**PROPOSED CONVENTIONS FOR
EVALUATING OBSERVER RELIABILITY:
*A Commentary on Two Articles by Birkimer and Brown***

B. L. HOPKINS

UNIVERSITY OF KANSAS

The paper, "A Graphical Judgmental Aid . . .," (Birkimer and Brown, 1979a) proposes the adoption of conventions that I believe will be useful to the field of applied behavior analysis. I recommend that the conventions be adopted with certain precautions. I wish to take the liberty of restating the recommended conventions in a cookbook fashion in hopes this will facilitate adoption.

Birkimer and Brown address research in which data are collected on the percentage of intervals or percentage of trials in which some response is recorded as occurring by a human observer. They are concerned with interpretations of agreements and disagreements between two observers whenever interobserver reliability checks are conducted, and with the implications those interpretations have for conclusions about experimental effects.

Birkimer and Brown recommend that:

1. graphic presentations of the data include points representing the data of both observers for days on which reliability checks are conducted;
2. the following index of reliability be calculated:

$$\text{Percentage of Disagreements} = \frac{\text{Number of Disagreements}}{\text{Number of Intervals or Trials}} \times 100$$

3. the percentage of disagreements be plotted on the graphic presentation as a vertical band

centered about the mean of the two observers' data points.

The authors note that this band defines, on that day, a range in which the observers disagree about the occurrence of the response. Below the band lies the percentage of intervals or trials on which the observers agree that the response occurred. Above the percentage of disagreement band lies the percentage of intervals or trials on which the two observers agree that the response did not occur.

Birkimer and Brown further recommend that:

4. the following chance disagreement percentage be calculated:

$$\text{Chance Percentage Disagreement} = \frac{(O_1 \times N_2) + (N_1 \times O_2)}{T^2} \times 100$$

where one observer records the response as occurring in O_1 of the intervals and not occurring in N_1 of the intervals, and the second observer records the response as occurring in O_2 intervals and not occurring in N_2 intervals. T is the total number of intervals.

5. the chance percentage disagreement be plotted as a vertical range, also centered about the mean of the two data points.

The chance disagreement range is the percentage of intervals on which the two observers would be expected to disagree if they randomly recorded the response to occur and not occur the indicated number of times. Below the chance

Requests for reprints should be addressed to B. L. Hopkins, Department of Human Development, University of Kansas, Lawrence, Kansas 66045.

disagreement range lies the chance agreement on occurrences range. The chance agreement on nonoccurrences range lies above the chance disagreement range.

Graphing the chance disagreement and agreements ranges allows for easy comparisons between the obtained percentages and those expected by chance. Hopefully the obtained disagreement percentage will be smaller than that expected by chance and the obtained agreement percentages greater than those expected by chance.

Finally, Birkimer and Brown recommend that:

6. there be no overlap between the disagreement range(s) obtained under one condition and that (those) obtained under a different condition before concluding that there is an experimental effect.

These recommended conventions in "A Graphical Judgmental Aid . . ." are simple and elegant. One need not calculate all of the obtained and chance percentages of agreement and present them in tabular form as recommended by Hopkins and Hermann (1977). Instead, the percentage disagreement and chance percentage disagreement can be calculated and plotted as described. The result will include all of the information contained in the convention recommended by Hopkins and Hermann, plotted in a form that makes visual inspections easy.

Much of the attendant discussion in "A Graphical Judgmental Aid . . ." is confusing. Birkimer and Brown's discussion of scored interval reliability, unscored interval reliability, chance scored interval reliability, chance unscored interval reliability, disagreements on scored intervals, and disagreements on unscored intervals is extraneous to the recommendations.

Birkimer and Brown's argument that their formulas for computing chance probabilities for agreements on scored and unscored intervals are superior to those presented by Hopkins and Hermann (1977) is specious. The Hopkins and Hermann formulas address the number of agree-

ments on occurrences and nonoccurrences relative to the total number of observation intervals. The Birkimer and Brown formulas provide the number of agreements on occurrences relative to the number of intervals in which either of the observers record the response as occurring and the probability of agreements on nonoccurrences relative to the number of intervals in which either of the observers record the response as not occurring. The two sets of formulas are clearly different. However, both sets are clearly interpretable, as are a number of other possible measures such as phi and kappa (see Hartmann, 1977).

In any case, the reader should understand that the conventions recommended by Birkimer and Brown involve the percentages of agreement and chance percentages of agreement defined by Hopkins and Hermann. The agreement on occurrences range and agreement on nonoccurrences range resulting from the Birkimer and Brown conventions are calculated by using the Hopkins and Hermann formulas for occurrence and nonoccurrence reliabilities, respectively. The chance agreements on occurrences and chance agreements on nonoccurrences ranges resulting from the Birkimer and Brown conventions can be calculated by using the chance occurrence reliability and chance nonoccurrence reliability formulas presented by Hopkins and Hermann, not by the Birkimer and Brown formulas for chance scored-interval reliability and chance unscored-interval reliability.

Adoption of the recommended conventions might protect the researcher and readers from concluding that some experimental effect had occurred when apparent differences in data might reasonably be attributed to variations in observers' behaviors. Although I see these conventions as being useful to applied behavior analysis, I strongly recommend that a certain precaution be employed along with them.

Baer, Wolf, and Risley (1968) argued that a distinguishing characteristic of applied behavior analysis is a concern with producing results that are socially important. Concluding that an ap-

parent change in responding is attributable to something other than perturbations of observers' behaviors should not be interpreted as guaranteeing that the change is socially important.

The concern to produce socially important results places the applied behavior analyst in a particular obligatory position with respect to consumers. The position is different than that occupied by many social sciences. The position amounts to accepting a contract to produce behavior change, but not just any behavior change. The contract states, "Produce at least an X change in behavior at a cost not to exceed Y , given current standards of measurement."¹ Costs should be considered in the most general sense to include response and other costs as well as dollars.

Such contracts have generally been implicit. Certainly, they are not regularly reported in *JABA*. However, they will surely become explicit as we and our consumers better learn to quantify our reinforcers and results. As this occurs, the Birkimer and Brown convention regarding conclusions about changes of responding can be modified. The modified convention, as applied to a contract with a consumer, might read, "The disagreement ranges during treatment conditions will lie beyond X ."

For now, the Birkimer and Brown recommendation regarding conclusions about changes in responding might best be considered a necessary but not sufficient test. All reported data should pass the test. However, passing the test in no way substantiates the social significance of the effect. All reported changes in responding should further be submitted to our implicit and imprecise tests for social importance.

It could be argued that the Birkimer and Brown test would be unnecessary if our tests for social importance were adequate. That is probably correct. However, the tests for social importance are still poor. I have just scanned sev-

eral back issues of *JABA* and have found at least five sets of data that probably would not pass the Birkimer and Brown test. In other words, we have published, as experimental effects, changes in data that could reasonably be attributed to unreliable observations. We should protect ourselves and our consumers from such errors.

"Back to Basics . . ." (Birkimer and Brown, 1979b) offers three simple methods that approximate the formula published by Yelton, Wildman, and Erickson (1977). Their formula could be used to determine if an obtained set of agreements between two observers is statistically reliably better than would be expected by chance. The approximations could be used if a researcher were interested in that question and didn't have a convenient computer or didn't care to deal with very large numbers.

I doubt that the question is a useful one. However, I also doubt that statisticians and some applied behavior analysts who have statistical repertoires will be quickly convinced of the lack of utility.

Baer (1977) argued that applied behavior analysis is not concerned with small effects; and that statistical inference, therefore, is not useful. The strongest form of his argument was that the use of statistical inference could be harmful to the field by causing researchers and readers to accept, as important, effects that are statistically reliable but socially unimportant. These arguments apply to the "Back to Basics . . ." recommendations.

Hopkins and Hermann (1977) pointed out that statistical inference procedures, rather than providing objective decisions, in fact require subjective judgments about what are acceptable alpha-levels (or levels of significance or confidence intervals, if you prefer) given various sample sizes. Suppose you are just beginning some research on a response that observers are recording as occurring in about 50% of the 20 observation intervals you program each day. You conduct your first reliability check and find that the observers disagree on 7 (35%) of the

¹This argument, in various trappings, must have been made thousands of times. However, I first heard it stated in this form by Gordon Fitch and John Hanna.

intervals. Is that a satisfactory level of agreement? You might consult Table 1 in "Back to Basics . . ." and conclude that it is not.

Given this unhappy state of affairs and the prospects of not having your eventual paper accepted for publication, you might change something. Need you rewrite your observational definitions or better train your observers? Not if you follow the "Back to Basics . . ." recommendation. Looking down the 50% column of Table 1, you see that 35% disagreement is barely acceptable if you simply collect data for 50 intervals or, to be on the safe side, you elect to program 100 intervals for which 40% disagreement is acceptable. The methods offered in "Back to Basics . . ." can provide a means to make relatively poor observational reliability statistically acceptable without improvements in observation technology.

Practical considerations must also apply to decisions about the acceptability of observer reliability. It seems reasonable to ask that such reliability pass the test described by Birkimer and Brown in "A Graphical Judgmental Aid. . . ." This might again be considered as a necessary but not sufficient requirement for acceptability. Standards of acceptability will be different in different areas of research. The standards tests are certainly still imprecise, and the standards will change. However, we would seem

to have more to gain by spending our research efforts on improving our observation technology than by simply programming more observation intervals to produce some arbitrary statistical significance.

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.
- Baer, D. M., Wolf, M. M., and Risley, T. R. Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, 1968, **1**, 91-97.
- 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. (a)
- Birkimer, J. C. and Brown, J. H. Back to basics: Percentage agreement measures are adequate, but there are easier ways. *Journal of Applied Behavior Analysis*, 1979, **12**, 535-543. (b)
- 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.
- 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 2 April 1979.

(Final Acceptance 20 July 1979.)