ON THE NOT SO RECENT INVENTION
OF INTEROBSERVER RELIABILITY STATISTICS:
A Commentary on Two Articles by Birkimer and Brown

DONALD P. HARTMANN AND WILLIAM GARDNER

UNIVERSITY OF UTAH

The two articles by Birkimer and Brown (1979a, 1979b) are interesting additions to the bibliography of papers concerned with interobserver reliability published in the *Journal of Applied Behavior Analysis*. The growing length of this collection attests to the continuing interest of applied behavioral researchers in assessing and improving the quality of observational data. Unfortunately, by failing to reference or discuss parallel material outside the behavioral tradition, the authors tend to perpetuate the myth that scholarly concern with interobserver reliability coincides with the history of applied behavior analysis. An examination of the historical antecedents of the principal topics discussed by Birkimer and Brown (1979a, 1979b) may help to counter this myth. The topic of their first paper (1979a), methods of measuring interobserver reliability, has resurfaced with some regularity at least since the early 1940's (see Fleiss, 1975, for a review of methods of measuring agreement between two judges for occurrence-nonoccurrence data). The topic of their second paper (1979b), assessing the statistical significance of two-by-two tabled data, is based directly on the Exact Probability Test which was described by R. A. Fisher in 1934. For both topics dealt with by Birkimer and Brown, scholarly interest clearly precedes the inauguration of the *Journal of Applied Behavior Analysis*. But enough of history.

The first article by Birkimer and Brown (1979a) makes two solid contributions that

seem unlikely to arouse controversy among *JABA* readers. The first contribution is their elegant graphic method of distinguishing between the three agreement statistics commonly applied to two-by-two table reliability data. The second contribution is their recommendation to present reliability data from both observers and the disagreement range directly on the primary data display (see Figure 1 in Birkimer and Brown, 1979a).¹

The authors also recommend the use of the disagreement range to establish the "believability" of experimental effects. We feel that their specific proposal is not likely to improve the evaluation of experimental results by applied behavioral researchers. We suggest that it may lead to unacceptable levels of Type II error.

The authors state that "If overlap exists between these disagreement ranges, the apparent change in target behavior is not enough, given the degree of observer disagreements, to permit certainty that an effect of treatment has been shown." This statement is not informative since, as the authors are well aware, there is no procedure that does permit certainty that an effect of treatment has been shown. Since the evaluation of experimental results always yields a decision with only a likelihood of certainty, and with resulting costs in Type I or Type II error, it is critical that we use valid means of estimating the confidence we can place in our decision. Such

Reprints may be obtained from Donald P. Hartmann, Department of Psychology, University of Utah, Salt Lake City, Utah 84112.

¹The upper bound of the disagreement range is equal to the mean ($\bar{X}$) of the two observers' rate of the target behavior plus one half their number or percent of disagreements (.5D). The lower limit of the disagreement range is equal to $\bar{X} - .5D$.

decisions involve judgments concerning the reliability, level, slope, and variability exhibited in the data. This is true whether we perform a calculation or use some form of visual inspection of the data. A criterion based solely upon the reliability of the data cannot do justice to the complexity of the decision.

It seems likely to us that the adoption of Birkimer and Brown's criterion would lead to an unwarranted increase in Type II errors. Consider the application of this recommendation to an investigation in which the range of target behaviors during both baseline and treatment phases was 20 percentage points and in which agreement reliability was 80% (the latter figure corresponds to a disagreement range of 20%). Given these not uncommon circumstances, the difference between phases may have to exceed 40 percentage points (e.g., a change in rate of the target behavior from 30% to 70%) in order to be judged believable!

In their second article, Birkimer and Brown (1979b) recommend the use of significance tests to aid in evaluating two-by-two interobserver reliability data, and develop three simple methods for assessing statistical significance for these data including their clever 50-10-10 (90) rule. Some *JABA* readers with an antistatistical bent may object to the authors' recommendation to use statistical devices as a judgmental aid. Other readers, like ourselves, may approve of their recommendation, but object to the space devoted to developing "easy methods" for conducting a statistical test that already has a widely

used and simple solution. The test of significance suggested by Birkimer and Brown (1979b) can be closely approximated by the χ^2 test of independence for two-by-two tables with as few as 20 entries (Camilli and Hopkins, 1978), the smallest sized table discussed by the authors.² The test is described in most introductory statistics books, and is easily performed by hand or with the aid of an inexpensive calculator.

REFERENCES

- 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)
- Camilli, G. and Hopkins, K. D. Applicability of chi-square to 2×2 contingency tables with small expected frequencies. *Psychological Bulletin*, 1978, **85**, 163-167.
- Fisher, R. A. *Statistical methods for research workers* (5th ed.). Edinburgh: Oliver & Boyd, 1934.
- Fleiss, J. L. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 1975, **31**, 651-659.
- Siegel, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

Received 4 May 1979.

(Final Acceptance 20 July 1979.)

²For two-by-two data tables with less than 20 entries, see Siegel (1956).