JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1979, 12, 535-543 NUMBER 4 (WINTER 1979)

BACK TO BASICS: PERCENTAGE AGREEMENT MEASURES ARE ADEQUATE, BUT THERE ARE EASIER WAYS

JOHN C. BIRKIMER AND JOSEPH H. BROWN

UNIVERSITY OF LOUISVILLE

Percentage agreement measures of interobserver agreement or "reliability" have traditionally been used to summarize observer agreement from studies using interval recording, time-sampling, and trial-scoring data collection procedures. Recent articles disagree on whether to continue using these percentage agreement measures, and on which ones to use, and what to do about chance agreements if their use is continued. Much of the disagreement derives from the need to be reasonably certain we do not accept as evidence of true interobserver agreement those agreement levels which are substantially probable as a result of chance observer agreement. The various percentage agreement measures are shown to be adequate to this task, but easier ways are discussed. Tables are given to permit checking to see if obtained disagreements are unlikely due to chance. Particularly important is the discovery of a simple rule that, when met, makes the tables unnecessary. If reliability checks using 50 or more observation occasions produce 10% or fewer disagreements, for behavior rates from 10% through 90%, the agreement achieved is quite improbably the result of chance agreement.

DESCRIPTORS: chance agreement, chance reliability, interobserver agreement, observational data, observational technology, percentage agreement, reliability

THE PROBLEM: ARE PERCENTAGE AGREEMENT MEASURES ADEQUATE?

Kelly (1977) surveyed all research published in the *Journal of Applied Behavior Analysis* from 1968 through 1975 and found that the majority of studies involved observational data as opposed to mechanically collected or permanent-product data. Most of these studies used trial scoring, interval recording, or momentary time-sampling to record data and summarized interobserver agreement or reliability via the calculation of some measure of percentage agreement. Percentage agreement measures used included point-by-point reliability, often referred to as interval-by-interval or as moment-by-moment reliability, based on the proportion of all observation occasions for which the two observers agree regarding whether or not target behavior occurred. (This proportion converted to a percentage is the reliability measure.) Percentage agreement measures used also included occurrence reliability, the ratio of the number of occasions both observers agree the behavior occurred to the sum of those occasions plus occasions on which they disagreed (converted to a percentage), and nonoccurrence reliability, the ratio of agreements regarding the nonoccurrence of target behavior to those agreements plus disagreements.[1]

Recent articles dealing with the percentage agreement measures disagree on whether we should continue using them (Baer, 1977*a;* Hopkins and Hermann, 1977) or not (Hartmann, 1977; Yelton, Wildman, and Erickson, 1977);

Reprints of this article are available from John C. Birkimer, Department of Psychology, University of Louisville, Louisville, Kentucky 40208.

[1] Other methods of calculating reliability exist (Kelly, 1977). This paper addresses issues involving the most frequently used reliability calculation procedures for the most frequently used recording procedures.

whether, if any are used, all three should be presented (Hopkins and Hermann, 1977) or not (Hawkins and Dotson, 1975); how best to protect against accepting, as evidence of true observer agreement, levels which might have resulted from "chance" or random agreement (Hartmann, 1977; Hopkins and Hermann, 1977; Kratochwill and Wetzel, 1977; Yelton *et al.*, 1977).

We believe there are only two major problems regarding the three percentage agreement measures raised in those articles. The first problem is that, under conditions of low reported rates of target behavior, researchers and consumers must be aware of the magnitude of observers' disagreements relative to their agreements on occurrences and, when reported rates are high, they must be aware of the magnitude of disagreements relative to agreements on nonoccurrences. In Birkimer and Brown (1979), we showed that comparing the percentage of disagreements to the percentage of agreements on occurrences and to the percentage of agreements on nonoccurrences is simpler and easier than calculating the three reliability percentages, though they are adequate, and presented a procedure for doing this graphically. That procedure summarizes all the reliability data usually collected and has the added advantage of helping to identify apparent experimental effects which are not great enough to be trustworthy.

The second problem raised by the articles cited above is the problem of accepting "chance" agreement. We need a procedure to help researchers and consumers be certain that obtained interobserver agreement is not likely the result of chance agreement. This paper supplies three such procedures. The procedures are based on our belief, shared by Baer (1977*a*) and at least partially by Kratochwill and Wetzel (1977) that such procedures should be as little removed from past practices, require as little statistical sophistication, and be as closely related to our primary data as possible. (Such procedures are more likely to be adopted, understood, and useful.)

PROTECTING AGAINST CHANCE AGREEMENT

Chance agreements on occurrences, chance disagreements, and chance agreements on nonoccurrences would occur if two observers responded randomly during any reliability check. Formulas for calculating the chance probabilities for each, and the resulting chance probabilities of each of the three percentage agreement reliabilities, are given in Birkimer and Brown (1979).[2] These formulas, however, give the mean value for each which would occur by chance; for each there exists a theoretical distribution of values which *could* occur if observers responded randomly. (In fact, chance responding could produce perfect agreement, and would on some percentage of occasions.) The question that must be answered, then, for any degree of obtained observer agreement, however summarized, is not "how much better than chance" it is, but rather "how likely is this agreement if observers are really responding randomly?"

The issue is that *any* obtained disagreement rate, occurrence agreement rate, and nonoccurrence agreement rate, and thus *any* of their combinations in percentage agreement measures, have some real, determinable probability of occurring if observers are responding randomly. We must, then, as researchers and as consumers, protect ourselves and others from accepting as evidence of observer agreement levels of obtained agreement which would be likely as a result of chance agreements. (Hopkins and Hermann's suggestion that the calculated "chance" reliabilities be taken as minimum values which

[2] Hopkins and Hermann's formulas for chance occurrence and nonoccurrence reliabilities are not totally correct. The appropriate denominator for chance occurrence reliability is the chance occurrence agreement rate plus the chance disagreement rate, and the denominator for chance nonoccurrence reliability is the chance nonoccurrence rate plus the chance disagreement rate. Neither is appropriately T^2 as they suggested. The formulas we presented are theirs with this correction.

obtained values must exceed only guarantees that values accepted would be in the top half of the distribution of values possible by chance. Such values would occur by chance 50% of the time, so their suggestion is scant protection against accepting chance agreement as evidence of true agreement.)

Hopkins and Hermann's statement that larger sample sizes make any obtained level of agreement less probable by chance is true but is not a criticism of probability generating statistics. Such statements are descriptive of the way events in the environment occur. Of coins tossed in the air, 100% will land "heads" 1 time in 2 if only one coin is tossed, but only 1 time in 1,024 if ten coins are tossed. Any valid probability generating formula must reflect the fact that increasing sample sizes reduces the probabilities of discrepant outcomes.

We are not arguing here for statistical significance tests as they are typically used, to identify experimental effects. The experimental designs currently in use admirably demonstrate such effects in the single subject. The logic of our recommendation is, however, similar to that underlying such tests. When we obtain from a reliability check two observers' occurrence agreement rate, disagreement rate, and nonoccurrence rate (and, perhaps, convert these to percentage agreement measures) there is always the possibility that observers responded randomly (our values occurred by chance) and the possibility that they did not (our values reflect true observer agreement). If we determine the probability of our obtained values, assuming they resulted from chance responding, and we find that probability is trivially low, then we can reject the assumption they resulted from chance and conclude that our measures show true observer agreement.

A formula provided by Yelton *et al.* (1977), a useful adaptation of Fisher's Exact Probability Test (Siegel, 1956), permits calculation of the probability of any obtained level of occurrence agreement having resulted from chance or random observer responding. We do not agree with their suggestion that the probability itself should be used as the primary measure of observer agreement; the probability value does not compare disagreements to agreements on occurrences and to agreements on nonoccurrences, can be reasonable when observers disagree on many more than 10% of observation occasions, and is too far removed from basic data to be a preferred summary of observer agreement. The formula is quite valuable, however, because it can be used to protect against accepting as true those levels of observer agreement which may have resulted from random observer responding.

The Yelton *et al.* (1977) formula, given the number of observation occasions and the two observers' reported rates of target behavior, can be used to calculate the probability of any obtained number of agreements on occurrences. For any given number of observation occasions (N) and observer one and observer two reported numbers of target behaviors (X and Y), however, the number of disagreements (and the number of agreements on nonoccurrences) is determined once the number of agreements on occurrences (Z) is known. (Given N, X, and Y, with $X \geqq Y$, and Z agreements on occurrences, there will be $D = X + Y - 2Z$ disagreements and $N - D - Z$ agreements on nonoccurrences.) What the Yelton *et al.* formula actually calculates is the probability of a particular number of agreements on occurrences, disagreements, and agreements on nonoccurrences, given N, X, and Y. This has a very useful implication: once the chance probability of a given level of agreement on occurrences is calculated, the probability of the resulting numbers of disagreements and of agreements on nonoccurrences is identical. That probability is also the probability of each of the three percentage agreement reliabilities which could be calculated from these agreements and disagreements. Thus, if we determine the probability of any one of these six measures occurring as a result of chance, and find it improbable, we know all our measures are equally improbably the result of chance.

Tables 1 and 2 were developed by applying the Yelton *et al.* formula to the various X and Y reported occurrences which would produce median (or, since they are the same, mean) reported rates of target behavior of 5%, 10%, 20%, and so forth through 95%. This was done for sample sizes (or numbers of observation occasions) of 20, 50, and 100.

The formula was applied first, for each sample size and median reported rate of occurrences, to values of $X = Y$, to find the number of occurrence agreements which were "acceptably improbable" as a result of chance. "Acceptably improbable" was defined as $p \leqq .05$ in Table 1, a result which would occur by chance one time in 20, and $p \leqq .01$ in Table 2, a result which would occur by chance one time in 100. We then applied the formula to values of $X = Y + 1$, then to $X = Y + 2$, and so forth until Y was enough less than X to represent observers disagreeing by 10% on their reported rates of target behavior ($X = Y + 2$ for $N = 20$, $X = Y + 5$ for $N = 50$, and $X = Y + 10$ for $N = 100$). In every calculation, the Yelton *et al.* formula was first applied to the case of perfect agreement on occurrences, then to each succeeding level of less-than-perfect agreement. The resulting probabilities were summed, from the perfect agreement case down, so the lowest level of occurrence agreement which was still "acceptably improbable" by chance was found.

Finally, since Y's differing from X by odd integers produce medians not quite equal to the table heading values, we calculated these for the cases where the medians would be just closer to 50% than the table heading values, referred to as "more central $X = Y + 1$, $Y + 3$, etc.," and for the cases where the medians would be just further from 50% than the table heading values, "less central $X = Y + 1$, $Y + 3$, etc."

Thus, for N = 20 and Median Reported Percentage of Target Behavior of 20%, $X = Y + 1$ "more central" involves $X = 5$ and $Y = 4$ for a true Median Reported Percentage of 22.5%, with $X = Y + 1$ "less central" involving $X = 4$ and $Y = 3$ for a Percentage of 17.5%. For $N = 50$ the more central Percentages are 1% nearer 50% with the less central 1% more distant and for $N = 100$ the more central and less central Percentages are .5% nearer and further from 50%, respectively. In the tables *less central* and *more central* are abbreviated l.c. and m.c., respectively.

We converted from occurrence agreements to disagreements after calculating but before constructing the tables. The tables are read as follows: The median (or mean) percentages of target behavior are listed across the top. (Note that we show percentages greater than 50% in parentheses. The full table is symmetric around 50%, so greater percentages are read back from 50% to the left.) The three sample sizes are listed down the left. Table entries are the number of acceptable disagreements. This number of disagreements is improbable enough as a result of chance to be acceptable evidence of true observer agreement, with improbable defined as occurring one time or fewer in 20 for Table 1, and as one time or fewer in 100 for Table 2. (We prefer Table 2, in keeping with Baer's [1977*b*] urging to try to avoid "Type 1" errors, accepting as true those results which are not.) Occasionally the probabilities for a number of disagreements produced by X and Y values change enough, as these vary around $X = Y$, to make acceptable one more disagreement, or to make the generally acceptable number unacceptable by one. These cases are included in the table with a brief explanation of when they apply, with the unexplained entry applying in all other cases.

Procedure 1

Researchers could use Table 2 to guarantee that they accept agreement rates as true evidence of agreement only when those rates are great enough to be improbable as a result of chance. A conservative usage would be to check obtained numbers of disagreements against the acceptable number for the tabled sample size equal to or just smaller than one's own sample size. Thus, true sample sizes *below* 50 but above 20 would

Table 1

Acceptable ($p \leq .05$) numbers of disagreements for various median (or mean) reported percentages of target behavior, for $N = 20$, $N = 50$, and $N = 100$ observation occasions. Table applies from case where X (one observer's reported percentage of behavior) $= Y$ (other observer's percentage) through case where X and Y differ by 10%. Unexplained number of disagreements applies in all cases, except those listed for explained number. See text for explanation of m.c. (more central) and l.c. (less central).

N	*5% (95%)*	*10% (90%)*	*20% (80%)*	*30% (70%)*	*40% (60%)*	*50%*
20	None 0 when X=Y	0	2 when l.c. X=Y+1 3	4	4 when l.c. X=Y+1 5	5
50	None 0 when X=Y+1 and m.c. X=Y+2	5 6 when X=Y+4	10 when l.c. X=Y+1 or Y+3 11	14 when l.c. X=Y+1 or Y+3 15	17	18
100	6 when l.c. X=Y+1, Y+3, or Y+5 7	13 14 when X=Y+6 Y+8, or Y+10 15 when m.c. X=Y+9	25 26 when X=Y+8 or Y+10	34	38 39 when m.c. X=Y+9	40 41 when X=Y+7 or Y+9

Table 2

Acceptable ($p \leq .01$) numbers of disagreements for various median (or mean) reported percentages of target behavior, for $N = 20$, $N = 50$, and $N = 100$ observation occasions. Table applies from case where X (one observer's reported percentage of behavior) $= Y$ (other observer's percentage) through case where X and Y differ by 10%. Unexplained number of disagreements applies in all cases, except those listed for explained number. See text for explanation of m.c. (more central) and l.c. (less central).

N	*5% (95%)*	*10% (90%)*	*20% (80%)*	*30% (70%)*	*40% (60%)*	*50%*
20	None	0	1 2 when X=Y+2	2 when l.c. X=Y+1 3	3	3 4 when X=Y+2
50	None 0 when X=Y+1 or mc. X=Y+2	4 when l.c. X=Y+1, Y+3, or Y+5 5	9	12 when l.c. X=Y+1, Y+3 13	14 when l.c. X=Y+1 15	15 16 when X=Y+4
100	4 when lc. X=Y+1 5 6 when X=Y+6 7 when mc. X=Y+7	12 13 when m.c. X=Y+5 through Y+9	22 when l.c. X=Y+1, Y=3, or Y+5 23	30 when l.c. X=Y+1, Y+3, or Y+5 31	35	37 38 when X=Y+10

be checked against the $N = 20$ limits. A similarly conservative procedure would be to require the number of disagreements considered acceptable, for median percentage behaviors between table heading values, to be the smaller of the numbers shown for the two nearest table heading percentages. Thus, a median percentage behavior of 37% for $N = 50$ would require the tabled 13 or fewer disagreements for 30% rather than the tabled 15 or fewer for 40%. Alternatively, the trends in the tables appear linear enough that linear interpolation between N's and between Median Reported Percentage Target Behavior values would not be misleading.

The chance probability of any particular number of disagreements, for given sample size and reported rates of target behavior, is identical to the chance probability for the resulting numbers of occurrence agreements and for the resulting numbers of nonoccurrence agreements, and is thus identical to the chance probabilities of any of the percentage agreement statistics which are based on them. Researchers, then, can check the chance probability of their obtained disagreement percentages to be sure their point-by-point, occurrence, and/or nonoccurrence reliabilities are acceptably unlikely the result of chance. This is one procedure to solve the problem of accepting "chance" agreement. The percentage agreement measures are, then, adequate for assuring that obtained levels of agreement are unlikely the result of chance.

Procedure 2: An Easier Way

A second procedure to solve the problem of chance agreements, an easier way, is to follow our earlier recommendation to present disagreements, agreements on occurrences, and agreements on nonoccurrences directly and to check directly the chance probability of obtained disagreements in Table 2. Only if obtained disagreements, for a given sample size and reported rate of target behavior, are acceptable by Table 1 or 2 would the agreement be taken as evidence of true observer agreement.

Procedure 3: The Easiest Way—50-10-10 (90) Rule

Fortunately, there is an even easier way to protect against the acceptance of agreement levels that are too probable by chance. Figure 1 shows the Table 2 acceptable disagreement values plotted for each sample size. The left-hand vertical axis is the percentage of all observation occasions, with the right-hand axis showing acceptable disagreement percentages. These are graphed, down from the top of the figure, so agreements of both sorts fall below the function, and acceptable disagreement percentages lie above it.

The figure leads to several conclusions. Note that acceptable disagreement percentages for $N = 20$ are zero (agreements totaling 100%) with response rates of 10% and 90%, and are 5% (agreements totaling 95%) for response rates of 20% and 80%. The sample size of 20, then, does not easily permit proof of true observer agreement for response rates below 20-30% or above 70-80%. Sample sizes of 50 and 100, on the other hand, permit substantial and very similar acceptable disagreement percent-

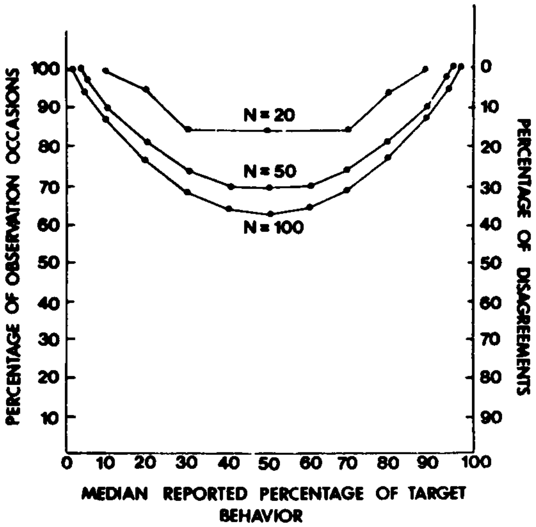

Fig. 1. Acceptable ($p \leq .01$) agreement percentages and disagreement percentages for various median reported percentages of target behavior and for various numbers of observation occasions.

ages. The figure indicates, then, that increasing sample size beyond 50 observation occasions during reliability checks has only a small effect on the observer agreement needed to demonstrate true agreement.

A less comforting implication of Figure 1 is that, with very low or very high response rates, very little disagreement may occur and still permit concluding true observer agreement has been shown. (Increasing sample size helps here, but, as the figure shows, with diminishing returns.)

The most valuable conclusion from Figure 1 is that disagreement percentages of 10% or less are clearly improbable enough to be evidence of true observer agreement, if sample sizes are 50 or greater, for reported response rates from 10% to 90%. This is a substantial and important discovery, for it can serve as a simple rule and make consulting Table 2 unnecessary, in most studies. We refer to this third solution to the problem of chance agreement as "the 50-10-10 (90) rule." If researchers consistently include at least 50 observation occasions in all reliability checks, obtain disagreement percentages of 10% or less, and have reported rates of target behavior between 10% and 90%, then they and consumers can conclude the obtained observer agreement is evidence of true observer agreement.

Suppose a researcher obtains reported target behavior percentages of 38% and 42% from two observers during a reliability check, with 4 disagreements, and there were 50 observation occasions during the recording session (it matters not whether these were intervals, moments, or trials). Following the first procedure we propose, the researcher calculates any or all of the percentage agreement reliabilities, checks the number of disagreements in Table 2, finds that for the Median Reported Percentage of Target Behavior of 40% with an N of 50, 4 disagreements are many fewer than the minimally ($p \leq .01$) acceptable 15, so concludes he/she has obtained true (non-chance) observer agreement. Following the second (easier) procedure we described, the researcher would not calculate the reliability percentages but simply show the percentages of agreements of both sorts and disagreements, probably graphically, as Birkimer and Brown (1979) recommended, and check the number of disagreements in Table 2, reaching the same conclusion as above. Following the third (easiest) procedure, using the 50-10-10 (90) rule, the researcher would simply note that there were 50 observation occasions, fewer than 10% disagreements, and a reported median behavior percentage between 10% and 90%, so correctly conclude that true observer agreement has been shown. (Note that checking Table 2 was unnecessary.)

A second researcher obtains a median reported percentage of target behavior of 88% over 50 observation occasions, but obtains 6 disagreements (12%) from the observers. He/she calculates (procedure 1) the reliability percentages and checks the 6 disagreements in Table 2, but finds that, while 9 would be acceptable for a behavior percentage of 80%, only 5 are acceptable for 90%. Following the "conservative" procedure we described he/she would evaluate the 6 against the acceptable 5 at 90% and conclude true observer agreement was not quite demonstrated. Using, instead, linear interpolation between the 9 acceptable at 80% and the 5 at 90%, the researcher would see that acceptable disagreements decrease by 4 from 80% to 90%, or by 4/10 for each one percent, so would be 5.8 for 88%, permitting the conclusion that he/she had shown (barely) that true (non-chance) observer agreement existed. Following our procedure 2, this researcher would calculate the percentages of agreements of both sorts and disagreements, proceed as in procedure 1 with Table 2, and reach the same conclusions. (Since this reliability check produced 12% disagreements, it violates the 50-10-10 (90) rule and forces the use of Table 2, as would use of fewer than 50 observation occasions or obtaining behavior percentages below 10% or greater than 90%.)

In Birkimer and Brown (1979) we recommended (Recommendation 4) graphing the limits of what we called "the chance disagreement range" around the median of the two observers' median reported percentage of target behavior. This was to permit comparison to the similarly graphed obtained "disagreement range," to show obtained disagreement was less than expected by chance. Our recommended procedures here are superior, in handling the problem of chance agreements, since these involve precise criteria ($p \leqq .01$) for determining that obtained agreement is unlikely the result of chance agreements. Our first three recommendations in that paper speak to other issues and still stand. Conveniently, the graphing of the disagreement range recommended there, along with a statement of the number of observation occasions, permits rapid determination of whether or not the 50-10-10 (90) rule has been met.

Meeting the "10% or fewer disagreements" criteria in the 50-10-10 (90) rule is conservative in that, with sample sizes greater than 50 and behavior percentages between 10% and 90%, disagreement percentages considerably greater than 10% are often acceptably nonchance. However, in Birkimer and Brown (1979) we showed, essentially, that claimed experimentally produced changes in target behavior must be greater than the percentage of disagreements obtained in reliability checking in order to be strong evidence of treatment effects. Meeting the 10% disagreement criterion would, then, assure that rate changes of 10% or greater would be believable. In general, acceptable agreement must permit demonstration of believable experimental effects, as addressed in our earlier paper, and be acceptably unlikely as a result of chance, as better addressed here.

In conclusion, then, complex statistical measures of observer agreement are unnecessary for interval recording, momentary time-sampling, and trial scoring data. Percentage agreement measures adequately compare disagreement percentages to the percentages of agreements on occurrences and on nonoccurrences and, if disagreements are checked in Table 2 for their "acceptability," can be shown to be improbable as a result of chance agreements, thus evidence for true agreement. Easier ways exist and are recommended.

RECOMMENDATIONS

1. Researchers collecting data by interval recording, momentary time-sampling, or trial scoring should, if they calculate the various reliability percentages, check their obtained number of disagreements in Table 2 to be sure their reliabilities are acceptably unlikely the result of chance agreement. (Procedure 1)
2. As an easier procedure, researchers can simply present their obtained percentages of agreements on occurrences, of agreements on nonoccurrences, and of disagreements, and check their obtained number of disagreements in Table 2 to assure acceptably non-chance agreement. (Procedure 2)
3. A third, still easier, option is to use routinely at least 50 observation occasions. If 10% or fewer disagreements are obtained, for median reported behavior percentages from 10% to 90%, then acceptable evidence of true (nonchance) agreement has been obtained. (Procedure 3; the 50-10-10 (90) rule. When the rule is violated, researchers should revert to procedure 2.)

REFERENCES

Baer, D. M. Reviewer's comments: Just because it's reliable doesn't mean that you can use it. *Journal of Applied Behavior Analysis*, 1977, **10**, 117-119. (*a*)

Baer, D. M. "Perhaps it would be better not to know everything." *Journal of Applied Behavior Analysis*, 1977, **10**, 167-172. (*b*)

Birkimer, J. C. and Brown, J. H. A graphical judgmental aid which summarizes obtained and chance reliability data and helps assess the believability of experimental effects. *Journal of Applied Behavior Analysis*, 1979, **12**, 523-533.

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

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.

Hopkins, B. L. and Hermann, J. A. Evaluating interobserver agreement 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 judgment: Some considerations in presenting observer agreement data. *Journal of Applied Behavior Analysis,* 1977, **10**, 133-139.

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

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

Received 31 May 1978.
(Final Acceptance 22 January 1979.)