JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1977, **10**, 127-131 NUMBER 1 (SPRING) 1977

A PROBABILITY-BASED FORMULA FOR CALCULATING INTEROBSERVER AGREEMENT[1]

ANN R. YELTON, BETH G. WILDMAN, AND MARILYN T. ERICKSON

REIDSVILLE, N.C., SCHOOLS, VIRGINIA TREATMENT CENTER FOR CHILDREN, AND VIRGINIA COMMONWEALTH UNIVERSITY

Estimates of observer agreement are necessary to assess the acceptability of interval data. A common method for assessing observer agreement, per cent agreement, includes several major weaknesses and varies as a function of the frequency of behavior recorded and the inclusion or exclusion of agreements on nonoccurrences. Also, agreements that might be expected to occur by chance are not taken into account. An alternative method for assessing observer agreement that determines the exact probability that the obtained number of agreements or better would have occurred by chance is presented and explained. Agreements on both occurrences and nonoccurrences of behavior are considered in the calculation of this probability.

DESCRIPTORS: observer agreement, probability-based formula, recording and measurement techniques, time sampling, reliability

Direct systematic observation of behavior has been a major data-collection technique for the behavioral sciences, especially in the areas of child development and behavior modification. Typically, observers collect these data by categorizing the occurrence or nonoccurrence of specific behaviors over short, fixed intervals of time, usually between 5 and 30 sec in length. To evaluate the replicability of data, observer agreement must be calculated. The per cent agreement method has been employed for ascertaining interobserver agreement. This paper describes the difficulties associated with this method for assessing agreement and suggests an alternative method.

There are three major weaknesses with the use of per cent agreement

$$\left(\frac{\text{number of agreements}}{\text{number of agreements} + \text{number of disagreements}}\right)$$

as a measure of observer agreement: (1) per cent agreement is directly affected by the frequency of behaviors, (2) the decision whether or not to include agreement on nonoccurrences of behavior can drastically affect obtained agreement percentages, and (3) similar agreement percentages do not necessarily reflect the same quality of observer agreement because the number of agreements that may have occurred by chance are not considered. Despite these difficulties, no alternative to this currently popular technique has been proposed.

A positive relationship has been observed between frequency of behavior and per cent agreement such that low and high frequencies are associated with lower and higher observer agreements, respectively (Johnson and Bolstad, 1973; Richards and Irwin, 1936). As the frequency of a behavior recorded for a session increases, so does the number of agreements expected by chance. If more than half of the intervals rated by each observer are scored as occurrences, some agreements would have to occur. For example, if one observer recorded a given behavior eight times, and a second observer recorded the same behavior six times in 10 intervals, they must have agreed at least four times (a minimum of 40% agreement). Because percentage of agreement may vary with frequency, observer agreement would not be stable over time if the frequency of

[1] Reprints may be obtained from Marilyn T. Erickson, Department of Psychology, Virginia Commonwealth University, Richmond, Virginia 23284.

behavior changes. This problem is especially severe in studies in which the frequency of behavior changes substantially across experimental phases.

Behavior can be classified dichotomously into occurrences and nonoccurrences, and categories may be defined as either. Generally, only one of these classifications is chosen as the behavior to be recorded. However, the same information would be contained in data recorded using either classification. Because frequency may affect observer agreement, the choice of which classification to use (occurrences or nonoccurrences) would influence the obtained per cent agreement.

Per cent agreement also may be influenced by how agreements are defined. One definition of agreement includes agreements on nonoccurrences as well as agreements on occurrences of the target behavior. However, the recommended definition is that agreements should be defined solely on the basis of agreements on the occurrence of behavior because occurrences are usually the data of interest (Bijou, Peterson, Harris, Allen, and Johnston, 1969; Johnson and Bolstad, 1973; O'Leary and Kent, 1973.) These data are summarized as either the number of intervals in which the behavior occurred when the number of intervals is held constant, or as the per cent of intervals in which the behavior occurred. Because these measures consider the total number of observation intervals, measures of observer agreement based solely on intervals in which an occurrence was scored are inappropriate. However, when agreements on nonoccurrences are included in the total number of agreements, high agreement percentages may be obtained even when there are few agreements on occurrence.

Probability-Based Assessment of Observer Agreement

An alternative approach to observer agreement would be to compute the probability of two observers agreeing a given number of times. The proposed model, based on probability theory, takes into account the rate of behavior, the definition of agreement, and the adequacy of obtained observer agreement. The formula for determining the agreement probability is:

$$\sum_{Z=A}^{Y} \frac{Y!}{Z!\,(Y-Z)\,!} \times \frac{X!}{(X-Z)\,!} \times \frac{(N-X)\,!}{(\;(N-X)-(Y-Z)\;)\,!} \times \frac{(N-Y)\,!}{N!}$$

where

A = the number of agreements on occurrence obtained,

N = the number of intervals,

X = the number of occurrences recorded by Observer 1, and

Y = the number of occurrences recorded by Observer 2.

The observers must be labelled such that Observer 1 recorded frequency of behavior equal to or greater than did Observer 2 (*i.e.,* $X \geqq Y$). Calculation of this formula may be accomplished by a simple computer program[2] or by using a table of factorials (Beyer, 1968; Pfeiffer and Schum, 1973). This formula assumes that each observer recorded at a constant and set rate, namely the number of behaviors recorded divided by the number of intervals of observation.

The above formula can best be understood by dividing the Y occurrences recorded by Observer 2 into Z agreements and $Y - Z$ disagreements with respect to the X occurrences recorded by Observer 1. The first term in the formula $\left(\frac{Y!}{Z!\,(Y-Z)\,!}\right)$ determines the number of possible ways that Z agreements (or alternatively, the $Y - Z$ disagreements) could be selected from the Y recorded occurrences of Observer 2. The second term $\left(\frac{X!}{(X-Z)\,!}\right)$ then gives the number of ways that the Z agreements could be assigned among the X occurrences recorded by Observer 1, or alternatively the number of ways Z agreements could be selected from Observer 1's X recorded occurrences. The third term $\left(\frac{(N-X)\,!}{(\;(N-X)-(Y-Z)\;)\,!}\right)$ determines the

[2] A copy of the program in Fortran may be obtained from the authors.

number of ways that the $Y - Z$ disagreements could be chosen among the $N - X$ nonoccurrences recorded by Observer 2. The product of these three terms gives the total number of combinations in which the (Y) occurrences recorded by Observer 2 agree exactly Z times with the (X) occurrences recorded by Observer 1. To obtain a probability, the number of combinations of the Y occurrences resulting in exactly Z agreements is divided by $\frac{N!}{(N - Y)!}$ (the reciprocal is the fourth term in the formula), which gives the total number of possible arrangements of Y occurrences in N intervals.

A probability is calculated not only for the obtained number of agreements, A, but also for any possible greater number of agreements. These are summed together so that the formula yields the probability of obtaining at least A agreements when X, Y, and N are given.

The observation protocols for two observers are presented in Table 1. Observer 1 recorded four occurrences in 10 intervals, and Observer 2 recorded three occurrences. The observers agreed on the occurrence of the behavior in two intervals. The probability of obtaining two agreements ($Z = 2$) when $X = 4$, $Y = 3$, and $N = 10$ is

$$\frac{3!}{2!1!} \times \frac{4!}{2!} \times \frac{6!}{5!} \times \frac{7!}{10!} = (3)\,(4 \times 3)\,(6)\frac{1}{10 \times 9 \times 8} = 0.3.$$

Observer 2 recorded three occurrences, which are divided into two agreements and one disagreement with respect to the four occurrences recorded by Observer 1. There are three ways that the two agreements could have been selected from the three occurrences recorded by Observer 2. There are $4 \times 3 = 12$ ways that the two agreements could be assigned to the four occurrences recorded by Observer 2. Note that there are four possible ways in which the first occurrence recorded by Observer 2, which resulted in an agreement, could be paired with an occurrence recorded by Observer 1. Once this occurrence is assigned, there are three ways in which the second occurrence recorded by Observer 2 and simultaneously recorded by Observer 1 could be paired with a remaining occurrence recorded by Observer 1. There are six possible ways that the occurrence scored by Observer 2, and which was not recorded by Observer 1, could be assigned with respect to the six nonoccurrences recorded by Observer 1. The last term, $\frac{1}{10 \times 9 \times 8}$, refers to the number of ways in which the occurrences recorded by Observer 2 could be arranged within 10 intervals. To obtain the probability of obtaining two or more agreements, the probability of getting three agreements also must be calculated because in this example three agreements was a possible arrangement. The result for three agreements would be[3]

$$\frac{3!}{3!0!} \times \frac{4!}{1!} \times \frac{6!}{6!} \times \frac{7!}{10!} = (1) \times (4 \times 3 \times 2) \times \left(\frac{1}{10 \times 9 \times 8}\right) = 0.033.$$

The final results, obtained by adding the two probabilities together, is 0.333. By chance alone, the obtained arrangement or a better one would occur one third of the time.

The probabilities generated by this formula over each of the Y possible number of agreements forms an asymmetric, inverted U-shaped distribution. The formula gives the probability that the obtained number of agreements falls within the upper portion of this distribution bounded on the horizontal axis by the obtained number of agreements. With few agreements, most of the curve will be included and the probability of obtaining the observed number of agreements or better will be high (close to one). With many agreements, only a small tail of the curve will be included, and consequently the probability will be low (close to zero).

Both nonoccurrences and occurrences must be considered in calculating observer agreement in order to assess whether observation protocols re-

Table 1

Sample Observation Protocol

Observer 1	X		X	X			X				
Observer 2	X		X					X			

[3] For computational purposes, the reader is reminded that $0! = 1$.

flect similar external events. Observers must not only agree on the recording of occurrences in response to external events but also must agree on the recording of nonoccurrences. Some disagreements resulting from different definitions of behavior may be attributed to one observer defining an event as an occurrence, while the other observer defines the same event as a nonoccurrence of a given behavior. To obtain a low probability with the probability-based formula, samples of both agreements on occurrence and agreements on nonoccurrences must be included. If, for example, no agreements on nonoccurrences and three disagreements occurred in 10 intervals, the probability of seven agreements would be 1.

Table 2 presents examples that demonstrate various agreement probabilities obtained with the probability formula and the results obtained using per cent agreement based on all agreements and per cent agreement based on occurrence agreements only. In Examples A and B, Observer 1 recorded occurrences in every interval; thus, an agreement resulted for every occurrence recorded by Observer 2, making the probability of these agreements 1.0. In Example A, 50 agreements had to occur. In Example B, although the two observers agreed on occurrences in 99 of 100 intervals, they never agreed on nonoccurrences.

In Example C, the observers agreed on occurrences in eight intervals, agreed on nonoccurrences in one interval, and disagreed in one interval. The two observers disagreed one of two times when a nonoccurrence was scored. The rather high probability of 0.2 reflects the small sampling of nonoccurrences. An investigator who assigns importance to both agreements on occurrences and nonoccurrences of behavior could not conclude that observers agreed on both. If in 40 intervals the observers disagreed one of two times when a nonoccurrence was scored, the probability drops to 0.05 because the agreement on occurrences is more certain (Example D). Similarly, when the number of intervals used is 200, the probability reaches 0.01 (Example E).

In these extreme cases, when nonoccurrences or occurrences are not recorded or are recorded only once by at least one of the observers, the high probability obtained reflects poor sampling of the event. The investigator may conclude that agreement on the well-sampled event was good, but he cannot predict how well the observers

Table 2

Comparison of Formulae Used to Compute Observer Agreement

	Data				*Formulae*		
Examples	*Observer 1*	*Observer 2*	*Intervals*	*Agreements on Occurrence*	*Per Cent Agreement (Occurrences & Nonoccurrences)*	*Per Cent Agreement Occurrences*	*Probability Formula*
A	100	50	100	50	50.0	50.0	1.0
B	100	99	100	99	99.0	99.0	1.0
C	9	8	10	8	90.0	88.9	0.20
D	39	38	40	38	97.5	97.4	0.05
E	199	198	200	198	99.5	99.5	0.01
F	40	30	50	28	72.0	66.7	0.006
G	20	10	50	8	72.0	36.3	0.006
H	6	6	10	6	100.0	100.0	0.005
I	12	12	20	10	80.0	71.4	0.01
J	6	6	10	5	80.0	71.4	0.12
K	41	40	50	36	82.0	80.0	0.01
L	22	20	50	13	68.0	44.8	0.02
M	22	20	50	18	88.0	75.0	0.001
N	40	10	50	10	40.0	33.3	0.08

would have agreed had the event not occurred, and therefore cannot be confident that the observers responded to the same environmental events. Because it is theoretically arbitrary whether an observer has been instructed to record an event or its absence as an occurrence, per cent agreement with nonoccurrence, and the probability formula each are the same for the reciprocal cases of Examples F and G. Even with perfect agreement, there is a slight probability that the agreements occurred by chance (Example H).

Unlike per cent agreement, the probability formula is sensitive to the number of observation intervals. The probability of obtaining at least A agreements in N intervals is higher than the probability of obtaining $2A$ agreements or better in $2N$ intervals. When two observers maintain a moderate rate of agreement over many intervals, the probability that the agreements were due to chance decreases. Examples I and J provide an illustration. When both observers recorded six occurrences and agreed on five of them in 10 intervals (two disagreements), the probability was 0.12. However, when the two observers recorded at the same rate and agreed on the same proportion of occurrences in 20 intervals, the probability was 0.01.

The problem raised is how to handle many intervals of data. Data may either be combined across sessions, or probabilities may be combined for an overall measure of observer agreement. The first consideration to be made when contemplating the pooling of data must be the rates at which the observers recorded occurrences. If for any reason the rates are not similar, the data may not be pooled without inflating the probability that will be obtained, because, as stated above, the probability-based formula assumes that each observer records at a constant rate. Even if there is no reason to believe that the rate of recording occurrences of behavior rate has changed, it is always more conservative to calculate probability over a small number of intervals. With large numbers of intervals, relatively large differences in observer protocols will be significant (*i.e.,* unlikely to be due to chance). If the rate of observer recording may be estimated and the percentage of agreement on occurrence ascertained (*i.e.,* by making sample observations) one may calculate (by computer) many examples with varying numbers of intervals and thus estimate the number of intervals necessary to obtain a given probability. Examples K, L, M, and N are illustrations of values obtained using the probability-based formula.

The major advantage of the probability formula over per cent agreement is that it gives the exact probability of obtaining at least any given number of agreements. Acceptable levels of observer agreement should also be based on the likelihood that these agreements could have occurred by chance, which is the function of the above formula.

REFERENCES

Beyer, W. H. (Ed.) *Handbook of tables for probability and statistics* (2nd. ed.). Cleveland, Ohio: The Chemical Rubber Co., 1968.

Bijou, S. W., Peterson, R. F., Harris, F. R., Allen, K. E., and Johnston, M. S. Methodology for experimental studies of young children in natural settings. *Psychological Record,* 1969, **19**, 177-210.

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, Illinois: Research Press, 1973. Pp. 7-67.

O'Leary, K. D. and Kent, R. Behavior modification for social action: research tactics and problems. In L. A. Hamerlynck, L. C. Handy, and E. J. Mash (Eds.), *Behavior change: methodology, concepts, and practice.* Champaign, Illinois: Research Press, 1973. Pp. 69-96.

Pfeiffer, P. E. and Schum, D. A. *Introduction to applied probability.* New York: Academic Press, 1973.

Richards, T. W. and Irwin, O. C. The use of the clinical method in experimental studies of behavior. *Journal of Abnormal and Social Psychology,* 1936, **30**, 455-561.

Werry, J. S. and Quay, H. C. Observing the classroom behavior of elementary school children. *Exceptional Children,* 1969, **35**, 461-470.

Received 2 September 1975.
(Final acceptance 15 May 1976.)