A CAUTIONARY NOTE ON THE USE OF PROBABILITY VALUES TO EVALUATE INTEROBSERVER AGREEMENT

DONALD P. HARTMANN AND WILLIAM GARDNER

UNIVERSITY OF UTAH

Proposed methods of assessing the statistical significance of interobserver agreements provide erroneous probability values when conducted on serially correlated data. Investigators who wish to evaluate interobserver agreements by means of statistical significance can do so by limiting the analysis to every k^{th} interval of data, or by using Markovian techniques which accommodate serial correlations.

DESCRIPTORS: interobserver agreement, observational reliability, statistics, serial dependence

In the most recent round of the controversy over interobserver reliability (*Journal of Applied Behavior Analysis*, 1979, **12**, 491-571), a number of investigators suggested a potentially useful limiting value (or lower bound) for interobserver agreement: Obtained agreements must be statistically more numerous than chance agreements (e.g., Birkimer & Brown, 1979; Yelton, 1979). Although this general approach to evaluating interobserver agreement has elicited criticism (e.g., Hawkins & Fabry, 1979; Hopkins, 1979) our intent is to alert potential users to circumstances when this approach can be seriously misleading and to suggest alternative methods that are not misleading.

The results of simulations conducted in our laboratory (Gardner, Hartmann, & Mitchell, 1982) indicate that misleading results will occur when (a) traditional statistical methods, such as the Chi-square Test of Independence or the Fisher Exact Test, are used to calculate the statistical significance of interobserver agreements, and (b) when the data entering into these calculations are serially correlated. When each

observer's data were substantially serially correlated (i.e., lag one ϕ 's = $\pm .75$), the likelihood of obtaining statistically significant results *by chance alone* are inflated by as much as 500% using $\alpha = .05$ and by as much as 1700% using $\alpha = .01$. Even with more modest levels of serial correlation (i.e., lag one ϕ 's = $\pm .25$) chance levels of significance were increased by as much as 50% using $\alpha = .05$ and by as much as 350% using $\alpha = .01$.

These two conditions necessary to obtain erroneous results are not uncommon. The Fisher and Chi-square tests are the two recommended methods for assessing the significance of interobserver agreement (e.g., Birkimer & Brown, 1979; Hartmann, 1977; Yelton, Wildman, & Erickson, 1977) and serial correlation is a common characteristic of individual subject data (e.g., Hartmann, Gottman, Jones, Gardner, Kazdin, & Vaught, 1980). Thus, the use of significance tests to evaluate interobserver agreements could provide misleading results.

Fortunately, however, there are methods for obtaining accurate probability values available for investigators who may wish to use the statistical significance approach to evaluate interobserver agreements. Traditional statistical tests provide accurate results when they are conducted on serially *uncorrelated* data. A serially uncorrelated set of data can be extracted from

The work described in this article was supported in part by Research Grant HDMH 06914 from the National Institute of Mental Health, U.S. Public Health Service to Donna M. Gelfand and Donald P. Hartmann. Requests for reprints should be sent to Donald P. Hartmann, Department of Psychology, University of Utah, Salt Lake City, Utah 84112.

an originally serially correlated set by including only every k^{th} interval of data, where the $k - 1$ lag serial correlation is the last substantial lagged serial correlation. For example, if only the lag-1 serial correlation is substantial, include only every second observation in the reliability assessment; if the lag-1 and lag-2 serial correlations are substantial, include only every third observation. Alternatively, one could perform a valid significance test using the test for the independence of two Markov chains. (We appreciate the reviewer comments by J. M. Gottman, who pointed out the relevance of Markovian procedures in this context.) The details of this test may be found in Anderson and Goodman (1957), Billingsley (1961), or in Gardner and Hartmann (Note 1).

REFERENCE NOTE

1. Gardner, W., & Hartmann, D. P. *On Markov dependence in the analysis of social interaction*. Manuscript submitted for publication, 1981.

REFERENCES

- Anderson, T. W., & Goodman, L. A. Statistical inference about Markov chains. *Annals of Mathematical Statistics*, 1957, **28**, 89-110.
- Billingsley, P. *Statistical inference for Markov processes*. Chicago: University of Chicago Press, 1961.
- Birkimer, J. C., & 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.
- Gardner, W., Hartmann, D. P., & Mitchell, C. The effects of serial dependency on the use of χ^2 for analyzing sequential data. *Behavioral Assessment*, 1982, **4**, 75-82.
- Hartmann, D. P. Consideration in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, 1977, **10**, 103-116.
- Hartmann, D. P., Gottman, J. M., Jones, R. R., Gardner, W., Kazdin, A. E., & Vaught, R. S. Interrupted time-series analysis and its application to behavioral data. *Journal of Applied Behavior Analysis*, 1980, **13**, 543-559.
- Hawkins, R. P., & Fabry, B. D. Applied behavior analysis and interobserver reliability: A commentary on two articles by Birkimer and Brown. *Journal of Applied Behavior Analysis*, 1972, **12**, 545-552.
- Hopkins, B. L. Proposed conventions for evaluating observer reliability: A commentary on two articles by Birkimer and Brown. *Journal of Applied Behavior Analysis*, 1979, **12**, 561-564.
- Yelton, A. R. Reliability in the context of the experiment: A commentary on two articles by Birkimer and Brown. *Journal of Applied Behavior Analysis*, 1979, **12**, 565-569.
- Yelton, A. R., Wildman, B. G., & Erickson, M. T. A probability-based formula for calculating interobserver agreement. *Journal of Applied Behavior Analysis*, 1977, **10**, 127-131.

Received March 11, 1981

Final acceptance September 1, 1981