JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1977, 10, 103-116 NUMBER 1 (SPRING) 1977

CONSIDERATIONS IN THE CHOICE OF INTEROBSERVER RELIABILITY ESTIMATES

DONALD P. HARTMANN[1]

UNIVERSITY OF UTAH

Two types of interobserver reliability values may be needed in treatment studies in which observers constitute the primary data-acquisition system: trial reilability and the reliability of the composite unit or score which is subsequently analyzed, *e.g.*, daily or weekly session totals. Two approaches to determining interobserver reliability are described: percentage agreement and "correlational" measures of reliability. The interpretation of these estimates, factors affecting their magnitude, and the advantages and limitations of each approach are presented.

DESCRIPTORS: observational technology, reliability, validity, statistics, recording and measurement techniques, Cohen's kappa, generalizability theory, measurement theory, Spearman-Brown prophesy formula, correlational measures

Reliability is a necessary though not sufficient condition for validity. Thus, the likelihood of detecting a performance difference between treatment conditions, is a direct function of the reliability of the measures used. The importance of this relationship between validity and reliability seems to be underestimated by contemporary investigators employing human observers. Although observational technologies play an important role in current behavior therapies, the use of psychometrically sound measurement principles has not kept pace with the application of these observational techniques.

The reliability of observational data may be examined from a number of perspectives, such as interval consistency and stability over time and across situations and behavior. But the principal concern of many researchers is the reliability of their basic data-acquisition system—a human observer-recorder; that is, the "degree to which they can generalized from a given set of ratings to those that other raters might make" (Wiggins, 1973, p. 285). The present paper describes the principal methods of estimating the reliability of the human observer.

[1]I wish to thank Irwin Altman, David Born, Eugene Garcia, Donna Gelfand, Gary Gregor, Emily Herbert, Charles Turner, and my graduate students for their critical reading of earlier drafts of this manuscript. Reprints may be obtained from the author, Department of Psychology, The University of Utah, Salt Lake City, Utah 84112.

Research Design

The designs used in applied behavioral research involve some combination of observers, trials, sessions, subjects, conditions, and behaviors—what Cronbach, Glaser, Nanda, and Rajaratnam (1972) called design facets. For present purposes, it will be assumed that the facets include observers, trials (observation periods within a session), and sessions—a typical combination in studies employing a single-subject design. While this paper is specifically directed at reliability assessment in single-subject studies, much of the material is generalizable to group research by substituting "subjects" for "sessions" or "trials".

The researcher typically has three recording procedures from which to choose: event recording, which provides measures of the frequency of occurrence of the target behavior; duration recording, which provides measures of the duration of occurrence of the target behavior; and occurrence-nonoccurrence (interval) recording, which can provide estimates of both frequency and duration of the target behavior. The pur-

poses of a study, the nature of the dependent and independent variables, and other specific aspects of the experimental situation, determine which among these three techniques is the most advantageous (see Gelfand and Hartmann, 1975).

Issues Determining Nature of Reliability Assessments

Three decisions determine the nature of reliability assessment procedures (Johnson and Bolstad, 1973, pp. 10-17).

1. The first decision requires specification of the *score unit* on which reliability will be assessed. If the score unit is a narrowly defined specific target behavior such as soiling, then reliability with which soiling is scored is determined. On the other hand, the primary dependent variable may be a composite score, such as inappropriate behavior composed of a number of narrowly defined and specifically scored behaviors such as noncompliance, hitting, and stealing. In this case, reliability is appropriately determined for the composite score. If, as is often the case, the separate component behaviors making up a composite score are also analyzed, then reliability analyses should be conducted on each component behavior as well. As a general principle, reliability assessments should be conducted on the unit of behavior subject to visual or statistical analysis.

2. The second decision requires specification of the *time span* over which scores will be summed for purposes of reliability assessment. Reliability could be calculated on the scores in each of the recording intervals or trials for a session in which two or more observers independently collect data. This level of reliability will be referred to as trial reliability. Reliability also can be determined for longer temporal units of behavior, such as for condition scores, or more commonly for session scores. Reliability assessed on session scores (*e.g.,* the sum of scores for the multiple trials within a session) will be referred to as session reliability. Again, reliability should be assessed for at least the time span over which data are compiled for purposes of analyses.

3. The final decision concerns the type of summary reliability statistic applied to the data, which is the primary topic of the remainder of this paper.

Two general approaches to determining interobserver reliability, percentage agreement, and "correlational" reliability are employed in applied behavioral studies. Each technique has advantages and limitations, which are described separately for session scores and trial scores.

Session Reliability

Visual or statistical analysis in applied behavioral research is almost uniformly conducted on session scores, whether these be session means, session totals, or some session-based rate measure. These session scores are typically obtained in one of the following ways. First, session scores may be obtained by summing across the multiple recording intervals for which occurrence-nonoccurrence data are tabulated on the target behavior. For example, Kazdin and Klock (1973) obtained session scores for student "attentive" behavior, by summing scores over the 20 brief (15-sec) recording periods conducted each day. Second, session scores may be obtained by summing frequency or duration scores across multiple discrete trials, such as might be obtained if latency of response was timed for each of 30 daily requests made of a child. And finally, session scores may be obtained by summing either frequency or duration data across an entire observation period in which either time sampling or continuous observation of the target behavior has been conducted. For example, the number of helping incidents might be tallied during a 20-min free-play period. In all of these cases, the scores (whether means, rates/time, or totals) vary from zero to some positive value and can be considered to have the properties of a ratio scale.

Reliability assessments conducted on session scores indicate the degree to which we can generalize from the session scores obtained from one observer to those session scores that another observer might obtain. Depending on the index of session reliability used, session reliability also

can help determine whether variability within a condition (*e.g.*, baseline) is due to observer error. Session reliability is particularly important when session scores are the dependent variable and when statistical analyses such as correlational analyses are performed on session scores (*c.f.* Wahler, 1975).

Session scores are usefully viewed as composite or pooled scores, as they are composed of scores obtained from multiple real or hypothetical trials within a session. Like all composite scores, they reflect the characteristics of their component scores. Consequently, session scores reflect the adequacy of behavior definitions and the thoroughness of observer training in using these definitions. Also, because they are composite scores they will typically be more reliable than their component scores (Hartmann, 1976). While session scores frequently constitute the primary dependent variable in applied behavioral studies, they are rarely subject to formal reliability analysis. The reliability methods that might be applied to session scores include percentage agreement statistics and product-moment or intraclass correlations.

Percentage Agreement Reliability

Some investigators have described the interobserver reliability of session scores by dividing the smaller of the two scores obtained for a session by the larger, and multiplying this ratio by 100. For example, Schmidt and Ulrich (1969) calculated the interobserver agreement of decibel readings from the dial of a sound-level meter by calculating the mean decibel reading over 20-min observation periods for each observer. The smaller score was then divided by the larger score and multiplied by 100 to obtain a percentage agreement value. Similar procedures could be applied to the session scores given in Table 1. For these scores, the percentage agreement values range from $[100 \times (2/3)] = 67\%$ to $[100 \times (2/2)] = 100\%$. This method of calculating percentage agreement has its appeal primarily in its computational and interpretative simplicity and its utility in assessing whether the difference between session scores represents real change or merely observer error.[2] Unfortunately, it has a variety of limitations, including the lack of both a meaningful lower bound of acceptability and a value indicating no agreement. The value of this percentage agreement statistic also is heavily dependent on the specific rate of the behavior for the session in which it is calculated. When the behavior occurs at high rates, higher percentage agreement values result.

Table 1

Occurrence-Nonoccurrence data used to illustrate the determination of trial and session reliability.

	Trial (Observation Interval)					*Session*
Session	*1*	*2*	*3*	*4*	*5*	*Total*
OBSERVER 1						
1	0	1	0	0	1	2
2	1	0	0	1	1	3
3	1	1	1	0	0	3
OBSERVER 2						
1	0	0	1	0	1	2
2	1	0	0	1	0	2
3	1	1	1	1	0	4

A second percentage agreement statistic is the percentage of session scores for which the two observers completely agree—(number of sessions for which the two observers agree/number

[2]One must exercise some care in using session percentage agreement for this purpose. Assume, for example, that two session scores ($X_1 = 85$ and $X_2 = 118$) are compared to determine whether the difference of 33 points indicates a real (nonerror) change in the subject's performance. Also assume that 85% agreement was the minimum value obtained for all conditions, and that it was obtained for the two sessions in which Observer 1, the principal observer, obtained scores of 85 and 118 and Observer 2, the reliability checker, obtained scores of 100 and 100. The "confidence intervals" for the analyzed scores are 72 to 98 for X_1—$[85 \pm (1-0.85)85]$—and 100 to 136 for X_2—$[118 \pm (1-0.85)118]$. The use of these "confidence intervals" may result in the false conclusion that the two scores represent real change. A more conservative and hence preferable method of establishing "confidence intervals" is to use the value $X \pm [(1/0.85)-1]X$. This formula results in "confidence intervals" of 70 to 100 for X_1—$[85 \pm 15]$—and 97 to 139 for X_2—$[118 \pm 21]$. The use of these "confidence intervals" results in the correct conclusion that the difference between 85 and 118 could be due to observer error.

of sessions jointly observed) $\times$ 100. For the session data given in Table 1, this statistic has a value of $[100 \times (1/3)] = 33\%$. This statistic has very limited value as a measure of session reliability; it is not only extremely stringent in assessing agreement, but uses little of the information available in the data, *e.g.,* a difference of one between two observers' session totals is equivalent to a difference of 100, so far as this statistic is concerned.

Reliability Coefficient

The more traditional, although infrequently reported, measure of session interobserver reliability is the reliability coefficient r_{kk}. When calculated directly, this is simply the product-moment correlation based on the paired scores provided by the two observers for the sessions that are jointly observed. For the three pair of session scores given in Table 1, $r_{kk} = 0.50$ (though ordinarily r_{kk} would be based on a minimum of eight to 10 pairs of session scores). The coefficient typically ranges from 0.00 to +1.00. (Although the possible range of r_{kk} extends from -1.00 to $+1.00$, negative reliability coefficients are rare.) An $r_{kk} = 0.00$ indicates a lack of relationship between the two observers' ratings, whereas an $r_{kk} = +1.0$ indicates perfect agreement (in the sense of identical standard scores). The reliability coefficient calculated by the product-moment formula has precise mathematical interpretations: r_{kk} equals the proportion of total score variance not due to error and the degree of linear association between the two observers' data; r_{kk}^2 equals the proportion of variance of one observer's scores that is predictable from knowledge of the other observer's scores.

The advantages and limitations of the product-moment correlation are thoroughly described in standard texts on statistics and psychometric theory (*e.g.,* Gulliksen, 1950; Lord and Novick, 1968; McNemar, 1969; Nunnally, 1967); hence, they are described only briefly here. The advantages include the following:

First, r_{kk} indicates the degree of confidence that can be placed in the session scores, and the standard error of measurement—which is a function of r_{kk}—can be used to generate a confidence interval that indicates the smallest difference between session scores that can be interpreted meaningfully. For example, assume that $r_{kk} = 0.9$, the standard deviation of session scores = 10, and both observers have equal means for the jointly observed sessions. Any difference between two session scores, say the last day of baseline and the first day of treatment, greater than $9.0 = 2\,[2(10)^2(1 - r_{kk})]^{1/2}$ represents a real (nonerror) change in performance (McNemar, 1969, pp. 165-173).

Second, the reliability coefficient gives an accurate description of the degree of linear dependency or correlation in the observers' ratings. However, when r_{kk} is calculated by means of the product-moment formula, only random components contribute to error to reduce r_{kk}. Scores for two observers could differ by a constant across all sessions and r_{kk} could equal +1.00. If r_{kk} is calculated by means of the analysis of variance (the intraclass correlation coefficient), systematic error can also serve to lower r_{kk}, depending on the sources of variance included in the calculations (Winer, 1971, p. 283). Systematic errors should be taken into account in calculating r_{kk} if neither of the two observers functions as data collector for all sessions and some degree of observer bias or consistent error between observers is present.

Third, because of its extensive history in psychometric theory and applications, many of the properties of r_{kk} are well known. For example, r_{kk} can be tested for significance by means of the usual procedures used to test whether r departs significantly from zero. Similarly, observer bias, or the tendency of one observer to code more of the target behavior than the other observer, can be tested by the t-test of the difference between correlated scores.

The reliability coefficient also readily lends itself to potentially useful estimation functions in the preliminary stages of a study via the Spearman-Brown prophesy formula and the correction-for-attenuation formula. (See, for example,

Guilford and Fruchter, 1973, chapters 17 and 18.) For example, under particularly adverse observational conditions, such as might be experienced on a large and crowded playground, or for behaviors that for any reasons are difficult to discriminate, some investigators may find it more efficient to improve reliability by employing multiple observers whose scores will be pooled than to engage in lengthy observer training, an experimental analysis of observer behavior, or the purchase of costly recording equipment.[3] When such situations occur, r_{kk} (the session reliability for a single pair of observers) may be used in the Spearman-Brown prophesy formula for estimating the number of similarly trained observers whose scores could be pooled to achieve a specified degree of interobserver reliability. For example, if the session reliability for a single pair of observers is 0.5, pooled scores with a reliability of 0.8 could be achieved by using four observers [$4 = 0.8(1 - 0.5)/0.5(1 - 0.8)$]. If the number of observers is fixed, but two or more, this same formula may be used for estimating the interobserver reliability of the pooled observer scores. In the previous example, the reliability of the two observers' pooled scores is 0.67 or $2(0.5)/[1 + 1(0.5)]$. And a variant of the Spearman-Brown formula can be used for estimating the reliability of trial scores from knowledge of the reliability of session scores or *vice-versa.*

By correcting for attenuation, r_{kk} can estimate the number of observations required to detect a treatment effect or correlation of a specific magnitude or to determine the likelihood of detecting a treatment effect or correlation of a specific magnitude when the number of observations is fixed. Also, r_{kk} can estimate the magnitude of relationship between session scores and some other variable under conditions of improved reliability. For example, if the correlation between session performance scores and conditions of reinforcement is 0.4 with $r_{kk} = 0.36$, the correlation can be expected to increase to $0.6 = 0.4 (0.81)^{1/2}/(0.36)^{1/2}$ if r_{kk} is increased to 0.81.

The chief disadvantage of the reliability coefficient occurs in those rare situations when the variability of scores for one or both observers is zero, and r_{kk} is undefined. However, in such cases there is little basis for making any substantive comments concerning reliability. Additional problems occur in the interpretation of r_{kk} if observer errors are correlated, if the scatter-plot of the observers' ratings indicates nonlinear regression, and when the variability of ratings is either nonnormal or differ markedly across the score intervals (heteroscedasticity). Finally, the reliability coefficient, like any other correlational statistic is affected by the range of scores. If it can be assumed that the discrepancy between observers' ratings is independent of the score range, r_{kk} will increase directly as the range of scores increases.

Trial Reliability

Trial data suitable for reliability analysis in applied behavioral studies typically come in one of two forms: categorical data or occurrence-nonoccurrence ratings that take on values of zero or one and numerical data that take on a wider range of values. The former data stem from interval-recording procedure, in which each observer records the presence or absence of one or more target behaviors in brief, say 10-sec, recording intervals (see Table 1). The latter type of data results from duration and frequency recording procedures. Because the reliability procedures for these two types of data differ, they are described separately.

[3]Any mention of pooling observers' scores seems to be anathema to some applied researchers, perhaps because of their distaste for pooled data in group experimentation. Consequently, it might be worth recalling that most behavioral data involve pooling of some sort. For example, session performance scores, exam scores, and test scores are all pooled scores—and are more reliable than their components largely for the same reasons that pooled observers' scores are more reliable than the scores from a single observer. In the final analysis, the decision to pool or not to pool observers' scores is largely a pragmatic one. If pooling allow us to "get on" with the investigation of an important behavior, if it increases efficiency without jeopardizing rigor, then it should be seriously considered.

Estimates of trial reliability indicate the reliability of trial scores, whether they are duration, frequency, or occurrence-nonoccurrence scores. Reliability at this micro level of analysis primarily indicates the adequacy of the behavioral definitions, the thoroughness of observer training in the use of both these definitions and the observational hardware such as coding sheets, event recorders, and timers. Without a reasonable degree of interobserver reliability at the trial level, a study may not be interpretable because of the ambiguous meaning of the basic data. Trial reliability thus is important in most behavioral studies, but particularly so in those studies in which analyses are performed on trial data (*e.g.*, Patterson and Cobb, 1973).

Categorical Data

Categorical or occurrence-nonoccurrence data are usually summarized in a two-by-two or larger square table similar to that shown in Table 2. The letters A through D in that Table summarize the two kinds of agreements and two kinds of disagreements possible when scoring occurrence and nonoccurrence of a single target behavior. For example, the frequency in Cell B indicates the number of intervals for which both observers indicated occurrence for the target behavior. Each of the reliability statistics applicable to categorical data use the data in this summary table form. These methods include percentage agreement and related techniques including effective percentage agreement, and a group of correlational-like techniques including kappa (κ) and phi (ϕ).

Percentage agreement. Percentage agreement is by far the most commonly used statistic for summarizing two-by-two table data. With reference to Table 2, percentage agreement is given by the proportion of agreements $(B + C)/N$, multiplied by 100. For the data in Table 2, percentage agreement equals 70% or $100 \times (30 + 40)/100$. This value indicates the per cent of total observations the observers agreed. Percentage agreement ranges from 100%, in which case all entries in the summary table are agreements and $(B + C) = N$, to 0%, in which case all entries in the summary table are disagreements, and $(A + D) = N$.

Effective percentage agreement. If the primary focus of an experiment is directed toward occurrences of a behavior, the agreements contributed by Cell C (nonoccurrence of the target response rate by both observers) can be removed by calculating effective percentage agreement for occurrences (occurrence agreement), a statistic described by Jensen (1959). Such might be the case, for example, if two observers were coding the frequency of automobile accidents at a busy intersection in 15-min intervals. If accidents occurred at the rate of only one per 3-hr period, occurrence reliability, rather than percentage agreement reliability, might more adequately describe the reliability of their ratings because of the very large number of entries in Cell C.

Table 2

Two-by-Two Data Table Used in Summarizing Trial Reliability for Categorical Data

		Observer 2 (0)	*Observer 2* (1)	
Observer 1	Occurrence (1)	A 10	B 40	$A + B = 50$
	Nonoccurrence (0)	C 30	D 20	$C + D = 50$
		$A + C = 40$	$B + D = 60$	$N = 100$

Effective percentage agreement for occurrences is given by $[B/(A + B + D)] \times 100$. That is, percentage agreement is calculated on only those occasions in which either or both observers rate the target behavior as having occurred. For the data given in Table 2, effective percentage agreement for occurrences equals $[100 \times (40/70)]$ or 57%. Effective percentage agreement for occurrences indicates the percentage of these intervals in which both observers agreed that the target behavior occurred. Like percentage agreement, effective percentage agreement for occurrences ranges from 100% when the observers agree on all observed incidents of occurrence of the target behavior, to 0% when the observers disagree on all rated occurrences of the target behavior. With infrequently occurring behaviors, effective percentage agreement is not spuriously raised by the inclusion of Cell C frequencies and is a more sensitive measure of agreement for the occurrence category.

Effective percentage agreement also can be calculated on nonoccurrence, *i.e.*, $[C/(A + C + D)] \times 100$. For the data in Table 2, effective percentage agreement for nonoccurrences equals 50% or $[100 \times (30/60)]$. This statistic may be preferable when the focus is on nonoccurrence of a specific behavior. Its interpretation is analogous to the interpretation of effective percentage of agreement for occurrences.

Both effective percentage agreement statistics were designed to provide a more sensitive measure of observer reliability by excluding the contributions of the high-rate agreement cell (either Cell B or Cell C) whose agreements might be largely due to "chance" agreements. By "chance" agreements is meant the expected number or proportion of agreements that would be obtained when the observer's ratings were unrelated (independent). For example, for the data in Table 2, 30 of the 40 agreements in Cell B and 20 of the 30 agreements in Cell C would be expected if the observers' ratings were unrelated and Observer 1 rated the behavior as occurring $100 \times [(A + B)/N] = 50\%$ of the time, and Observer 2 rated the behavior as occurring $100 \times [(C + D)/N] = 60\%$ of the time. Whether these 50 expected agreements are chance or real agreements can only be determined with additional reliability assessments. In general, "chance" or expected agreements are totally dependent on the marginal values in the two-by-two summary table; that is, the values of $A + B$, $C + D$, $A + C$, and $B + D$. For any two-by-two table, the expected agreements are given by: $[(A + B)(B + D)/N] + [(C + D)(A + C)/N]$. Thus, if the number of recording intervals $(N) = 100$ and both observers rate the target behavior as occurring 90% of the time $[100 \times (A + B)/N = 100 \times (B + D)/N = 90\%]$, the expected number of agreements is 82, or a "chance" percentage agreement of 82%.[4] The next set of trial reliability statistics were specifically developed to handle the problem of expected agreements.

Correlational-like measures. The correlational-like measures of trial interobserver reliability include two somewhat different statistics, phi (ϕ) and kappa (κ). When the rate of occurrence of the target behavior is approximately equal for the two observers $[(A + B)/N \simeq (B + D)/N]$, these two statistics are nearly identical in value.[5] In most studies incorporating careful observer training, this requirement will be met, so the

[4]With $(A+B)/N = (B+D)/N$, the percentage of expected agreements is a curvilinear function of $(A+B)/N$. With $(A+B)/N = 0$ or 1.0, the percentage of expected agreements is 100%; when $(A+B)/N = 0.50$, the percentage of expected agreement is 50%. See Hartmann (Note 1) for further elaboration of this point.

[5]Phi, according to Cohen (1960) will estimate kappa within 0.02 of a point as long as $|(A+B)/N - (B+D)/N| < 0.20$. As the marginal frequency of occurrence for the two observers becomes more disparate, the difference between κ and ϕ increases with $\kappa < \phi$. Pi (π), initially described by Scott (1955), is a third correlational-like statistic sometimes used with categorical data; π like κ equals $(p_o - p_c)/(1 - p_c)$. Phi is identical to κ when $(A+B)/N = (B+C)/N$, but $\kappa > \pi$ when $(A+B)/N \neq (B+D)/N$ because of the slightly different manner of calculating p_c for κ and π. See Krippendorff (1970) and Fleiss (1975) for discussions of the comparative properties of ϕ, κ, and π, when used to index the reliability of categorical data.

two statistics can be used interchangeably. In cases where the two statistics differ in value, and hence in interpretation, the statistic associated with a specific interpretation will be indicated.

Kappa (κ), a statistic especially developed to measure the interobserver reliability of categorical data by Cohen (1960), is given by $(p_o - p_c)/(1 - p_c)$, where p_o is the proportion of observed agreements and p_c is the proportion of chance or expected agreements. For the data in Table 2, $p_o = (40 + 30)/100 = 0.70$, $p_c = (60 \times 50)/100^2 + (40 \times 50)/100^2 = 0.50$, and $\kappa = (0.70 - 0.50)/(1 - 0.50) = 0.40$. As can be seen from the numerator of the formula for kappa, $(p_o - p_c)$, the proportion of observer agreements is explicitly corrected for the proportion of chance or expected agreements. The denominator for kappa, $(1 - p_c)$ is similarly corrected for chance agreements. Thus, kappa indicates the proportion of agreements, corrected for chance agreements.

Kappa is at a maximum of +1.0 when no disagreements are present and both observers exhibit variation in the scoring categories. Kappa will equal zero when the proportion of chance agreements equals the number of observed agreements, and kappa will taken on negative values when the proportion of observed agreements is less than the proportion of chance agreements. The properties of kappa are extensively described by Cohen (1960, 1968) and Fleiss (1971, 1973).

Phi (ϕ), the product-moment correlation between two sets of dichotomous (yes-no or occurrence-nonoccurrence) data, is given by $(BC - AD)/[(A + B)(C + D)(A + C)(B + D)]^{1/2}$. For the data given in Table 2, $\phi = (1200 - 200)/[(50)(50)(60)(40)]^{1/2} = 0.41$. Phi ranges from −1.0 through 0.00 to +1.0. Phi equal to 0.00 indicates an absence of relationship between the two observers' ratings, and ϕ equal to +1.0 indicates complete agreement. Because ϕ is a product-moment correlation, the interpretations of r_{kk} in the section on session reliability also are appropriately made regarding ϕ. Under those conditions in which $\phi \simeq \kappa$, ϕ can also be interpreted as a corrected percentage agreement statistic. (See Haggard [1958] for a more extensive discussion of ϕ.)

Comparison of Measures of Trial Reliability

All the two-by-two table statistics share a number of advantages. For example, all require that each trial score be identified with one of the multiple time-locked recording intervals. Consequently, those intervals in which disagreements occur can be pinpointed and this information used for subsequent observer training.[6] Furthermore, all these measures are easily calculated, and with the possible exception of ϕ when it differs from κ, are readily interpreted. Finally, the two-by-two data can be readily tested for significance.[7]

There is one possible disadvantage to the use of kappa, ϕ, percentage agreement, and related statistics. All the two-by-two summary table reliability statistics completely confound random and systematic error. Lower reliability estimates (entries in Cells A and D in Table 2) are produced by random factors (such as periodic lapses of attention, temporary blocking of the observer's field of view, and occasional inclusive or exclusive coding errors). Systematic factors (such

[6]Time-locked data also can pose problems for trial reliability analysis if one of the observers "drops" an interval, so that all subsequent intervals are mismatched; thus, while Observer 1 is marking Interval 10, Observer 2 is marking Interval 11 with resulting high ratios of disagreement. Although data sets can often be realigned, lost intervals can pose a vexing problem. However, with the increased availability of inexpensive cassette recorders to signal observers by recorded numbers coordinated with the observation intervals (Whelan, Note 2), data sheets should rarely become unaligned.

[7]If it is desirable to test whether the two observers are agreeing more than would be expected on a chance basis, and the number of trials is large, χ^2 with 1 *df* can be determined and tested for significance. When the sample size is small and the assumptions of χ^2 cannot be met, Fisher's Exact Test (McNemar, 1969, p. 272 ff) can be used. To determine whether one observer is coding significantly more of the target behavior than is the other observer (observer bias), McNemar's test of the difference between correlated frequencies can be used (McNemar, 1969, p. 56).

as lack of agreement on the criteria for a response so that one observer consistently codes more of the behavior than a second observer) also contribute to lower reliability. If $A + B$ differs from $B + D$—as would be the case if one observer codes more of the target behavior than a second observer—these differences must be represented in either Cell A or Cell D, both of which are disagreement cells.[8]

The primary bases for choosing among the various two-by-two table reliability statistics include the accuracy with which they assess reliability, their relationship to formal reliability theory, the generality of their applicability and their relationship to session reliability.

Accuracy of reliability estimate. The primary concern with any estimate of reliability is that it reflects accurately and with minimum ambiguity the degree of reliability of the data assessed. As the data in Table 3 indicate, the measures of trial reliability differ markedly in value when applied to the same data, may change appreciably in value with changes in rate of the target behavior, and can provide substantially misleading esti-

[8]The increase in trial reliability expected when systematic factors are removed through retaining can be readily estimated. In the case of ϕ, $\phi/\phi_{max.}$ provides such an estimate (Guilford and Fruchter, 1973, pp. 306-310), and sensible estimates could readily be developed for the other two-by-two tables statistics. For example, $100 \times [B+(A \text{ or } D, \text{ whichever is smaller})]/N + 100 \times [C+(A \text{ or } D, \text{ whichever is smaller})]/N$, provides the maximum percentage of agreement possible with the marginal values fixed. Obtained percentage agreement divided by maximum percentage agreement then provides an estimate of the percentage of agreements to be expected when no systematic errors are present.

Table 3

Two-by-Two Table Data Used to Exemplify Limitations of Trial Reliability Statistics

Panel A

O_1 \ O_2	0	1	
1	5	5	10
0	85	5	90
	90	10	N=100

Percentage Agreement = 90%
Occurrence Agreement = 33%
Kappa = 0.44

Panel B_1

O_1 \ O_2	0	1	
1	5	45	50
0	45	5	50
	50	50	N=100

Percentage Agreement = 90%
Occurrence Agreement = 82%
Kappa = 0.80

Panel B_2

O_1 \ O_2	0	1	
1	25	25	50
0	25	25	50
	50	50	N=100

Percentage Agreement = 50%
Occurrence Agreement = 33%
Kappa = 0.00

Panel B_3

O_1 \ O_2	0	1	
1	14	36	50
0	36	14	50
	50	50	N=100

Percentage Agreement = 72%
Occurrence Agreement =56%
Kappa = 0.44

Note.—Panel A presents fictitious data obtained during the baseline phase of a treatment study; Panels B_1, B_2, and B_3 present data that might be obtained midway through a treatment phase. The underlined statistic in each of the B panels has the same value as in Panel A. The values of ϕ for these data are within 0.02 of the values of kappa.

mates of the reliability of trial scores. Consider Panel A of Table 3, which presents fictitious data obtained during the baseline phase with a problem such as correct pronunciation of *r* (a response that is originally produced infrequently but is to be accelerated as a result of treatment). The values of percentage agreement, effective percentage agreement for occurrence, and kappa are given below the table. Percentage agreement calculated on these data is a substantial 90%, whereas both occurrence agreement and kappa are noticeably lower. One might question which of these values most accurately represents the reliability of the data as the rate of the behavior increases with treatment. If the percentage agreement statistic does, then data similar to those presented in Panel B_1 would be obtained midway through treatment, as rate of target behavior increases. In this case, the three statistics are all substantial in magnitude and similar in value.

On the other hand, the agreements in the Panel A data might largely be due to the ease with which the observers rated nonoccurrences (and their difficulty in rating occurrences) coupled with the high rate of nonoccurrences of the target behavior. In this case, data similar to those presented in Panel B_2 of Table 3 would be obtained midway through treatment and the occurrence agreement statistic most adequately represents the reliability of trial scores. For Panel B_2 data, the three statistics are quite low in magnitude, but differ appreciably in value—from 50% for percentage agreement to 0.00 for kappa.

Finally, the agreements in the Panel A data might largely be due to the substantial number of chance or expected agreements produced by the highly divergent marginal values. [With rates of occurrence equal to 10 and of nonoccurrence equal to 90 for both observers, the expected number of agreements is $[90/100)(90/100) + (10/100)(10/100] \times 100 = 82$ even though the observers failed to attend to the target subject.] In this case, data similar to those presented in Panel B_3 in Table 3 would be obtained midway through treatment, and kappa most adequately represents the reliability of the trial scores. For the Panel B_3 data, the three statistics are intermediate in magnitude, but again differ appreciably in value; that is, from 72% for percentage agreement to 0.44 for kappa.

This analysis indicates the substantial differences in magnitude of observer agreement the three statistics yield and their ambiguous meaning when applied to a *single* set of two-by-two table data. It also highlights the necessity of inserting trial reliability probes (jointly observed trials) throughout a study. Only then can one obtain an accurate estimate of trial reliability. While all the measures provide varying results, they differ in the degree to which they might produce misleading optimism. Of the three applicable measures in this situation (nonoccurrence agreement is unlikely to be used here), percentage agreement consistently produces the highest index of agreement, with kappa and occurrence agreement yielding substantially lower values.[9] The tradition in science to accept conservative rather than liberal estimates suggests that percentage agreement is the least desirable of the three trial reliability statistics.

Relationship to formal reliability theory. Applied behavior analysts may at present find formal reliability theory surprisingly useful. For example, it may be useful to establish confidence intervals for a score, to determine the number of observers required to obtain pooled observer scores that attain some specified level of reliability, or to estimate the improvement in a correlation between trial scores and some other variable with improved interobserver reliability. These and other estimation functions described for r_{kk} in the section on session reliability can also be performed by ϕ for trial data. Comparable estimation formulas are not available for the percentage agreement statistics. In addition, both phi, and its close relation, kappa, are intraclass

[9]In three empirical studies comparing percentage agreement, the effective percentage agreement statistics and ϕ, Whelan (1974) found percentage agreement consistently higher than the remaining statistics. The intercorrelations between the measures ranged from 0.00 to 1.00, with ϕ generally yielding the lowest intercorrelation.

correlation coefficients (Fleiss, 1975), and as such are preferred measures of reliability (generalizability) in Cronbach's *et al.* (1972) liberalization of reliability theory. Kappa also has been shown to be related to statistics from information theory (Krippendorf, 1970).

Range of applicability. All of the reliability statistics discussed in this section are applicable to any mutually exclusive set of two-by-two table data. However, some applied researchers have displayed ambivalence concerning applications of the two effective agreement statistics. Deciding whether to employ occurrence reliability or nonoccurrence reliability evidently is not easy, particularly for behaviors whose rates vary from high to low during a study. This ambiguity of usage, together with the deceptive conclusions reached by choosing the wrong effective agreement statistic, represent then a potentially serious limitation of effective percentage agreement statistics. In general, effective percentage agreement statistics perhaps are best restricted to situations in which extreme rate behaviors do not undergo substantial changes in rate.

Occasionally, it may be desirable to estimate the interobserver reliability over a larger set of mutually exclusive target behaviors. Both percentage agreement and kappa are applicable here, as they can be used with any size square summary table. In addition, kappa has been developed for applications in which disagreements are differentially weighted (Cohen, 1968), and when multiple observers are used, not all of whom observe at the same time (Fleiss, 1971).

Continuous Data

Continuous trial data may result from either of two sources. First, duration or frequency data may be obtained from discrete trial responding. For example, duration data may be obtained on the latency of response to each of a set of specific requests, or frequency data may be obtained on the number of self-stimulatory responses following each trial of a learning task. Second, a period of observation may be artificially divided into smaller recording intervals (trials) and frequency or duration data obtained for each of these artificially constructed trials. For example, event recording of tantruming during a 40-min observation period may be artificially divided into 10, 4-min observation periods. Breaking a session into smaller units assists in identifying behavioral incidents that pose scoring problems, provides a more stringent test of reliability, and provides a more efficient method of determining session reliability (by estimating session reliability from trial reliability).

The techniques generally used for determining the interobserver reliability of trial frequency and duration scores are the same as those described for session reliability data. For the trial data presented in Table 4, percentage agreement values defined as $100 \times$ (smaller frequency/larger frequency) range from $67\% = [100 \times (2/3)]$ to $100\% = [100 \times (2/2)]$ and $[100 \times (3/3)]$. Percentage agreement defined as $100 \times$ (number of trials for which both observers agree/number of trials jointly observed) for these same data is $[100 \times (2/5)] = 40\%$ while r_{xx} is $+0.94$. The discussion of these techniques in the section on session reliability is directly applicable to the issues concerning the trial reliability of duration and frequency scores. That information may be generalized to this section by making the following substitutions: substitute r_{xx} for r_{kk} and "trial" for "session".

Acceptable Values of Trial Reliability

No entirely agreed upon set of rules for deciding on an acceptable value of trial (or session) reliability has yet been formulated. Percentage agreement of 80% for trial reliability seems to have some consensus among applied behavioral researchers. Gelfand and Hartmann (1975) rec-

Table 4

Frequency data used to illustrate the calculation of trial reliability for continuous data.

	Observation Period				
	1	*2*	*3*	*4*	*5*
Observer 1	4	3	5	2	3
Observer 2	5	3	6	2	2

ommend that ϕ, kappa, and r_{xx} should exceed 0.60. As a general rule, the fewer the number of data points, the smaller the behavioral change expected, and the greater the variability of the target behavior, the higher must be the interobserver reliability for there to be a reasonable likelihood of detecting the change produced by treatment. While analytic procedures are available for determining the correlational reliability required to detect an expected treatment effect with a specific probability (Cohen, 1969), these procedures require more information than is generally available.

Relationship Between Trial and Session Reliability

Because session scores are composites formed of trial scores, there is a relationship between the reliability of trial scores and the reliability of session scores. In the case of ϕ or r_{xx} performed on trial scores, the relationship to r_{kk} is formal and mathematically precise. The reliability of trial scores indexed by these correlational measures will, in most cases, provide a lower-bound estimate of the reliability of session scores (Hartmann, 1976). Thus, if the average value of ϕ or $r_{xx} = 0.6$, r_{kk} in most cases will be substantially higher. Percentage agreement statistics calculated on trial scores do not relate formally to session reliability scores.[10]

Because of the substantial number of factors relevant to choosing one of the four reliability statistics for occurrence-nonoccurrence data, the principal advantages and limitations of these four statistics are summarized in Table 5.

Whichever reliability method seems most suitable for a particular study, reliability statistics should be presented in a manner that allows easy translation from one statistic to another. For two-by-two table data, investigators might present the proportional rates of occurrence of the target behavior as rated by each of the observers, the proportion of observations in the occurrence-occurrence cell (Cell B), and the total number of jointly observed trials. This information would permit readers to reconstruct the table and calculate any of the four reliability statistics. Similarly, the value of r_{xx}, the standard deviation of trial scores, and the mean difference between the two observers' trial ratings would provide readers with the information necessary to estimate the average and minimum values of the percentage agreement statistic for continuous trial scores. (In the case of session scores, replace r_{xx} with r_{kk}, and "trials" with "sessions" in the previous sentence.)

CONCLUSIONS

The tradition in applied behavioral research has been to perform trial reliability analyses, usually by means of the percentage agreement statistic. The primary intent of these analyses is presumably to provide a more stringent estimate of reliability than is provided by session-score

[10]To demonstrate the difference between correlational and percentage agreement measures of reliability and their relationship to session reliability, a simple simulation study was undertaken. The study was based on 100, 5-sec intervals and simulated occurrence-nonoccurrence measures. Observers' scores were generated by using a table of random numbers, with digits 1 to 8 coded as occurrence and 9 and 0 as nonoccurrence. Trial reliability measures yielded the following results when applied to these data: percentage agreement was 65% (slightly lower than the 68% expected); percentage effective agreement for occurrences was 64%; percentage effective agreement for nonoccurrences was 10%; kappa was −0.03; and ϕ was −0.04. Three measures of session reliability (r_{kk}) were also calculated on these same data: r_{kk} was calculated on the 20, 25-sec collapsed trials, on the 10, 50-sec collapsed trials, and on the 5, 100-sec collapsed trials. These analyses were undertaken to provide an analogue to the reliability of composite (session) scores. In the three analyses, the values of r_{kk} varied from −0.01 to −0.18. Since the two observers' data were uncorrelated (independent), there would seem little doubt that kappa, ϕ, and to a lesser degree, nonoccurrence agreement, were far more accurate in describing the degree of interobserver reliability at the trial level than were percentage agreement or occurrence agreement. Approximately the same values for all statistics, with the exception of percentage agreement and occurrence reliability, would have been obtained if the digits 1 to 9 had been coded as occurrences, and the digit 0 had been coded as nonoccurrences. In that case, however, percentage agreement and occurrence agreement would have been approximately 82%!

Table 5

Summary of Interobserver Reliability Methods Suitable for Occurrence-Nonoccurrence Trial Data

Measure of Reliability	*Formula*	*Tests of Significance Available for*		*Ease of*		*Effected by*		*Functions in*		*Possibly Inflated by Extreme Rates of Occurrence*	*Range of Applicability*	*Relationship to Session Reliability*
		Independence	*Observer Bias*	*Computation*	*Interpretation*	*Random Error*	*Systematic Error*	*Estimation Equation*	*Formal Reliability Theory*			
Percentage Agreement	$[(B+C)/N] \times 100$	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Moderate to Broad	Informal
Effective Percentage Agreement	$[B/(A+B+D)] \times 100$; $[C/(A+C+D)] \times 100$	Yes[a]	Yes	Yes	Yes	Yes	Yes	No	No	Slightly	Moderate to Broad	Informal
Phi	$BC-AD/[(A+B)(C+D)(A+C)(B+D)]^{1/2}$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Moderate	Formal
Kappa	$(p_o-p_c)/(1-p_c)$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Broad	Formal

[a]The test of significance involves all data in the two-by-two summary table.

reliability. Unfortunately, the material in the present paper suggests that under certain circumstances, percentage agreement may provide insufficient or misleading information about session scores, and spuriously high, rather than stringent, estimates of trial reliability. In addition, reliance on percentage agreement trial reliability has tended to restrict behavioral researchers' contact with the extensive body of theory and experience, which have their bases in correlational analysis and test theory. Perhaps the general ill repute of psychological tests has indeed resulted in the baby being thrown out with the bath water. This paper's aim was to provide applied behavior analysts with information relevant to the choice of reliability assessment methods, some of which may be unfamiliar because of the methods' historical roots in psychological testing and measurement.

REFERENCES NOTES

1. Hartmann, D. P. *Assessing the quality of observational data.* Paper presented at the Western Psychological Association Meeting, San Francisco, April 1974.
2. Whelan, P. A. *Reliability of human observers.* Unpublished doctoral dissertation, University of Utah, Salt Lake City, Utah, 1974.

REFERENCES

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.

Cohen, J. *Statistical power analysis for the behavioral sciences.* New York: Academic Press, 1969.

Cronbach, L. J., Glaser, G. C., Nanda, H., and Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley, 1972.

Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological 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.

Gelfand, D. M. and Hartmann, D. P. *Child behavior analysis and therapy.* New York: Pergamon Press, 1975.

Guilford, J. P. and Fruchter, B. *Fundamental statistics in psychology and education.* 5th ed.; New York: McGraw-Hill, 1973.

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

Haggard, E. A. *Intraclass correlation and the analysis of variance.* New York: Dryden Press, 1958.

Hartmann, D. P. Some restrictions in the application of the Spearman-Brown prophesy formula to observational data. *Educational and Psychological Measurement,* 1976, **36,** 843-845.

Jensen, A. R. The reliability of projective techniques: methodology. *Acta Psychologica,* 1959, **16,** 108-136.

Johnson, S. M. and Bolstad, O. D. Methodological issues in naturalistic observation: Some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, and E. J. Mash (Eds.), *Behavior change: methodology, concepts and practice.* Champaign, Ill.: Research Press, 1973.

Kazdin, A. E. and Klock, J. The effect of nonverbal teacher approval on student attentive behavior. *Journal of Applied Behavior Analysis,* 1973, **6,** 643-654.

Krippendorff, K. Bivariate agreement coefficients for reliability of data. In E. F. Borgatta and G. W. Bohrnstedt (Eds.), *Sociological methodology.* San Francisco: Jossey-Bass, 1970. Pp. 139-150.

Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores.* Menlo Park, Calif.: Addison-Wesley, 1968.

McNemar, Q. *Psychological statistics.* 4th ed.; New York: Wiley, 1969.

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

Patterson, G. R. and Cobb, J. A. Stimulus control for classes of noxious behavior. In J. F. Knutson (Ed.), *The control of aggression.* Chicago: Aldine, 1973. Pp. 145-199.

Schmidt, G. W. and Ullrich, R. E. Group contingent events and classroom noise. *Journal of Applied Behavior Analysis,* 1969, **2,** 171-179.

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

Wahler, R. G. Some structural aspects of deviant child behavior. *Journal of Applied Behavior Analysis,* 1975, **8,** 27-42.

Wiggins, J. S. *Personality and prediction: Principles of personality assessment.* Reading, Mass.: Addison-Wesley, 1973.

Winer, B. J. *Statistical principles in experimental design.* 2nd ed.; New York: McGraw-Hill, 1971.

Received 19 September 1974.
(Final acceptance 15 May 1976.)