JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1978, 11, 523-527 NUMBER 4 (WINTER 1978)

A METHOD FOR COMBINING OCCURRENCE AND NONOCCURRENCE INTEROBSERVER AGREEMENT SCORES

FRANCIS C. HARRIS AND BENJAMIN B. LAHEY[1]

UNIVERSITY OF GEORGIA

Various statistics have been proposed as standard methods for calculating and reporting interobserver agreement scores. The advantages and disadvantages of each have been discussed in this journal recently but without resolution. A formula is presented that combines separate measures of occurrence and nonoccurrence percentages of agreement, with weight assigned to each measure, varying according to the observed rate of behavior. This formula, which is a modification of a formula proposed by Clement (1976), appears to reduce distortions due to "chance" agreement encountered with very high or low observed rates of behavior while maintaining the mathematical and conceptual simplicity of the conventional method for calculating occurrence and nonoccurrence agreement.

DESCRIPTORS: reliability, interobserver agreement, combining occurrence and nonoccurrence scores, critical assessment of commonly used procedures

The field of applied behavior analysis currently relies heavily on data collected by human observers (Kelly, 1977). These data are typically considered reliable if two independent observers reach an "acceptable" level of agreement on the occurrence and/or nonoccurrence of a target behavior, using more-or-less standard observation methods. The demonstration of an acceptable level of interobserver agreement (and, presumably, of objectivity) is crucial to applied behavior analysis, but therein lies the problem. No current method of calculating interobserver agreement has been widely accepted, although several have been proposed. The need for a standard method by which interobserver agreement can be computed has been discussed in a recent series of articles in the *Journal of Applied Behavior Analysis* (Baer, 1977; Hartmann, 1977; Hopkins and Hermann, 1977; Kratochwill and Wetzel, 1977; Yelton, Wildman, and Erickson, 1977).

Percentage Agreement Statistics

One commonly used statistic in interval recording has been overall percentage agreement. This typically is determined by counting the number of intervals in which the observers agree on occurrences *and* nonoccurrences, dividing by the total number of observation intervals, and multiplying the quotient by 100. This statistic has face validity, in that it gives the percentage of intervals in which observers agreed that the behavior occurred and did not occur.

Overall percentage agreement generally has been considered to be susceptible to misinterpretation, however, when a relatively high or low number of intervals is scored. This can be considered to be due to the probability of "chance" agreements being high. For example, if in a 100-interval observation session each observer scored 10 occurrences, but only two in the same intervals, the overall agreement percentage would be 84% [(two agreements on occurrence + 82 agreements on nonoccurrence = 84 agreements) ÷ 100 intervals = 84%]. In this case, the high number of unscored recording intervals can be assumed to result in a high frequency of chance agreements that inflate the agreement score. In the case of high rates of recorded behavior, a high number of intervals would be marked by both observers. If the two observers

[1]Reprints may be obtained from Benjamin B. Lahey, Psychology Clinic, Dept. of Psychology, University of Georgia, Athens, GA 30602.

were randomly marking at a high rate of occurrences, a large number of intervals would be marked by *both* observers and would, therefore, be counted as agreements. If agreements on occurrences *and* nonoccurrences were included and given equal weight in the calculation of the agreement score, a high score would be obtained, even though the records of the two observers were unrelated (random). The same reasoning would apply in the case of low-rate behaviors in which the high number of chance agreements on unmarked intervals would inflate the overall agreement score (Hartmann, 1977). Hopkins and Hermann (1977) stated the same point in a different way: "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" (p. 122).

One method of reducing the threat of such chance agreements, currently used by many behavior analysts, is to calculate the interobserver agreement for scored intervals only when the observed rate of behavior is low and for unscored intervals only when the observed rate is high. It typically is calculated by dividing the number of intervals in which the observers agree on occurrences (nonoccurrences) by the total number of intervals in which at least one observer scored an occurrence (nonoccurrence). This method may overcompensate, however, by throwing out *all* of the agreement data on unscored or scored intervals, respectively. In the example cited above, not all of the agreements on unscored intervals could be assumed to be "chance" agreements. In addition, this method is not appropriate for the many studies in which rates of observed behavior vary (Hartmann, 1977). This is due to the absence of an objective method for determining the frequency at which one score should be used instead of the other. The alternative of reporting occurrence *and* nonoccurrence agreement scores for each session would result in an unnecessary inconvenience to the research consumer (Kratochwill and Wetzel, 1977).

Recently, several investigators have suggested alternate methods for dealing with the problem of chance agreement. Hopkins and Hermann (1977) suggested that overall percentage agreement might be interpretable if it were compared to the overall agreement percentage expected by chance and presented formulas for calculating agreement scores that would be expected by chance. Minimum criterion for an acceptable level of agreement would be an obtained score greater than that expected by chance alone. In the example cited above, obtained agreement would be 84% and chance agreement would be 82% (Hopkins and Hermann, 1977). Thus, satisfactory agreement would be obtained (by a margin of 2%) even though the observers could agree on only two occurrences while they disagreed on 16. Furthermore, if, in a 100-interval session each observer scored 45 occurrences but only 21 in the same intervals, the overall percentage agreement score would be 52%, and the score expected by chance would be 51%. Thus, according to Hopkins and Hermann, adequate agreement would have been obtained even though there was only 52% overall agreement. The mathematically derived minimum criterion of chance agreement is appealing, but there is no reason to believe that it is any more *useful* to the behavior analyst than some arbitrary, but conventional level such as 80% agreement. This is similar to the clinical *versus* statistical significance issue. In the above example, the proposed statistical criterion was met, but many behavior analysts would not consider the data to be "reliable" (useful) because of the relatively low proportion of occurrence agreements.

A method of calculating occurrence (nonoccurrence) agreement that permits comparison to a score expected by chance also has been described by Hopkins and Hermann (1977). The number of intervals scored (unscored) by both observers is divided by the total number of intervals (regardless of how many were scored by

either). This percentage then can be compared with the one expected by chance, with acceptable agreement being any score greater than chance. For our first example, the occurrence agreement percentage equals 2% and the chance percentage equals 1%. Thus, according to Hopkins and Hermann, adequate agreement would have been reached. This could be misleading for the same reasons that were presented for the Hopkins and Hermann overall percentage agreement statistic. Like the conventional occurrence and nonoccurrence agreement percentages, those described by Hopkins and Hermann minimize chance agreements by not considering nonoccurrence (occurrrence) agreements. They differ from the conventional agreement percentages in that the divisor is always the number of intervals in the session. This makes the possible range of the statistics dependent on the number of intervals scored. For example, a "perfect" agreement percentage could be 10% for one session and 90% for another.

Correlation-Like and Probability-Based Methods

Correlation-like measures have been proposed to minimize the chance agreement problem (Hartmann, 1977). Essentially, they express a comparison between observed and expected interobserver agreement, but in a manner that is more mathematically complicated than the formulas of Hopkins and Hermann (1977). They can assume any value between −1.0 and +1.0. Interpretation of them generally requires greater statistical sophistication than statistics that use the simple 0% to 100% scale. Kratochwill and Wetzel (1977) pointed out that another disadvantage ". . . is that their 'novel' feature could cause investigators to employ them to the exclusion of simpler statistical aids that could adequately represent observer agreement" (p. 138).

A probability-based formula that gives the exact probability of obtaining at least any given number of overall agreements has been put forth by Yelton, Wildman, and Erickson (1977). Interpretation of this statistic also requires greater statistical sophistication than does those using the 0% to 100% scale. In addition, its novelty could cause investigators to use it in lieu of simpler statistical aids, such as percentage agreement scores. Furthermore, its cumbersome mathematics make it unlikely to be adopted by many behavior analysts.

The correlation-like and probability-based methods differ from the more commonly used methods, in that they each provide a formal method of comparing an obtained agreement score with one expected by chance, rather than describing the *degree* of agreement. As with the Hopkins and Hermann (1977) method, the issue in evaluating these methods is the same as the issue of statistical *versus* clinical significance. The correlation-like and probability-based formulas tell us whether obtained interobserver agreement exceeds a mathematically determined minimum standard of "significance", rather than assessing the extent to which the degree of interobserver agreement reaches some conventional level of "usefulness". Both involve pure assumptions: one involves a mathematical model of chance agreements; the other involves a conventional standard of utility.

Combining Occurrence and Nonoccurrence Percentage Agreement Scores

The formulas suggested for calculating interobserver agreement in the Spring 1977 issue of the *Journal of Applied Behavior Analysis* seek to minimize the chance agreement problem using different, but statistically sound, procedures. Baer (1977) noted the arbitrariness of all methods of calculating interobserver agreement and suggested that the choice of a standard method be based on "(1) the avoidance of allowing the reliability of occurrence from influencing the reliability of nonoccurrence and *vice versa;* and (2) by the apparent, face meaning of the estimate's calculation technique" (p. 117). The separate calculation of conventional occurrence and nonoccurrence agreement percentages fits Baer's criteria perfectly. (1) It minimizes the likelihood of allowing occurrence agreement to

influence nonoccurrence agreement, and *vice versa;* and (2) it has good face validity ". . . Two observers watching one subject, and equipped with the same definition of behavior . . . agree about its occurrence X% of the relevant intervals, and about its nonoccurrence Y% of the relevant intervals" (Baer, 1977, p. 118).

In addition to the problems with this procedure already noted, however, the interpretation of separately calculated coefficients of occurrence and nonoccurrence agreement is uncertain. We have no guidelines as to how much "weight" to give to each coefficient at differing observed rates of behavior.

If, in studies in which behavior levels vary over time, a single agreement score is required to summarize interobserver agreement and simplify the task of research consumers (Kratochwill and Wetzel, 1977), some combination of occurrence and nonoccurrence agreement scores that differentially weights each score on the basis of the observed behavior frequency would seem appropriate. Although some difficulties are associated with it, such a statistic has been proposed by Clement (1976):

$$\text{Interobserver agreement} = (A \times B) + (C \times D)$$

where

A is the number of agreements for occurrences divided by the number of time samples marked by the "standard" observer;

B is 1.00—(occurrences marked by the "standard" observer divided by the total number of time samples);

C is the number of agreements for nonoccurrences divided by the number of nonoccurrences indicated by the "standard" observer; and

D is 1.00—(nonoccurrences indicated by the "standard" observer divided by the total number of time samples).

In essence, Clement's formula provides a weighted mean of indices of occurrence and nonoccurrence agreement, with weight assigned to these two indices according to the frequency at which behavior is recorded. Proportionately greater emphasis is placed on occurrence agreement when relatively few intervals are scored and proportionately greater emphasis is placed on nonoccurrence agreement when a relatively high number of intervals is scored. This compensates for distortions due to to "chance" agreements with high- or low-rate behaviors without eliminating any data.

Clement's formula, therefore, offers a solution to the chance agreement dilemma inherent in other formulas of interobserver agreement. Two modifications of his formula are apparently needed, however, to bring it more in line with conventional thinking in applied behavior analysis. First, the A and C terms in Clement's equation provide inaccurate occurrence and nonoccurrence agreement scores. They should be calculated by dividing the number of agreements on occurrences (or nonoccurrences) by the total number of intervals marked (or left unmarked) by *either* observer, rather than by only one observer (the "standard" observer). Clement's formula overestimates agreement by providing an incomplete divisor. Second, the weighting factor should be the mean of the occurrences recorded by both observers, rather than arbitrarily designating one person as the standard observer. Since data are being combined to yield one agreement score, differentially assigning more or less weight to one observer's score is inappropriate. The modified formula for weighted agreement is therefore:

$$WA = (O \times U) + (N \times S) \times 100$$

where

O is the occurrence agreement score, *i.e.,* the number of occurrence agreements divided by (the number of occurrence agreements + the number of occurrence disagreements);

U is the mean proportion of unscored intervals, *i.e.,* (the proportion of intervals not scored by Observer 1 + proportion

of intervals not scored by Observer 2) divided by 2;

N is the nonoccurrence agreement score, *i.e.,* the number of nonoccurrence agreements divided by (the number of nonoccurrence agreements + the number of nonoccurrence disagreements);

S is the mean proportion of scored intervals, *i.e.,* (the proportion of intervals scored by Observer 1 + proportion of intervals scored by Observer 2) divided by two.

More simply, this formula may be conceptualized as occurrence agreement weighted by the average rate of nonoccurrence, plus nonoccurrence agreement weighted by the average rate of occurrence.

For example, if in a 100-interval observation session one observer scored 25 occurrences, the other scored 30, they agreed on occurrences 20 times and nonoccurrences 65 times, and disagreed 15 times each on occurrences and nonoccurrences:

$$\begin{aligned} \mathrm{WA} &= \left(\frac{20}{20+15}\right)\left(\frac{0.75+0.70}{2}\right) \\ &\quad + \left(\frac{65}{65+15}\right)\left(\frac{0.25+0.30}{2}\right) \times 100 \\ &= (0.57)(0.72) + (0.81)(0.28) \times 100 \\ &= 64\% \end{aligned}$$

Note that in this example, 72% of the weight is assigned to the occurrence agreement score and 28% is assigned to the nonoccurrence score.

This formula differs from that of Clement's (1976), then, in calculating separate agreement coefficients for marked and unmarked intervals using a complete divisor (the total number of intervals marked by either observer, rather than just one observer) and by using the mean of the occurrences recorded by both observers as the weighting factor, rather than the occurrences recorded by one observer.

The score yielded by the above formula must always be between 0% and 100%, such that a convention similar to that of an adequate score being approximately 80% or greater could be adopted. The weighted agreement formula yields a single score that minimizes chance agreement and makes use of all the available interobserver agreement information by combining occurrence and nonoccurrence interobserver agreement scores. It also permits evaluation of interobserver agreement on the familiar 0% to 100% scale. It appears to be an especially useful, efficient, and convenient method for expressing interobserver agreement in studies in which the frequency of the target behavior varies considerably. Furthermore, it is only a slight departure from the widely understood and regularly used method of calculating interobserver agreement percentages. It appears, therefore, to provide a reasonable, conventional method for assessing interobserver agreement when using interval data.

REFERENCES

Baer, 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.

Clement, P. G. A formula for computing interobserver agreement. *Psychological Reports,* 1976, **39,** 257-258.

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

Hopkins, B. L. and Hermann, J. A. Evaluating interobserver reliability of interval data. *Journal of Applied Behavior Analysis,* 1977, **10,** 121-126.

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.

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

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.

Received 13 October 1977.
(Final Acceptance 28 July 1978.)