*RELIABILITY IN THE CONTEXT OF THE EXPERIMENT:
A Commentary on Two Articles by Birkimer and Brown*

ANN RUBINSOHN YELTON

REIDSVILLE CITY SCHOOLS

Two sources of variability must each be considered when examining change in level between two sets of data obtained by human observers; namely, variance within data sets (phases) and variability attributed to each data point (reliability). Birkimer and Brown (1979a, 1979b) have suggested that both chance levels and disagreement bands be considered in examining observer reliability and have made both methods more accessible to researchers. By clarifying and extending Birkimer and Brown's papers, a system is developed using observer agreement to determine the data point variability and thus to check the adequacy of obtained data within the experimental context.

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

The widespread use of human observers to obtain data in the study of behavior has sparked considerable interest and concern about the adequacy of such data. Human observers of behavior are subject to bias, judgment, inattention, and other sources of random and systematic error; thus it is imperative to assess the reliability of the collected data before such data may be considered valid or useful. In order to determine the reliability of data, two observers record simultaneously and independently and their protocols are then compared with respect to agreement; researchers who show acceptable levels of agreement assume reliable data.

Two approaches have been used to judge observer agreement. The first centers around comparing numbers of agreements and disagreements against some standard. The second approach involves comparing obtained protocols with what would be expected by chance if the two observers had responded randomly at their observed rates. The former approach may produce adequate levels of agreement when in fact the observers are responding at chance levels; also, the selection of adequate levels is arbitrary

and the obtained reliability is dependent upon the definition of agreement and the rate of the observed behavior. On the other hand, the chance methods may show nonchance responding even when the protocols are very dissimilar.

Birkimer and Brown (1979a, 1979b) have brought the two approaches together and have made each approach more accessible to researchers. Their major contribution, however, lies in their suggestions that both methods be used and that the adequacy of data be judged in the context of the magnitude of observed experimental effect. Thus very reliable data are needed to judge small changes in behavioral level between experimental conditions and relatively large numbers of disagreements can be tolerated with large changes in behavioral level.

Since Birkimer and Brown have presented their arguments in two separate papers, the practical integration of the approaches is not complete. Also, some clarifications and modifications of the techniques are needed. There are two sources of variability that must each be considered in the study of behavioral data; namely, the variance of data within phases and the variance of each data point estimated by observer reliability. Researchers must first determine that their data are adequate by comparing obtained

Reprints are available from the author, Reidsville City Schools, 920 Johnson Street, Reidsville, North Carolina 27320.

observer protocols with chance levels. Then they must estimate the variability for each data point; the disagreement band reflects the probable range of behavior for each point. Finally, the variability with phases and of data points must be used to compare observed change in level of behavior between phases. This approach provides researchers with assurance that their data are acceptable and provides a system that will put observer agreement in the context of the experiment.

JUDGING ADEQUACY OF OBSERVER AGREEMENT

Data that result from observers responding at chance levels are clearly not adequate for further analyses. Greater than chance levels of responding by observers is the first requisite for judging data acceptable and is necessary but not alone sufficient. Back to Basics (Birkimer and Brown, 1979b) examines the Yelton, Wildman, and Erickson (1977) method for comparing observer protocols to chance levels, and makes this method practical and accessible to researchers under most circumstances. The table presented is a laudable innovation and may often allow researchers to avoid the cumbersome calculations associated with the Yelton *et al.* probability-based formula. The 50-10-10 (90) rule further simplifies the comparison of data to chance levels in many circumstances. However, the conservative nature of both the table and rule must be emphasized. When the table shows nonchance agreement or the 50-10-10 (90) rule is applicable, the researchers are assured that the observers responded significantly better than expected by chance. When the table suggests chance levels (or the rules are violated), researchers should use the original formula to check the acceptability of their data because the table and rule will reject nonchance data.

A few clarifications may be helpful to researchers. First, for any given average rate of recorded behavior, the more divergent X and Y (the data from Observers 1 and 2), the more

unlikely that the number of agreements obtained were due to chance. Thus the table is valid beyond the restrictions listed (*i.e.*, X and Y differ by at most 10%). Second, the reader should be made aware that the Yelton *et al.* formula computes the probability of the obtained number of agreements or *better* (or, alternatively, the obtained number of disagreements or fewer) and thus goes beyond the Fisher Exact Test. Third, interpolation assuming a straight line as proposed is not valid; in the example presented concerning the "second researcher," if $X = 44$ and $Y = 44$, the probability of six or fewer disagreements would be .018 and thus *not* acceptable at a .01 level as suggested. Finally, it must be stated again that whenever the use of these conservative rules and table does not show acceptance, researchers should use the original formula.

USING DISAGREEMENT BANDS TO ASSESS VARIANCE WITHIN DATA POINTS

After it has been determined that the data obtained resulted from observations that were significantly above chance levels, researchers should examine the variability of their data. Even at nonchance levels, large numbers of disagreements may occur between observers which contribute to error or variability of data points. The use of disagreement bands is a useful visual aid for examining the probable range within which lies the behavioral rate. The assumptions underlying the disagreement band concept should be made clear, and these same assumptions lead to further understanding of the disagreement band concept. The first assumption is that if two observers agree that a behavior occurred, then the probability of its occurrence is 100%. Similarly, the probability of behavior is 0% when two observers agree on nonoccurrence. When exactly one observer records the behavior, the probability of occurrence is assumed to be 50% (we assign each observer an equal chance of being correct). Birkimer and

Brown create the most conservative band which includes all levels at which the "true" rate of behavior would occur under the above assumptions. However, each rate may be assigned a probability; for example, it is more likely that the rate would be near the center of the disagreement range than at either of the extremes. In the example provided by Birkimer and Brown (1979a) in their Figure 1, the true rate would fall between 20 and 30 100% of the time (the disagreement range), between 21 and 29 99% of the time, and between 22 and 28 95% of the time.

Contrary to Birkimer and Brown's suggestion, the disagreement band should be centered at the mean level for each phase (or, alternatively, around the projected best fit line). The variability found during reliability checks is assumed to be representative of the variability of each data point. Birkimer and Brown have ignored all data in their experiment except from days on which reliability was assessed; thus their data are reduced to four single data points that do not necessarily best represent data from each phase. One may visualize their procedure applied to data from phases that were not different. If reliability had been checked at extreme days, an experimental effect would have been concluded.

If several reliability checks are made within a phase, researchers may use the above assumptions combined with probability theory to determine a probable disagreement band.¹ However, researchers may use the most discrepant data to determine their band. Or, to keep the procedure simple as Birkimer and Brown suggest, researchers may be conservative and use the complete disagreement band.

USING DISAGREEMENT BANDS FOR DETERMINING MINIMALLY ACCEPTABLE LEVELS OF TREATMENT EFFECTS

A significant experimental effect is determined by examining differences between two or

more sets of data; researchers statistically examine whether one set of data is likely to have been drawn from a population defined by the data of another set. Researchers may inspect their data visually, perform mean comparisons, or may consider trend and other factors by comparing one data set to a best fit line predicated by a second set of data. Change in level between data sets is judged in the context of the data variance. In addition to the within-phase variability, Birkimer and Brown have made researchers aware of the variability of each data point; thus we may declare two sets of observational data (such as two experimental phases) significantly different only if they deviate significantly both with respect to the observed within-phase variance and the variance of data themselves (*i.e.*, disagreement range).

Clearly, then, neither the variance of the data within phases nor the variance associated with each data point (the disagreement range) alone can be used for judging experimental effects. Researchers should select the larger of these two variances when comparing the level of one phase with that of another. This procedure assures that the variance within phases used for analyses is at least as large as the variance associated with each data point. If visual analyses are used for analyses, researchers should at a minimum draw a disagreement range around the mean of each phase and check that the means of comparative phases fall outside the disagreement range.

The procedure suggested by Birkimer and Brown for comparing differences between phases relative to the disagreement band is far too conservative and does not take into account the variability of data within each phase. Just as data points may overlap between two phases, disagreement ranges may overlap. Researchers should examine whether the mean (or projected best fit line) for one set of data overlaps the disagreement range from another phase. The disagreement ranges are centered around the means (or projected best fit line) of each data set.

¹This procedure may be obtained from the author.

TWO ILLUSTRATIONS

Figure 1 is a reproduction of Birkimer and Brown's Figure 2. Since this example fits the 50-10-10 (90) rule, we are assured that the observer responded at better than chance levels. Probable disagreement bands are drawn around the mean of each phase to show the range in which behavior would occur 99% of the time. When the total disagreement band is 10, a 99% confidence band would have a span of 8. Since the mean level of each treatment phase falls outside the probable disagreement band of baseline and reversal phases, and since the converse is also true, the change in levels between adjacent phases is not likely due to data point variability. The researcher must finally compare differences between phases. Visually an experimental effect appears likely. *T* tests reject effects at .01 levels but do find differences at .05 levels.

Figure 2 shows another example of simulated experimental data. For the first phase, one reliability check was made; Observer 1 recorded 51 incidents of behavior and Observer 2, 53. They disagreed 34 times. Following the above procedure, the observers' protocols are compared with chance responding. Using the table, we may reject the hypothesis that the two observers did

not differ from chance responding at the .01 level. Next we construct the disagreement range around the phase mean of 50. The disagreement band (total possible range) is from 33 to 67. However, the rate would fall within a range of 16 99% of the time so a disagreement range is constructed from 42 to 58. This range, however, overlaps the mean from phase 2, thus we must conclude that our data are too variable to judge the observed change in level. Note that if the mean of phase 2 had been slightly lower, the data could have been judged adequate. If the researcher had ignored the variability of the data points (*i.e.*, disagreement range), a *t* test on the data would have suggested an experimental effect at .01 significance level.

In summary, researchers should follow these steps, which integrate Birkimer and Brown's two articles and include the modifications suggested in this paper.

1. Researchers should first make certain that their data are reliable. Two observers recording independently must obtain significantly fewer disagreements than could be expected by chance before data from these observers may be used for judging treatment effects. This criterion may

Fig. 1. Probable disagreement ranges applied to data from Birkimer and Brown's Figure 2 and centered at the means of each phase. Data point variability allows further analysis of changes in level.

Fig. 2. Probable disagreement ranges applied to simulated data and centered at the means of each phase. Data point variability results in rejection of significant change in level despite significant *t* tests on within-phase data.

be met by using the 50-10-10 (90) rule, the tables, or the probability-based formula.

2. After the data are judged reliable, the variability attributed to each data point should be determined by finding the disagreement range. This range should be centered at the mean of the data set rather than the particular data point at which reliability was assessed. Researchers may use a probable disagreement range rather than the conservative total band.

3. Experimental effects are judged by comparing mean levels of data sets within the context of the observed data variability. For acceptable variability, the disagreement range from one set of data should not overlap the mean from a comparative set. If this situation does not hold true, then the variance of each data point is concluded to be too great to judge experimental effects "believable." Otherwise, the

data from each set may be compared to determine whether statistically significant changes in level have been obtained.

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)
- 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 16 March 1979.

(Final Acceptance 20 July 1979.)