JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1977, 10, 121-126 NUMBER 1 (SPRING) 1977

EVALUATING INTEROBSERVER RELIABILITY OF INTERVAL DATA[1]

B. L. HOPKINS AND JAIME A. HERMANN

UNIVERSITY OF KANSAS AND
UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

Previous recommendations to employ occurrence, nonoccurrence, and overall estimates of interobserver reliability for interval data are reviewed. A rationale for comparing obtained reliability to reliability that would result from a random-chance model is explained. Formulae and graphic functions are presented to allow for the determination of chance agreement for each of the three indices, given any obtained per cent of intervals in which a response is recorded to occur. All indices are interpretable throughout the range of possible obtained values for the per cent of intervals in which a response is recorded. The level of chance agreement simply changes with changing values. Statistical procedures that could be used to determine whether obtained reliability is significantly superior to chance reliability are reviewed. These procedures are rejected because they yield significance levels that are partly a function of sample sizes and because there are no general rules to govern acceptable significance levels depending on the sizes of samples employed.

DESCRIPTORS: interobserver reliability, interval data, statistical inference, chance agreement, reliability criteria

Much research involving applied behavior analyses employs data collected by observers who record the occurrence of responses during short time intervals (*e.g.,* Ayllon and Roberts, 1974; Glynn and Thomas, 1974; Knapczyk and Livingston, 1974). Such research assesses the reliability of observations by having two observers simultaneously record the same responses. The two records are compared interval-by-interval to determine the percentage of intervals in which the two observers agree that the behavior did or did not occur.

This index might be called *overall reliability* and is defined by:

$$R_{overall} = \frac{O_{1\&2} + N_{1\&2}}{T} \times 100 \qquad (1)$$

where

$O_{1\&2}$ = the number of intervals in which *both* Observer 1 and Observer 2 record the response as occurring;

$N_{1\&2}$ = the number of intervals in which *both* Observer 1 and Observer 2 record the response as not occurring; and

T = the total number of intervals for which the two observers' records are compared.

For example, if two persons simultaneously observe for 100 intervals, and both record some response as occurring in the same 63 intervals and do not record the response as occurring in the same 17 intervals (during the remaining 20 intervals, one or the other, but not both, records the response as occurring), the overall index of reliability would be:

$$R_{overall} = \frac{63 + 17}{100} \times 100$$

$$R_{overall} = 80\%$$

The ratio of intervals of agreement to total intervals is commonly multiplied by 100 to yield a percentage.

[1]This manuscript is part of a paper, "Problems in Experimental Design and Data Analysis", presented at the American Psychological Association Meetings in Montreal, Canada, 1973. Preparation of this manuscript was supported in part by SRS grant 59-P-35116. Reprints may be obtained from B. L. Hopkins, Department of Human Development, University of Kansas, Lawrence, Kansas 66045.

Bijou, Peterson, and Ault (1968) mentioned that the above index of reliability may be difficult to interpret whenever responses are recorded as occurring in either a large percentage or a small percentage of intervals. Table 1 is a hypothetical example of the problem that can exist for responses recorded as occurring in only a few intervals. Each observer has recorded the response as occurring in one of the 10 intervals. The observers agreed by making similar observations in eight of the 10 intervals. However, they failed to agree on intervals in which the response is recorded as occurring. Such records would cause doubt that the observers are, in fact, agreeing on occurrences of the response.

Similar problems exist for responses recorded as occurring in most intervals. Table 2 is a hypothetical example of such a problem. The two observers recorded the response as occurring in 90% of the intervals. Moreover, the observers agreed that the behavior did or did not occur in 80% of the intervals. Nevertheless, the observers failed to agree on the intervals in which the response is not recorded as occurring. This discrepancy is crucial. The observers might be recording two entirely different but relatively high-rate behaviors, and interval-by-interval comparison of their records would yield many intervals of agreement simply because both are recording some response as occurring in most intervals.

Table 1

A hypothetical example of records obtained by two observers recording a behavior as occurring in a small percentage of intervals.

	Short Intervals									
Observer 1		B								
Observer 2									B	
Overall agreement	✓	✓	✓	✓	✓	✓	✓	✓		
$O_{1\&2}$										
$O_{1 \text{ or } 2}$			✓						✓	

Table 2

A hypothetical example of records obtained by two observers recording a behavior as occurring in a large percentage of intervals.

	Short Intervals									
Observer 1	B	B		B	B	B	B	B	B	B
Observer 2	B	B	B	B	B	B	B	B		B
Overall Agreement	✓	✓		✓	✓	✓	✓	✓		✓
$N_{1\&2}$										
$N_{1 \text{ or } 2}$			✓						✓	

Because of these problems, Bijou *et al.* (1968) recommended that an index of *occurrence reliability* be computed for very low-rate behaviors and an index of *nonoccurrence reliability* for high-rate behaviors. However, as Hawkins and Dotson (1975) noted, their recommendations have not been widely adopted by researchers. The calculation definitions for these indices are:

$$R_{\text{occurrence}} = \frac{O_{1\&2}}{T} \times 100 \qquad (2)$$

and

$$R_{\text{nonoccurrence}} = \frac{N_{1\&2}}{T} \times 100 \qquad (3)$$

In the example of Table 1, there are no intervals in which both observers record the response as occurring and two intervals in which either observer records the behavior as occurring. Thus, there is 0% agreement on occurrences of the behavior. Similarly in Table 2, although there is 80% agreement on the overall reliability index, the two observers fail to agree on intervals in which the response does not occur and there is, therefore, 0% agreement on the nonoccurrence index.

Routine methods are available to compare obtained percentages of agreement to agreement that would be expected by a random-chance model. The chance model assumes that the two observers record the response as occurring in the same number of intervals as it is empirically determined to occur. However, the model further

assumes that the recording of instances of the response are randomly distributed over intervals. It is then possible to determine whether the empirically determined reliability as obtained by two actual observers is superior to reliability that might be obtained by chance.

Computation formulae for these chance-reliability indices can be deduced from the basic theorems of probability theory for independent events (Feller, 1957). They are:

$$\text{Chance } R_{overall} = \frac{(O_1 \times O_2) + (N_1 \times N_2)}{(T)^2} \times 100 \quad (4)$$

$$\text{Chance } R_{occurrence} = \frac{O_1 \times O_2}{(T)^2} \times 100 \quad (5)$$

$$\text{Chance } R_{nonoccurrence} = \frac{N_1 \times N_2}{(T)^2} \times 100 \quad (6)$$

where

O_1 = the number of intervals in which Observer 1 records the response as occurring;

O_2 = the number of intervals in which Observer 2 records the response as occurring;

N_1 = the number of intervals in which Observer 1 records the response as not occuring;

N_2 = the number of intervals in which Observer 2 records the response as not occurring; and

T = the total number of intervals for which the two observers' records are compared.

Suppose that two observers are recording on-task behavior for a retarded child in a classroom and that they simultaneously observe for 100, 10-sec intervals. Further suppose that both record the response as occurring in 90 intervals and that the index of overall reliability, as calculated by formula (1) above, indicates that they agree on 80% of the intervals. We can employ formula (4) to determine if the obtained percentage of agreement is better than would be obtained by chance:

$$\text{Chance } R_{overall} = \frac{(90 \times 90) + (10 \times 10)}{(100)^2} \times 100 = 82\%$$

Indeed, the obtained percentage of agreement, 80, would be less than the 82% expected by chance.

Each of the three computational formulae, (4), (5), and (6), is constructed in such a way that chance agreement varies with the empirical per cent of intervals in which the response is recorded as occurring. By assuming various proportions of intervals in which a behavior is recorded as occurring, the entire chance functions can be developed. For example, if two observers are recording a behavior as occurring in 10% of the observation intervals,

$$\begin{aligned}\text{Chance } R_{overall} &= [(pO_1 \times pO_2) + (pN_1 \times pN_2)] \times 100 \\ &= [(0.10 \times 0.10) + (0.90 \times 0.90)] \times 100 \\ &= [0.01 + 0.81] \times 100 = 82\%.\end{aligned} \quad (7)$$

If the behavior is being recorded as occurring in 30% of the intervals,

$$\begin{aligned}\text{Chance } R_{overall} &= [(0.30 \times 0.30) + (0.70 \times 0.70)] \times 100 \\ &= [0.09 + 0.49] \times 100 = 58\%.\end{aligned}$$

Formula (7) is equivalent to formula (4) but has been transformed to deal with proportions of intervals, rather than actual numbers of intervals to allow for easy computations for the hypothetical cases. The terms pO_1, pO_2, *etc.*, are the proportion of intervals in which Observer 1 records the behavior as occurring, the proportion in which Observer 2 records the behavior as occurring, *etc.*

The entire function for chance overall reliability is plotted in Figure 1. The function is qualitatively similar to one published by Hawkins and Dotson (1975) but is more exact than theirs. To use this function, first determine the per cent of intervals in which the observers are recording the behavior as occurring. Find this per cent on the horizontal axis. Project a straight line vertically from that point. The point at which the projected line intercepts the function provides the per cent reliability that would be obtained by chance by projecting a horizontal line from that point on the function to the vertical axis. For example, as determined on Figure 1, chance agreement is about 73% for a response recorded

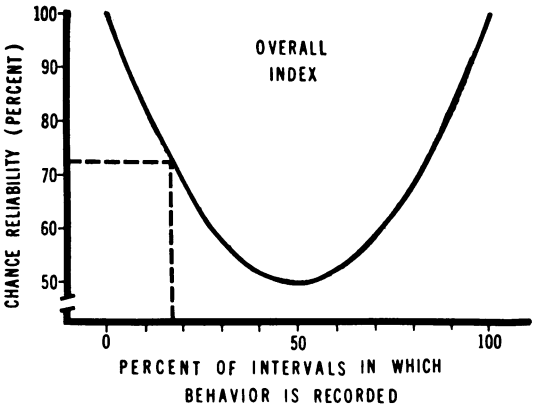

Fig. 1. Overall chance reliability as a function of the per cent of intervals in which a behavior is recorded as occurring.

as occurring in about 16% of the intervals of observation.

The function in Figure 1 assumes that the two observers are recording the response as occurring in about the same per cent of intervals. If there are discrepancies in the per cent of intervals in which the response is reported to occur by the two observers, then the calculation formula (4) should be employed, rather than Figure 1.

Figure 2 is the chance-reliability functions developed for the occurrence and nonoccurrence indices. The function for the overall index and the function for the occurrence index converge as the per cent of intervals in which the response is recorded approaches 100. Similarly, the overall and nonoccurrence functions converge as the per cent of intervals in which the response is recorded approaches zero.

The functions of Figure 2 are used in the same way as described above for the function for the overall index. Thus, chance occurrence reliability for observers recording a response as occurring in 10% of intervals is only 1%. Chance nonoccurrence reliability for observers recording a response as occurring in 70% of intervals is only 9%. Again, the calculation formulae (5) and (6) should be used instead of the figures unless both observers are recording the response as occurring in about the same percentage of intervals.

Inspection of Figures 1 and 2 indicates that all three functions are continuous for the entire range of the per cent of intervals in which a response is recorded as occurring. Therefore, contrary to the Bijou *et al.* (1968) recommendation, all of the indices of reliability are interpretable, regardless of per cent of intervals in which observers record a response. The level of chance agreement simply changes as the per cent of intervals in which the behavior is recorded changes.

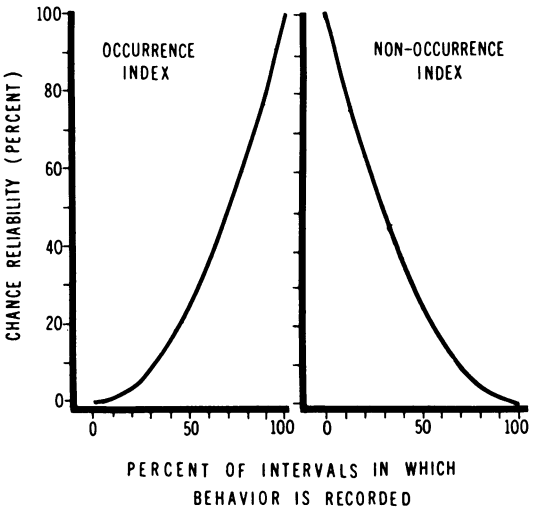

Fig. 2. Occurrence and nonoccurrence chance reliability as a function of the per cent of intervals in which a behavior is recorded as occurring.

Suppose, for example, that two observers independently record whether or not a child engages in some social interaction during each of 100, 10-sec intervals. Further suppose that each observer records the behavior as occurring in 20% of the intervals and that the empirical percentages of agreement for the two observers, as calculated by formulae (1), (2), and (3) are:

$$R_{overall} = 80\%$$
$$R_{occurrence} = 33\%$$
$$R_{nonoccurrence} = 78\%.$$

Chance agreement for these indices, as determined either by projection on Figures 1 and 2 or by calculation with formulae (4), (5), and (6) is:

$$\text{Chance } R_{overall} = 68\%$$
$$\text{Chance } R_{occurrence} = 04\%$$
$$\text{Chance } R_{nonoccurrence} = 64\%.$$

In fact, obtained reliability for all three indices is greater than would be expected by chance.

Once either the figures or the calculation formulae have been employed to determine that observations of independent observers are better than would be expected by chance, it is reasonable to question how much better. This, of course, is the kind of question for which inferential statistics might be appropriate. Several statistical procedures have been recommended for dealing with such problems. Cohen's k or kappa (Cohen, 1960) is a coefficient of interjudge agreement that excludes chance agreements. The Phi coefficient provides a measure of the correlation between the records of two different observers (Young and Veldman, 1965). Both of these descriptive statistics can be related to inferential statistics to yield an estimate of the probability that obtained reliability is superior to chance reliability. Similarly, Fisher's exact test or a Chi square can be used to compute the likelihood that agreements as good as those obtained could be attributed to chance (Siegel, 1956). However, these statistical procedures provide ambiguous answers to questions regarding how much better than chance a particular degree of reliability may be. Essentially, statistical significance increases and confidence levels decrease as sample sizes increase. Furthermore, agreement even only slightly superior to chance may become statistically significantly better than chance as sample sizes become large.

Consider again the example in which two observers record the social interactions of a child and their agreement is subsequently compared with chance agreement. Again, assume that the obtained index of overall reliability is 80% and chance overall reliability is 60%. Suppose we calculate kappa and then ask if the obtained kappa is significantly different from zero. If our data involve 100 intervals, the probability of obtaining the calculated kappa by chance is less than 0.006, while if exactly the same relationships held, but our data were based on only 10 intervals, the probability would be less than 0.23.

The other inferential statistics behave in exactly the same fashion as kappa. Unless researchers had some rule of thumb, or perhaps method based on experience, to determine what might be an acceptable level of significance for a given sample size in a particular area of research, the results of the inferential statistics provide virtually useless information. Moreover, for the large samples often involved in calculations of reliability in applied behavior analyses, the inferential statistics are particularly generous. Returning to our example, if obtained overall agreement were 80%, and if this were only slightly greater than a calculated chance agreement of 78%, kappa would still be significant at the 0.01 level if the observations were based on only 150 intervals.

At this time, there appears to be no satisfactory way to determine that obtained reliability is acceptably better than chance reliability. Therefore, the procedures for calculating chance reliability can only describe a lower boundary at which obtained reliability is unacceptable. In practice, researchers would generally demand higher degrees of interobserver reliability if the effects of independent variables are slight than if large effects were obtained, because apparent slight effects might simply be attributable to observation errors. However, there is no objective rule that allows a researcher to translate this consideration into a greater-than-chance lower limit for acceptable obtained reliability.

CONCLUSIONS

Four recommendations follow from the above considerations:

1. All publications dealing with interval data should report indices of interobserver agreement, as suggested by Bijou *et al.* (1968) and Hawkins and Dotson (1975).

2. Researchers should calculate and publish indices of random-chance interobserver agree-

ment against which obtained measures can be compared.

3. All indices (overall, occurrence, and non-occurrence) of reliability can be interpreted regardless of the per cent of intervals in which a response is recorded. This is contrary to the recommendations of Bijou *et al.* and of Hawkins and Dotson.

4. Until other statistics are developed, researchers should postpone considerations of how much better than chance is the obtained interobserver reliability.

REFERENCES

Ayllon, T. and Roberts, M. D. Eliminating discipline problems by strengthening academic performance. *Journal of Applied Behavior Analysis,* 1974, **7,** 71-76.

Bijou, S. W., Peterson, R. F., and Ault, M. H. A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. *Journal of Applied Behavior Analysis,* 1968, **1,** 175-191.

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

Feller, W. *An introduction to probability theory and its applications,* Vol. 1. New York: John Wiley & Sons, 1957.

Glynn, E. L. and Thomas, J. D. Effect of cueing on self-control of classroom behavior. *Journal of Applied Behavior Analysis,* 1974, **7,** 299-306.

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, New Jersey: Prentice-Hall, 1975. Pp. 359-376.

Knapczyk, D. R. and Livingston, G. The effects of prompting question-asking upon on-task behavior and reading comprehension. *Journal of Applied Behavior Analysis,* 1974, **7,** 115-121.

Siegel, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

Young, R. K. and Veldman, D. J. *Introductory statistics for the behavioral sciences.* New York: Holt, Rinehart & Winston, 1965.

Received 14 October 1974.
(Final acceptance 15 May 1976.)