi under only one limited set of conditions ($B = C = 0$). The question is: How many readers could see that this was so when they

considered the formulas? Most behavioral investigators do not have sufficient experience with these types of statistics to predict accurately the statistic distribution from the formula alone.

The present paper first briefly reviews and gives computational formulas for the most commonly discussed measure of observer reliability calculated for an interval coding system. An observational interval is a time segment, usually brief, during which a behavior is scored as either occurring or not occurring according to the coding system in use. Intervals may be continuous (one interval immediately following the previous interval) or time sampled (uncoded periods between intervals). The measures presented here are calculated on an interval-by-interval (or "trial") basis: Observers must agree on not only something happening, but also when it happened. Measures which do not analyze the data on a trial basis are labeled "total" or "session"—the measures consider only the total recorded frequencies by observers in calculating their concordance. Session measures for interval or frequency coding are not presented in the present paper; the interested reader could consult Hartman (1977). Second, an empirical comparison of the measures over a range of potential outcome values from a reliability check is given. Third, the question of the "best" measure is considered briefly.

COMPUTATIONAL FORMULAS

Measures of association between two or more sets of data represent a topic that has long been of interest to mathematicians. Yule's early articles (1900, 1912) discussed several possible approaches to the problem, and in the past seven decades a large number of statistics have been examined. The present review is limited to measures which have been used or proposed for use with observational data or which appear to have clear potential for consideration. Measures of primarily theoretical interest or measures which have been generally concluded to be inadequate as a measure of association (chi square, for instance) are not considered. An extensive literature exists on this topic and the interested reader is strongly recommended to sample it directly; basic references include Everitt (1977), Haggard (1958), Fleiss (1971, 1973, 1975), Goodman and Kruskal (1954, 1959, 1963, 1972), Kendall and Stuart (1961), and Sarndal (1974).

One common fashion of summarizing the incidence of agreement and disagreement between two observers is in terms of a two-by-two table such as that shown in Fig. 1. The different cell totals provide basic information on the patterns of coding of the two observers across observation intervals for a given behavior category. Cells A and D represent the number of agreements on the behavior's occurrence and nonoccurrence, respectively.

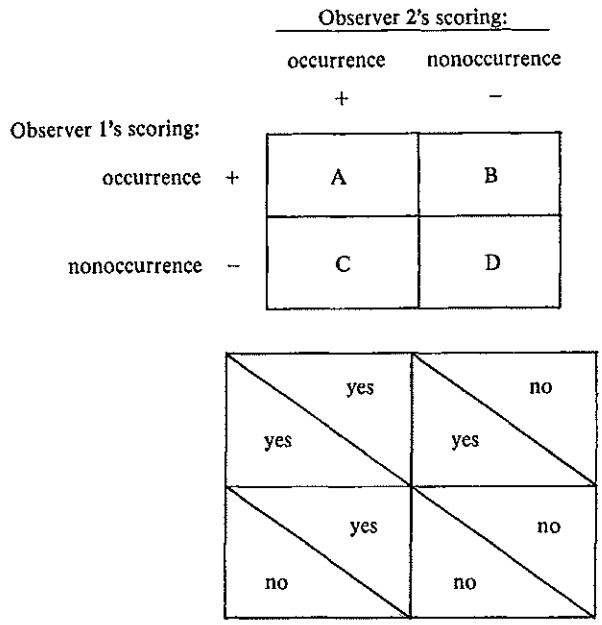

Fig. 1. Two by two matrix notation of interobserver agreement for a behavior category.

A = number of agreements on occurrence.
B = number of disagreements where observer 1 coded the category and observer 2 did not.
C = number of disagreements where observer 1 did not code the category and observer 2 did.
D = number of agreements on nonoccurrence.
N = A + B + C + D = number of intervals coded in observation.
O_1 = A + B = frequency of occurrence recorded by observer 1.
O_2 = A + C = frequency of occurrence recorded by observer 2.

Cells B and C represent the number of disagreements when one observer coded the behavior and the other observer did not. The sum of all cells, N (A + B + C + D), gives the total number of intervals coded in the observation. Two additional values of interest are the occurrence frequencies of the behavior recorded by Observer 1; O_1 (A + B), and by Observer 2, O_2 (A + C).

All computational formulas given in Fig. 2 are expressed in terms of the four cell values A, B, C, and D. It is hoped that this will allow for greater understanding when comparing alternative procedures. It should be noted that several of the formulas can be written more concisely by using the additional notation N, O_1, and O_2.

The first six formulas are all variations of agreement percentage. Percentage methods yield result varying between 0 (no association) and 100% agreement (complete association). There are actually only two basic formulas or approaches to percentage agreement: total agreement and specified event agreement (either occurrence or nonoccurrence agreement percentage). The first analysis considers both of the two possible types of agreements—agreement that something did happen and agreement that something did not happen. The second approach restricts interest, at least in initial analysis, to one of these types of agreement. The other formulas all represent manipulations of these two basic approaches, usually by assigning weights to certain cell values, i.e., considering some cells more important than others (in the extreme case by assigning a weight of zero to a cell and eliminating its influence on the association measure).[3]

The first formula given in Fig. 2 is the basic total agreement percentage: the sum of agreement divided by the sum of agreement plus disagreement. Although defensible as a measure of association for behaviors with moderate (40 to 60%) occurrence (Hawkins and Dotson, 1975; Kratochwill and Wetzel, 1977), total agreement methods are seldom seen as acceptable. The difficulty is that with either very frequent or very infrequent events, the probable large number of agreements on nonoccurrence or occurrence, respectfully, will produce high ("inflated") agreement values almost regardless of how "together" the observers are (e.g., Hawkins and Dotson, 1975; Hartmann, 1977; Repp *et al.*, 1976). For example, if both record the behavior as occurring 10% of the time ($O_1 = 10$, $O_2 = 10$) then the minimum total agreement percentage between them will be 80% agreement, even if they never agreed on the behavior happening (A = 0, B = 10, C = 10, D = 80, T% = 80).

The alternative percentage approach is to restrict attention to the type of event (either occurrence or nonoccurrence) of predetermined interest to the experimenter or of lower frequency. The occurrence and nonoccurrence

[3]The reader has probably realized that all percentage formulas could be viewed as variations of a total agreement formula, with specified event formulas simply being the special cases where one of the agreement cells is assigned the weight of zero. Although less parsimonious, the distinction made in the text is heuristically useful because it conforms to the common choices discussed in the behavioral literature regarding percentage measures of observer reliability.

Total Agreement Percentage: "percentage agreement" (Hartman, 1977); "agreement percentage" (Kent and Foster, 1977); "interval by interval" by interval" (Hawkins and Dotson, 1975); "point-by-point" (Kelly, 1977).

$$T\% = \frac{A + D}{A + B + C + D} \times 100. \quad (1)$$

Occurrence Agreement Percentage: "effective percentage agreement" (Hartman, 1977); "scored interval" (Hawkins and Dotson, 1975); "type I reliability" (Wahler *et al.*, 1976).

$$O\% = \frac{A}{A + B + C} \times 100. \quad (2)$$

Nonoccurrence Agreement Percentage: "unscored interval" (Hawkins and Dotson, 1975).

$$NO\% = \frac{D}{B + C + D} \times 100. \quad (3)$$

Averaged Occurrence/Nonoccurrence Agreement Percentage: "mean agreement" (Hawkins and Dotson, 1975).

$$M\% = \frac{\left[\frac{A}{(A + B + C)}\right] + \left[\frac{D}{(B + C + D)}\right]}{2} \times 100. \quad (4)$$

Weighted Total Agreement Percentage: "Weighted mean average" (Farkas, 1978); "Weighted average" (Sloat, 1978).

$$W\% = \frac{A + D}{A + D + 2(B + C)} \times 100. \quad (5)$$

Weighted Occurrence Agreement Percentage.

$$C\% = \frac{A}{A + B} \times 100. \quad (6)$$

Interobserver Agreement (Clement, 1976).

$$IOA = \left[\left(\frac{A}{A + B}\right)\left(1 - \frac{A + B}{A + B + C + D}\right)\right]\left[\left(\frac{D}{C + D}\right)\left(1 - \frac{C + D}{A + B + C + D}\right)\right]. \quad (7)$$

Weighted Agreement (Harris and Lahey, 1978; Taylor, 1980).

$$WA = \left\{\left(\frac{A}{A+B+C}\right)\left[\frac{B+C+2D}{2(A+B+C+D)}\right]\right\} + \left\{\left(\frac{D}{B+C+D}\right)\left[\frac{2A+B+C}{2(A+B+C+D)}\right]\right\} \quad (8)$$

Fig. 2. Computational formulas for measures of interobserver agreement.

Kappa (Cohen, 1960).

$$K = \frac{(A + D) - \left\{ \frac{[(A + B)(A + C)}{(A + B - C + D)]} + \frac{[(C + D)(B + D)}{(A + B + C + D)]} \right\}}{(A + B + C + D) - \left\{ \frac{[(A + B)(A + C)}{(A + B + C + D)} + \frac{[(C + D)(B + D)}{(A + B + C + D)]} \right\}} \quad (9)$$

Weighted Kappa (Cohen, 1968).

$$K_W = 1 - \left\{ \frac{(A+B+C+D)\,[W_1(A) + W_2(B) + W_3(C) + W_4(D)]}{W_1(A+B)(A+C) + W_2(A+B)(B+D) + W_3(A+C)(C+D) + W_4(C+D)(B+D)} \right\}$$

W_n = assigned weight. (10)

Occurrence Kappa (Kent and Foster, 1977).

$$K_0 = \frac{(A) - \left[\frac{(A+B)(A+C)}{(A+B+C)}\right]}{(A+B+C) - \left[\frac{(A+B+)(A+C)}{(A+B+C)}\right]}. \quad (11)$$

Nonoccurrence Kappa (Kent and Foster, 1977).

$$K_{NO} = \frac{(D) - \left[\frac{(C+D)(B+D)}{(B+C+D)}\right]}{(B+C+D) - \left[\frac{(C+D)(B+D)}{(B+C+D)}\right]}. \quad (12)$$

Phi: product–moment correlation for dichotomous data; *"V"* (Yule, 1912).

$$\varphi = \frac{(AD) - (BC)}{[(A+B)(C+D)(A+C)(B+D)]}. \quad (13)$$

Q (Yule, 1900, 1912).

$$Q = \frac{AD - BC}{AD + BC}. \quad (14)$$

r_{11} (Maxwell and Pilliner, 1968; Fleiss, 1975).

$$r_{11} = \frac{2(AD - BC)}{(A+B)(C+D) + (A+C)(B+C)}. \quad (15)$$

Fig. 2. Continued.

G Index (Holley and Guildord, 1964; Janson and Vegelius, 1979).

$$G = \frac{(A+D) - (B+C)}{A + B + C + D}. \tag{16}$$

Lambda (Goodman and Kruskal, 1954).

$$\lambda = \frac{2A - B - C}{2A + B + C}. \tag{17}$$

Pi [(Scott, 1955); "r_*" (Fleiss, 1965)].

$$\pi = \frac{4AD - B^2 - 2BC - C^2}{(2A+B+C)(2D-B+C)}. \tag{18}$$

Probability of A or More Agreements (Yelton *et al.*, 1977).

$$p_{A \geq A_0} = \sum_{Z=A}^{(A+C)} \left\{ \left[\frac{(A+C)!}{Z!\,(A+C-Z)!} \right] \left[\frac{(A+B)!}{(A+B-Z)!} \right] \left[\frac{(C+D)!}{(D-A+Z)!} \right] \left[\frac{(B+D)!}{(A+B+C+D)!} \right] \right\}$$

*$A+C \leq A+B$. (19)

Averaged Agreement Matrix

$$X_{2\times 2} = \begin{array}{|c|c|} \hline \bar{X}_A & \bar{X}_B \\ \hline \bar{X}_C & \bar{X}_D \\ \hline \end{array}$$

$\bar{X}_A$ = Average of cell A value across observations.
$\bar{X}_B$ = Average of cell B value across observations.
$\bar{X}_C$ = Average of cell C value across observations.
$\bar{X}_D$ = Average of cell D value across observations. (20)

Fig. 2. Continued.

agreement percentage formulas are more "conservative" than the total agreement percentage, which is described as too "liberal." The problem with specified event approaches is that at very low frequencies they may be too conservative. All percentage measures are in part determined by the size of the "base": the number of relevant intervals. Suppose that of 100 intervals, both observers score a behavior 9 times but are discordant on 2 occassions (A/B/C/D = 8/1/1/90). Their occurrence agreement percentage is 80% agreement. Now imagine that only 4 instances were coded, but again they

were discordant on 2 (A/B/C/D = 2/1/1/96). Now their occurrence agreement percentage is 50%. Yet in both cases the observers made only 2 errors. Part of the problem is that at the lower frequency, each error is worth 25% of the total possible (note that the total agreement percentage is the same in both cases: 98%). Since often the behaviors of particular interest to a behavior therapist may have a low absolute frequency, the investigator may be unduly penalized by occurrence agreement formulas.

As a compromise Hawkins and Dotson (1975) proposed averaging occurrence and nonoccurrence agreement percentages for a mean agreement. Farkas (1978) pointed out a possible bias inherent in this, due to there typically being an unequal number of intervals involved in the two component measures, and proposed as a solution assigning each agreement score as weight based on the number of intervals on which it was computed. Sloat (1978) demonstrated that this procedure led to a formula equivalent to simply doubling the weight given to the error cells' contribution to the total agreement percentage, i.e., adding an additional penalty factor for mistakes.

The sixth formula in Fig. 2 is one example of a weighted occurrence agreement percentage measure. The rationale for its use is typically that one observer is assigned a "criterion" status and the only errors processed in analysis are when the "regular" observer fails to detect/record a behavior coded by the criterion observer. The choice of who is regular and who is criterion is determined by the researcher and not the obtained data; formula (6) could also be written as

$$C\% = \frac{A}{A + C} \times 100.$$

There seems little justification for ignoring the occurrence of one type of observer disagreement, but this method does have an important advantage: It is quicker and easier to tally manually than either total agreement or occurrence agreement percentage. Only information on two of the cells from the fourfold table are extracted from the raw observations. The person computing reliability has to scan only one observation until he finds an occurrence indicated and then check to see if the same interval is marked in the second observation. However limited, this measure is used in some projects, probably because of the ease and speed with which it can be calculated.

All percentage agreement formulas have the general form of

$$\frac{\text{agreements}}{\text{agreements} + \text{disagreements}} \times 100 = \%\ \text{agreement}.$$

Differences in formulas grow out of differences in interpretation of what should be counted as an "agreement" and what should be counted as a "disagreement." Major advantages of percentage measures are their ease of calculation and interpretation. Although the question of "how much" agreement is necessary, good, or reasonable is not resolved (or a priori resolvable) for any of the measures in Fig. 2, there seems some consensus among behavioral investigators that average agreement at or above 70% is necessary, above 80% is adequate, and above 90% is good. It is, of course, much more difficult to achieve these standards using occurrence than total agreement formulas. The most often cited limitation of percentage formulas is the absence of any correction factor for differences in chance agreement at different levels of response frequency. The issue here is not simply one of chance agreement between observers—some degree of spurious concordance is always a possibility (and probably a reality). The problem is that at different response frequencies, the probability of chance agreement differs, and to fail to correct for this introduces a bias into the measure of association. One answer has been to use correlational measures of association which are usually described as "controlling" for chance agreement. The nature of this control is to devalue the concordance achieved at high or low rates as contrasted with "moderate" rates (approximately 50% occurrence).

Clement's formula [Eq. (7), Fig. 2] is an alternative approach to the problem of differences in chance agreement due to unequal numbers of scored and unscored intervals, which assumes that chance agreement is least likely with respect to the least frequent event (occurrence or nonoccurrence). Clement presumes that one observer will be a "criterion," although this can be a random choice. His measure of interobserver agreement varies between 0 (when both A and D are zero) and + 1 (when both B and C are zero). Harris and Lahey (1978) have offered a modified version of Clement's formulas which weights the two agreement components to correct for "chance" agreement.

Kappa, pi, and lambda are all derived from the same theoretical formula (Cohen, 1960):

$$\text{coefficient} = \frac{P_o - P_e}{1 - P_e},$$

where P_o is the observer proportion of agreement and P_e is the expected proportion (chance agreement). The differences in computational formula result from differences in the interpretation of the expected proportion term, P_e. Weighted kappa is a derivative formula which allows the

investigator to state in advance that some types of error are more important than others. Weighted kappa was developed for use with ratings which had more than two possible values and has not been applied to observational data. Occurrence and nonoccurrence kappas are derivative formulas given by Kent and Foster (1977). An extensive literature has grown on the use and properties of kappa and weighted kappa (Cicchetti and Fleiss, 1977; Everitt, 1968; Fleiss and Cicchetti, 1978; Fleiss *et al.*, 1969, 1979; Landis and Koch 1977a,b). Fleiss (1965) derived a measure of association r^*, equivalent to pi, and recommended that its use be limited to those cases where the marginal distributions are evenly balanced. Lambda considers only agreements and disagreements on occurrence, as does occurrence kappa. Obviously the formula could also be calculated for nonoccurrence. Kappa, lambda, and phi all vary between $+1$ (complete association) and -1 (complete dissociation). Also, all three measures reach unity only if both error cell frequencies equal zero, a situation Kendall and Stuart (1961) referred to as "absolute" association (all O_1's are O_2's and all O_2's are O_1's), in contrast to measures which would reach unity if only one error went to zero ("complete" association).

Yule defined a coefficient of association Q in 1900, given in Eq. (14) (Fig. 2), which also varies between $+1$ and -1. Q shows only complete association, going to unity if either the B or the C error cell frequency is zero. Another measure of association given by Yule in 1912 also varies between $+1$ and -1 but shows absolute association. This measure V is equivalent to phi: the product-moment correlation for dichotomous data (Nunnally, 1967). Phi can be interpreted in a variety of ways, including as a derivation of chi square corrected for sample size. A maximum value of unity is possible only if the marginal distributions are equal, if the unequal maximum value is less than $+1$. Fleiss (1973) gave detailed consideration to the use of phi as a measure of association. He also suggested that values less than 0.30 or 0.35 may be taken as indicating no more than trivial association.

Maxwell and Pilliner (1968) developed a measure of association from an analysis of variance model and then extended it to the case of dichotomously scored data. The computational formula is given in Eq. (15) (Fig. 2). The resulting statistic also varies between $+1$ and -1 and shows absolute association. Holley and Guilford (1964) propose the G index varies between $+1$ and -1 and shows absolute association. Janson and Vegelius (1979) classify both phi and the G index as E coefficients.

Another approach to a measure of interobserver association is the computation of the probability of obtaining the observed number of agreements or more by chance alone. Yelton *et al.* proposed a statistic in 1977 based on this approach. Their computational formula, given in Eq.

(19) (Fig. 2), is equivalent to Fisher's Exact Probability Test (Siegel, 1956). The computational formula given is a factorial expansion and somewhat unrealistic to calculate by hand, even with the aid of an electronic calculator.

DISTRIBUTION EFFECTS

The most direct manner for the mathematically unsophisticated investigator to gain an understanding of the different measures of interobserver agreement is to calculate values for a number of possible observation outcomes. Doing so will illustrate how each measure varies as a function of the occurrence frequency, error frequency, error distribution, and nonoccurrence frequency. Occurrence frequency simply refers to how often a behavior is occurring. Behavioral investigators may refer to high rate behaviors as occurring in excess of 80% of coded intervals, low rate as categories occurring in less than 20% of coded intervals, and medium rate as occurring in between 40 and 60% of coded intervals. Obviously there are not absolute standards as to what is a "high" or a "low" rate of an action. In terms of the fourfold table of interobserver outcome, the two sums A + B and A + C represent the estimates of behavior frequency by observer 1 and 2, respectively.

Error frequency refers to the number of disagreements between the two observers: the sum of cells B and C. Error balance refers to the ratio between the two error cells: B/C (or C/B). If errors are "balanced" between the two observers, approximately equal numbers of errors are being made by both and the error cells will be approximately equal (B = C). Skewed error distributions are when either of the error cells is disproportionately larger than the other (B > C or B < C). The significance is that skewed balances more likely suggest a type of error is being made that might bias results (Hartmann, 1977; House, 1980). Nonocurrence frequency is exactly analogous to occurrence frequency: the occasions a behavior is not recorded (B + D) and (C + D).

In order to contrast several common measures of interobserver agreement across a wide range of possible outcomes of a reliability check, the data in Table I were assembled. For a hypothetical observation generating 100 intervals of data, a range of possible outcomes of observer concordance was generated informally, dictated by several considerations: sampling a wide range of occurrence frequency, nonoccurrence frequency, error frequency, and error balance. The relatively few actual examples of formula scores in the literature are universally based on only a small number of possible matrices of interobserver agreement. In case the reader is wondering why the authors did not simply compute all outcome

possibilities, it should be pointed out that the number of unique outcomes of a two by two matrix grows as an accelerating function of the sum of the cells. For $N = 100$, there are 176,851 possible solutions (R. Forcade and J. M. Cook, personal communication). Of this rather large universe of outcomes, 100 were sampled,[4] and for each of these the following measures of observer concordance were calculated: kappa, phi, lambda, IOA, occurrence percentage, nonoccurrence percentage, mean percentage, total percentage, weighted occurrence percentage, and weighted total percentage.

The reader is encouraged to examine Table I and note how the measure values change as a function of the outcome. Several trends become apparent with some study. For kappa, phi, and IOA, two influences can be noted. The values of these measures are greatly influenced by the overall level of occurrence of the behavior. These four measures take on their greatest values for a given number of observer disagreement when agreements are equally divided between occurrences and nonoccurrences. For instance, kappa for the outcome 70-6-4-20 is 0.73, while kappa for 50-6-4-40 is 0.80. Corresponding values for phi are 0.73 and 0.80, and for IOA 0.85 and 0.90. The effect of correcting for chance agreement is simply to make it more difficult to obtain high levels of observer agreement at either high or low frequencies.

M% also assumes higher values at roughly balanced rates of occurrence and nonoccurrence. T% and W% assume exactly the same value for any fixed number of agreements regardless of how the agreements are divided among occurrences and nonoccurrences, O%, NO%, C%, and lambda fluctuate nonsymmetrically since they utilize data from only three cells.

A more disturbing trend with kappa, phi, and IOA is that these measures assume greater values when the balance of observer disagreements is very skewed. For instance, kappa for 70-6-4-20 is 0.73, while kappa for 70-10-0-20 is 0.74. Corresponding values for phi are 0.73 and 0.76, and for IOA 0.85 and 0.98. This trend toward high values with skewed error distributions is of concern because the investigator would usually be more concerned over this type of observer error, as it is more likely to reflect serious bias in coding (House, 1980).

Lambda has the advantage of assuming higher values for a given number of observer disagreements when the two error cells are approximately balanced. O%, T%, M%, NO%, and W% assume the same value regardless of the error cell balance. C% increases with the error cell balance in the examples in Table I, but would decrease if based on cells A, C, and D rather than A, B, and D.

[4]An expanded version of Table I with values for 210 reliability check outcomes is available from the authors.

Table I. Interobserver Agreement Calculated by 11 Procedures for 210 Possible Results of a Reliability Check on 100 Intervals

Agreement–disagreement matrix				Interobserver agreement									
A	B	C	D[a]	κ	ϕ	λ	IOA	O%	NO%	M%	T%	W%	C%
100	0	0	0	–[b]	–	1.00	–	100	–	–	100	100	100
90	5	5	0	−0.05	−0.05	0.89	0.05	90	0	45	90	82	95
90	10	0	0	−0.00	–	0.89	–	90	0	45	90	82	90
90	8	2	0	−0.03	−0.04	0.89	0.02	90	0	45	90	82	92
80	10	10	0	−0.11	−0.11	0.78	0.09	80	0	40	80	67	89
80	20	0	0	−0.00	–	0.78	–	80	0	40	80	67	80
30	70	0	0	−0.00	–	−0.08	–	30	0	15	30	18	30
99	0	0	1	1.00	1.00	1.00	100	100	100	100	100	100	100
98	1	0	1	0.66	0.70	0.99	1.00	99	50	75	99	98	99
90	5	4	1	0.13	0.14	0.90	0.24	91	10	91	91	83	95
90	9	0	1	0.17	0.30	0.90	1.00	91	10	51	91	83	91
90	8	1	1	0.15	0.20	0.90	0.51	91	10	51	91	83	92
80	10	9	1	−0.01	−0.01	0.79	0.18	81	5	43	81	68	89
30	35	34	1	−0.51	−0.51	−0.07	0.18	30	1	16	31	18	86
80	0	0	20	1.00	1.00	1.00	1.00	100	100	100	100	100	100
75	5	0	20	0.86	0.87	0.94	0.99	94	80	87	95	90	94
75	3	2	20	0.86	0.86	0.94	0.92	94	80	87	95	90	96
70	5	5	20	0.73	0.73	0.87	0.83	88	67	78	90	82	93
70	10	0	20	0.74	0.76	0.87	0.98	88	67	78	90	82	88
60	10	10	20	0.52	0.52	0.71	0.73	75	50	63	80	67	86
60	15	5	20	0.53	0.54	0.71	0.80	75	50	63	80	67	80
60	20	0	20	0.55	0.61	0.71	0.95	75	50	63	80	67	75
50	15	15	20	0.34	0.34	0.54	0.64	63	40	52	70	54	77
50	25	5	20	0.37	0.41	0.54	0.77	63	40	52	70	54	67
50	30	0	20	0.40	0.50	0.54	0.93	63	40	52	70	54	63
30	25	25	20	−0.01	−0.01	0.09	0.49	38	29	34	50	33	55
30	50	0	20	0.19	0.33	0.09	0.88	38	29	34	50	33	38
70	0	0	30	1.00	1.00	1.00	1.00	100	100	100	100	100	100
60	5	5	30	0.78	0.78	0.85	0.88	86	75	81	90	82	92

50	10	10	30	0.58	0.58	0.67	0.78	71	60	66	80	67	83
40	15	15	30	0.39	0.39	0.45	0.70	57	50	54	70	54	73
50	0	0	50	1.00	1.00	1.00	1.00	100	100	100	100	100	100
45	5	0	50	0.90	0.90	0.89	0.95	90	91	91	95	90	90
45	3	2	50	0.90	0.90	0.89	0.95	90	91	91	95	90	94
40	5	5	50	0.80	0.80	0.78	0.90	80	83	82	90	82	89
30	15	5	50	0.59	0.60	0.50	0.78	60	71	66	80	67	67
30	10	10	50	0.58	0.58	0.50	0.78	60	71	66	80	67	75
20	15	15	50	0.34	0.34	0.14	0.64	40	63	52	70	54	57
20	30	0	50	0.40	0.50	0.14	0.70	40	63	52	70	54	40
40	0	0	60	1.00	1.00	1.00	1.00	100	100	100	100	100	100
39	1	0	60	0.98	0.98	0.97	0.99	98	98	98	99	98	98
35	5	0	60	0.89	0.90	0.87	0.93	88	92	90	95	90	88
30	5	5	60	0.78	0.78	0.71	0.88	75	86	81	90	82	86
30	9	1	60	0.78	0.79	0.71	0.85	75	86	81	90	82	77
30	10	0	60	0.78	0.80	0.71	0.85	75	86	81	90	82	75
20	10	10	60	0.52	0.52	0.33	0.73	50	75	63	80	67	67
10	15	15	60	0.20	0.20	−0.20	0.50	25	67	46	70	54	40
5	20	15	60	−0.00	−0.00	−0.56	0.35	13	63	38	65	48	20
0	20	20	60	−0.25	−0.25	−1.00	0.15	0	60	30	60	43	0
20	0	0	80	1.00	1.00	1.00	1.00	100	100	100	100	100	100
19	1	0	80	0.97	0.97	0.95	0.96	95	99	97	99	98	95
18	1	1	80	0.94	0.94	0.89	0.96	90	98	94	98	96	95
16	2	2	80	0.86	0.86	0.78	0.91	80	95	88	96	92	89
14	3	3	80	0.79	0.79	0.65	0.84	70	93	82	94	89	82
12	4	4	80	0.70	0.70	0.50	0.78	60	91	76	92	85	75
10	5	5	80	0.61	0.61	0.33	0.71	50	89	70	90	82	67
10	0	0	90	1.00	1.00	1.00	1.00	100	100	100	100	100	100
8	1	1	90	0.88	0.88	0.78	0.90	80	98	89	98	96	89
6	2	2	90	0.73	0.73	0.50	0.77	60	96	78	96	92	75
5	3	2	90	0.64	0.64	0.33	0.66	50	95	73	95	90	63
5	5	0	90	0.64	0.69	0.33	0.55	50	95	73	95	90	50
5	4	1	90	0.64	0.66	0.33	0.60	50	95	73	95	90	56
4	3	3	90	0.54	0.54	0.14	0.60	40	94	67	94	89	57
2	8	0	90	0.31	0.43	−0.33	0.28	20	92	56	92	85	20

Table I. Continued.

Agreement–disagreement matrix				Interobserver agreement									
A	B	C	D[a]	κ	ϕ	λ	IOA	O%	NO%	M%	T%	W%	C%
0	10	0	90	−0.00	–	−1.00	0.10	0	90	45	90	82	0
6	0	0	94	1.00	1.00	1.00	1.00	100	100	100	100	100	100
5	1	0	94	0.90	0.91	0.82	0.84	83	99	91	99	98	83
4	1	1	94	0.79	0.79	0.60	0.81	67	98	83	98	96	80
4	2	0	94	0.79	0.81	0.60	0.69	67	98	83	98	96	97
3	2	1	94	0.65	0.66	0.33	0.62	50	97	74	97	94	60
3	3	0	94	0.65	0.70	0.33	0.53	50	97	74	97	94	50
2	2	2	94	0.48	0.48	−0.00	0.52	33	96	65	96	92	50
2	3	1	94	0.48	0.50	−0.00	0.43	33	96	65	96	92	40
2	4	0	94	0.48	0.57	−0.00	0.37	33	96	65	96	92	33
1	3	2	94	0.26	0.26	−0.43	0.28	17	95	56	95	90	25
1	4	1	94	0.26	0.29	−0.43	0.24	17	95	56	95	90	20
1	5	0	94	0.27	0.40	−0.43	0.22	17	95	56	95	90	17
0	3	3	94	−0.03	−0.03	−1.00	0.03	0	94	47	94	89	0
0	4	2	94	−0.03	−0.03	−1.00	0.04	0	94	47	94	89	0
0	5	1	94	−0.02	−0.02	−1.00	0.05	0	94	47	94	89	0
0	6	0	94	−0.00	–	−1.00	0.06	0	94	47	94	89	0
3	0	0	97	1.00	1.00	1.00	1.00	100	100	100	100	100	100
2	1	0	97	0.80	0.81	0.60	0.68	67	99	83	99	98	67
1	1	1	97	0.49	0.49	−0.00	0.51	33	98	66	98	96	50
1	2	0	97	0.49	0.57	−0.00	0.35	33	98	66	98	96	33
0	3	0	97	−0.00	–	−1.00	0.03	0	97	49	97	94	0
0	2	1	97	−0.01	−0.01	−1.00	0.02	0	97	49	97	94	0
1	0	0	99	1.00	1.00	1.00	1.00	100	100	100	100	100	100
0	1	0	99	−0.00	–	−1.00	0.01	0	99	50	99	98	0
0	0	0	100	–	–	–	–	0	100	50	100	100	–

[a]Sum of A + B + C + D = 100. Matrices are constructed so that B ≥ C.
[b]Calculation involved an uninterpretable term, usually division by zero.

Kappa, phi, IOA, T%, M%, and W% are symmetrical with respect to agreements on occurrence versus nonoccurrence: The value for 20-5-5-70 will be the same as the value for 70-5-5-20. Lambda, O%, C%, and NO% are asymmetrical—the first three being solely functions of cells A, B, and C, and NO% solely a function of B, C, and D.

All measures have an upper limit of 1 (or 100) which they will assume whenever B and C are zero. Two measures, IOA and C%, register complete as well as absolute association. IOA goes to unity whenever cell C is zero regardless of the values assumed by A, B, and D. C% would go to 100 whenever cell B was zero (not reflected in Table I).

THE CHOICE OF AN INDEX OF ASSOCIATION

The practical question facing the behavioral observer is which measure of observer concordance to use. Hartman (1977) maintains that correlational measures are preferable to percentage measures on the basis of their mathematical properties (correction for chance agreement) and their adaptability to generalizability theory (Cronbach *et al.*, 1972). Fleiss (1975) reviews several correlational measures and recommends kappa and r_{11} over other possibilities (including phi) on the basis of their statistical properties. Another set of considerations concerns such practical questions as ease of computation. Percentage measures are generally easier to calculate than correlations. This may be an important factor to the average investigator. House *et al.* (1980) demonstrated that occurrence percentage is calculated not only more rapidly than kappa and phi, but also more accurately by students using electronic calculators. Unfortunately, except for a few technical notes (e.g., Knapp and Loveless, 1976; McQueen, 1975), few authors have addressed practical procedural issues affecting coding and reliability.

Rather than champion a candidate for "best measure," we would like to pose several considerations about which an investigator might think. This failure to voice a preference is based on a conviction that there is no "best measure." All measures have to be used with certain cautions in mind, hence the following general considerations:

(1) All the measures reviewed in this paper are ultimately based on the same data: the fourfold tables of observer concordance. Although never proposed to our knowledge, an interesting possibility would be reporting the mean values of cells A, B, C, and D across observations [formula (20)].

(2) The decision of which measure to use should not be based solely on statistical factors. Mathematics is a tool for accomplishing the task at hand; sometimes we become confused as to the relative importance of the

tool and the task—a type of intellectual snobbery. Simple measures that are easily understood and interpreted have much to recommend them as long as their limitations are borne in mind. One problem with complex measures is that we do not understand them well enough to recognize their limitations. In the words of Baer (1977): "just because it's reliable doesn't mean you can use it."

(3) The decision should not be unduly influenced by appeals to as yet unsubstantiated theoretical conceptions. For instance, generalizability theory may be a valuable conceptual framework for viewing the problems of observer accuracy and validity, as Hartmann (1977) and others (Jones *et al.*, 1975; Mitchell, 1979) have argued. But to date this value is more potential than realized. An equally compelling case could be made for viewing observer reliability within the frame of human vigilance research (Mackie, 1977) where percentage measures of agreement are typical. The point is that neither theoretical position has a clear empirically demonstrated utility for observational research.

(4) There are issues beyond the choice of a measure that merit thought. Seldom are agreement data on individual observations reported; more typically a summary statistic is given for a series of observations—mean percentage agreement, median kappa values, etc. The choice of summary procedure (mean versus median, for instance) may affect the reported results as much as the measure chosen. Also, unfortunately, data are often reported collapsed across observation categories as well as across observations. A single summary statistic is reported. This sacrifices a great deal of information, may obscure areas of greater or lesser uncertainty, and reduces the reader's evaluation to an oversimplified "good enough/not good enough" discrimination. The computational method is only one of several factors affecting reliability (cf. House and House, 1979; Kaydin, 1977).

(5) The problems of measuring observer agreement are unlikely to disappear. As long as observer records remain a mainstay of behavioral data, there will be the necessity to grapple with the issues raised here. With certain problems investigators have been very ingenious in devising permanent product measures (Hughes *et al.*, 1978). Audiotape records that can be "observed" over and over until acceptable agreement is obtained have been explored as an alternative to direct observation (Christensen, 1979; Johnson and Bolstad, 1975; Johnson *et al.*, 1976). Despite these alternatives, it seems likely that direct behavioral observation will continue to be an important method.

(6) These difficulties should not obscure the tremendous value and utility of observational procedures. House (1978) has argued for recognition of the robustness of observational measures. Direct sampling of relevant

actions in natural settings is and will undoubtedly remain the major strength of behavioral assessment.

REFERENCES

Bear, D. M. Reviewer's comment: Just because it's reliable doesn't mean that you can use it. *Journal of Applied Behavior Analysis,* 1977, *10,* 117–119.

Christensen, A. Naturalistic observation of families: A system for random audio recording in the home. *Behavior Therapy,* 1979, *10,* 418–422.

Cicchetti, D. V., and Fleiss, J. L. Comparison of the null distributions of weighted kappa and the C ordinal statistic. *Applied Psychological Measurement,* 1977, *1,* 195–201.

Clement, P. W. A formula for computing inter-observer agreement *Psychological Reports,* 1976, *39,* 257–258.

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

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin,* 1968, *70,* 213–220.

Cronbach, L. J., Glaser, G. C., Nanda, H., and Rajaratnam, N. *The Dependability of Behavioral Measurements: Theory of General Profiles.* New York: Wiley, 1972.

Everitt, B. S. Moments of the statistics kappa and weighted kappa. *British Journal of Mathematical and Statistical Psychology,* 1968, *21,* 97–103.

Everitt, B. S. *The Analysis of Contingency Tables.* New York: Wiley, 1977.

Farkas, G. M. Correction for bias present in a method of calculating interobserver agreement. *Journal of Applies Behavior Analysis,* 1978, *11,* 188.

Fleiss, J. L. Estimating the accuracy of dichotomous judgments. *Psychometrika,* 1965, *30,* 469–479.

Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psyhcological Bulletin,* 1971, *76,* 378–382.

Fleiss, J. L. *Statistical Methods for Rates and Proportions.* New York: Wiley, 1973.

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

Fleiss, J. L., and Cicchetti, D. V. Inference about weighted kappa in the nonnull case. *Applied Psychological Measurement,* 1978, *2,* 113–117.

Fleiss, J. L., Cohen, J., and Everitt, B. S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin,* 1969, *72,* 323–327.

Fleiss, J. L., Nee, J. C. M., and Landis, J. R. Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin,* 1979, *86,* 974–977.

Goodman, L. A., and Kruskal, W. H. Measures of association for cross-classification, Part I. *Journal of the American Statistical Association,* 1954, *49,* 732–764.

Goodman, L. A., and Kruskal, W. H. Measures of association for cross-classifications, Part II. *Journal of the American Statistical Association,* 1959, *54,* 123-163.

Goodman, L. A., and Kruskal, W. H. Measures of association for cross-classifications, Part III, Approximate sampling theory. *Journal of the American Statistical Association,* 1963, *58,* 310-364.

Goodman, L. A., and Kruskal, W. H. Measures of association for cross-classifications, Part IV, Simplification of asymptotic variances. *Journal of the American Statistical Association,* 1972, *67,* 415-421.

Haggard, E. A. *Intraclass Correlation and the Analysis of Variance.* New York: Dryden, 1958.

Harris, F. C., and Lahey, B. B. A method for combining occurrence and nonoccurrence agreement scores. *Journal of Applied Behavior Analysis,* 1978, *11,* 523–527.

Hartmann, D. P. Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis,* 1977, *10,* 103–116.

Hartmann, D. P. A Note on reliability: Old wine in a new bottle. *Journal of Applied Behavior Analysis,* 1979, *12,* 298.

Hawkins, R. P., and Dotson, V. A. Reliability scores that delude: An Alice in Wonderland trip through the misleading characteristics of interobserver agreement scores in interval recording. In E. Ramp and G. Semb (Eds.), *Behavior Analysis: Areas of Research and Application.* Englewood Cliffs, N.J.: Prentice-Hall, 1975, 539-376.

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

House, A. E. Naturalistic observation: Formal and informal difficulties. *Child Study Journal,* 1978, *8,* 17-28.

House, A. E. Detecting bias in observational data. *Behavioral Assessment,* 1980, *2,* 29-31.

House, B. J., and House, A. E. Frequency, complexity, and clarity as covariates of observer reliability. *Journal of Behavioral Assessment,* 1979, *1,* 149-165.

House, A. E., Farber, J., and Nier, L. L. Accuracy and speed of reliability calculation using different measures of interobserver agreement. Paper presented in poster session, Association for Advancement of Behavior Therapy, New York, November 1980.

Hughes, H., Hughes, A., and Dial, H. A behavioral seal: An apparatus alternative to behavioral observation of thumbsucking. *Behavioral Research Method and Instrumentation,* 1978, *10,* 460-461.

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

Johnson, S. C. Hierarchical clustering schemes. *Psychometrika,* 1967, *32,* 241-254.

Johnson, S. M., and Bolstad, O. D. Rectivity to home observation: A comparison of audio recorded behavior with observers present or absent. *Journal of Applied Behavior Analysis,* 1975, *8,* 181-185.

Johnson, S. M., Christensen, A., and Bellamy, G. T. Evaluation of family intervention through unobtrusive audio recordings: Experiences in "bugging children." *Journal of Applied Behavior Analysis,* 1976, *9,* 213-219.

Jones, R. R., Reid, J. B., and Patterson, G. R. Naturalistic observation in clinical assessment. In P. McReynolds (Ed.), *Advances in Psychological Assessment, Vol. 3,* San Francisco: Jossey-Bass, 1975, pp. 42-95.

Kaydin, A. E. Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis,* 1977, *10,* 141-150.

Kelly, M. B. A review of the observational data-collection and reliability procedures reported in *The Journal of Applied Behavior Analysis. Journal of Applied Behavior Analysis.* 1977, *10,* 97-101.

Kendall, M. G., and Stuart, A. *The Advanced Theory of Statistics. Vol. 2. Inference and Relationship.* New York: Hafner

Kent, R. N., and Foster, S. L. Direct observational procedures: Methodological issues in naturalistic settings. In A. R. Ciminero, K. S. Calhoun, and H. E. Adams (Eds.), *Handbood of Behavioral Assessment.* New York: John Wiley & Sons, 1977, pp. 279-328.

Knapp, T. J., and Loveless, S. E. A simple procedure for determining reliability scores in interval recording. *Behavior Therapy,* 1976, *7,* 557-558.

Kratochwill, T. R., and Wetzel, R. J. Observer agreement, credibility, and judgment: Some considerations in presenting observer agreement data. *Journal of Applied Behavior Analysis,* 1977, *10,* 133-139.

Landis, J. R., and Koch, G. G. A one-way components of variance model for categorical data. *Biometrics,* 1977a, *33,* 671-679.

Landis, J. R., and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics,* 1977b, *33,* 159-174.

Mackie, R. R. (Ed.). *Vigilance.* New York: Plenum, 1977.

Maxwell, A. E., and Pilliner, A. E. G. Deriving coefficients of reliability and agreement for ratings. *British Journal of Mathematical and Statistical Psychology,* 1968, *21,* 105-116.

McQueen, W. M. A simple device for improving inter-rater reliability. *Behavior Therapy,* 1975, *6,* 128-129.

Mitchell, S. K. Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin,* 1979, *86,* 376-390.

Nunnally, J. C. *Psychometric Theory.* New York: McGraw-Hill, 1967.

Repp, A. C., Deitz, D. E., Boles, S. M., Deitz, S. M., and Repp, C. F. Differences among common methods for calculating interobserver agreement. *Journal of Applied Behavior Analysis,* 1976, *9,* 109–113

Sarndal, C. E. A comparative study of association measures. *Psychometrika,* 1974, *39,* 165–187.

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

Siegel, S. *Nonparametric Statistics.* New York: McGraw–Hill, 1956.

Sloat, K. M. C. A comment on "Correction for bias present in a method of calculating interobserver agreement," Unpublished paper. Kamehomeha Early Education Program, 1978.

Taylor, D. R. An expedient method for calculating the Harris and Lahey weighted agreement formula. *The Behavior Therapist,* 1980, *3,* 3.

Wahler, R. G., House, A. E., and Stambaugh, E. E. *Ecological Assessment of Child Behavior.* New York: Pergamon, 1976.

Yelton, A. R., Wildman, B. G., and Erickson, M. T. A probability-based formula for calculating interobserver agreement. *Journal of Applied Behavior Analysis,* 1977, *10,* 127–131.

Yule, G. U. On the association of attributes in statistics. *Philosophical Transactions of the Royal Society, Series A,* 1900, *194,* 257.

Yule, G. U. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society,* 1912, *75,* 579–642.