

APPLIED BEHAVIOR ANALYSIS AND INTEROBSERVER RELIABILITY:

A Commentary on Two Articles by Birkimer and Brown

Robert P. Hawkins AND Bernard D. Fabry

West Virginia University

To the extent that human foibles can readily influence data, we are likely to ask the question, "Is that what I would have seen if I had been there?" This is a personalized version of the question of generalizability across observers (Cronbach, Gleser, Nanda, and Rajaratnam, 1972), and it is one we are prone to ask whether the data are of the type that we call scientific (obtained by carefully described, replicable methods and usually quantitative) -as when it is reported that "The woman initiated conversations at a mean rate of .02 per hour" or of the type we call impressionistic-as when it is reported "She doesn't seem to talk much." It is a question asked by the general public as well as by those labeled "scientist."

Though we, as consumers of science, usually cannot be present personally to make our own observations, we are reassured to know that at least two persons observed the event (s) of interest and gave similar reports. If the two observers also did their observing repeatedly, were unlikely to be biased, and did their observing and recording independently, we are especially reassured. As journal reviewers, our reassurance takes the form of differentially reinforcing the reporting of research that includes repeated assessments of the agreement between two or more unbiased, independent observers of the same event(s).

Applied behavior analysts-well known as stubborn skeptics-have been particularly frugal in their dispensing of approval for scientific behaviors related to assessing generalizability (agreement, reliability) across observers. The result is that our methods for assessing generalizability across observers have advanced considerably beyond those used by most other behavioral scientists.

TRADITIONAL vs. BEHAVIOR ANALYTIC
APPROACHES TO RELIABILITY

It may be worthwhile to review here the differences between traditional and behavior analytic practices related to assessment of interobserver reliability. First, a common traditional practice in behavioral science is to assess interobserver agreement *prior to* conducting the research of interest but not during the research itself, on the assumption that the reliability is in the nonhuman aspects of the measurement system. This approach probably evolved

from experience with standardized tests, with which it has often been found that a particular test can produce very similar results when administered by different persons (observers) to the same subject, at least when one is conducting a study specifically to demonstrate such agreement. Though this traditional practice is still common in many fields of behavioral science and can occasionally be seen in applied behavior analysis (e.g., White, 1975), we prefer to directly assess the interobserver reliability of the very data from which we will be drawing conclusions, the data from the study itself. The importance of this has been demonstrated in a variety of studies showing the influence of various environmental factors on observers' data (see Kent and Foster, 1977; Johnson and Bolstad, 1973; Wildman and Erickson, 1977).

Another traditional practice is to summarize interobserver reliability data by means of a single correlation coefficient that represents the degree to which the data reported by two observers matched on repeated occasions or across several subjects. Although correlation coefficients have some advantages (Hartmann, 1977), they also have several disadvantages. First, correlation coefficients are typically used as an index of reliability across many sessions. This gives the researcher and research consumer no information as to what level of agreement was obtained on any particular session or even a small group of sessions, and this information is needed under some circumstances to be discussed later. Second, because correlation coefficients are highly dependent upon variance in behavior, their usefulness in assessing interobserver agreement within any experimental condition involving relatively stable behavior is very limited. Third, if one observer *consistently* inflates or deflates the data, for whatever reasons, the correlation coefficient will suffer no decline to reflect this (though this could be detected with a *t* test, as Hartmann points out). Fourth, correlation coefficients take the researcher too far from the raw data; thus, as judgmental aids (Michael, 1976), they have the disadvantage of reducing the likelihood of the researcher's being controlled by the raw data. Or, as Baer (1977b) put it, correlations "cook numbers to produce highly abstract outcomes" (p. 117). By contrast, behavior analysts typically assess agreement between observers by some method that facilitates attention to their agreement or

disagreement on each of many occasions (trials, samples, intervals, and sessions). Besides the fact that this probably results in more careful attention to the whole measurement methodology, it also keeps applied behavior analysis closer to being a "people's science" that can be understood and appreciated by the educated lay public.

Finally, traditional practice requires little reporting of the methodology by which interobserver agreement was assessed. Such factors as how observers were trained, on how many occasions two observers recorded, how these occasions were spaced across the study, whether this was during the study itself, or whether any attempt was made to have them record independently are often considered trivial details that need not be reported. Behavior analysts have taken such matters much more seriously, though even they often omit some methodological details of importance.

The articles by Birkimer and Brown (1979a, 1979b) continue this healthy concern with methodological details. First, they, like several others before them, have been dismayed by the fact that the most popular interobserver reliability scores are highly dependent upon the rate of the behavior so that low agreement scores are extremely improbable at very high or low rates of behavior (Hawkins and Dotson, 1975; Hopkins and Hermann, 1977). They have proposed two alternative solutions to this problem. In the "Graphical Aid" paper (1979a), they suggest plotting the range of disagreement expected by chance (actually, the mean of a distribution of disagreement scores varying in probability), given the obtained "rate" or incidence of the behavior, so that the obtained disagreement range can be visually compared to the chance range. In the "Easier Ways" paper (1979b), they present their preferred solution the use of some simple methods (a rule, a table, or if necessary, simple computations) to assess whether the obtained reliability differs significantly from chance.

A second concern they have addressed (like Hawkins and Dotson, 1975) is the limited value of most interobserver reliability assessments for evaluating the believability of experimental effects. Their solution to this problem is an ingenious method of graphically, yet rather simply, presenting the following wealth of information regarding any data point: the absolute level of both the primary and the secondary observer's data and thus the degree of agreement between them in terms of the total session scores they report (as suggested by Hawkins and Dotson); the number of occurrences reported by the primary observer but not confirmed by the secondary observer; the number of occurrences reported by the secondary observer but not

confirmed by the primary observer;¹ the number of occurrences on which they agreed; and the number of nonoccurrences on which they agreed. In addition, from that information presented graphically the reader who wished to could derive the interval-by-interval (I X I) agreement score that was traditional in applied behavior analysis for several years; interval-by-interval disagreement; the kappa and phi that Hartmann (1977) suggests; and Pearson product-moment correlation (across sessions, but not within them).

THE FUNCTIONS TO BE SERVED BY INTEROBSERVER RELIABILITY ASSESSMENT

In order to judge the value of Birkimer and Brown's proposals it is useful to consider the functions that interobserver reliability scores are to perform. In asking the question "Is that what I would have seen if I had been there?" we are indicating skepticism regarding at least four aspects of the data (Hawkins and Dotson, 1975). First, we (especially behavior analysts) are concerned whether the behavior has been defined adequately, whether the definitional aspect of measurement is replicable in (generalizable to) other scientists' "laboratories." Hawkins and Dobes (1977) suggest that a definition should generally be objective, clear, and complete to maximize replicability; and they provide data indicating that high interobserver reliability scores have not assured behavior analysis (or probably other behavior sciences) of replicable definitions. Second, we are concerned with the competency with which those definitions are employed to generate data, especially such factors as the training of observers, their conscientiousness, and their lack of bias. A third concern in any intervention study is whether we can believe the experimental effects shown (or not shown). Finally, a fourth concern is whether we can believe the absolute level of the behavior, aside

¹ Their use of the term "agreement on occurrences" in Figure 1 (1979a) may be misleading. The term refers to just those occasions on which both observers reported that the behavior occurred and is reported as a percentage of the *total* occasions (intervals, trials, or samples) in the session only because all of the data are reported in those units. It is unlike the more familiar "occurrence agreement" or scored-interval (S-I) agreement coefficient for which the number of agreements on occurrences (or, those occasions on which both observers reported that the behavior occurred) is divided by just the number of agreements and disagreements on occurrences, *not* the total number of occasions. The same analysis applies to their term, "agreement on nonoccurrences" which is similarly unlike the unscored-interval (U-I) coefficient of agreement.

from any effects shown. This is a somewhat different concern than that regarding the believability of the experimental effects, because in nonintervention, descriptive studies (e.g., Walker and Hops, 1976; White, 1975) there is no effect to be evaluated.

RELATIONSHIP OF PROPOSED METHODS TO THE FUNCTIONS OF INTEROBSERVER RELIABILITY

Birkimer and Brown's "Easier Ways" paper, in providing an easy means by which researchers can evaluate the probability or statistical significance of their obtained interobserver reliability, given the incidence of the behavior, probably makes interobserver reliability scores more consistent yardsticks by which to investigate all four of those concerns. For instance, the very high reliabilities reported by Hawkins and Dotson (1975)—under conditions where definitions were obviously inadequate, observers were obviously performing incompetently, and effects were quite possible by chance—would very likely be found too probable by chance alone had Birkimer and Brown's or Hopkins and Hermann's (1977) procedures for assessing the probability of an agreement score been applied. Thus, such inferential statistics *could* help behavior analysis protect itself from what are essentially unreliable (dependent on rate of behavior, not just on observer agreement) reliability scores. But the question then is *should* we use inferential statistics. Our own opinion is that we should not, even though the evaluation of an interobserver reliability score clearly implies evaluation of the probability of that score.

Our reasons for rejecting the use of inferential statistics here are twofold. First, we think Birkimer and Brown's Graphical Aid serves the same purpose (plus others) adequately, as we will discuss shortly, particularly when there are several reliability checks during a study. Second, the extensive involvement of behavior analysis in inferential statistics will distract it from its main subject matter, the development of a technology of behavior (Baer, 1977a). It is improbable at the .00001 level that Birkimer and Brown's "Easier Ways" will be the last word regarding the reliability (deviation from chance) of reliability scores. Instead, someone will soon point out, as did Ryan (1962) several years ago regarding multiple significance tests within the same study, that one should really test the reliability of a reliability score only with a test that takes into account the number of reliability assessments done in the study. Someone else will question—and already has (Owings, Note 1)—the appropriateness of the Melton, Wildman, and Erickson (1977) model for evaluating probability. Other approaches will be proposed to replace the Birkimer and Brown approach.

And all of these publications will produce more, resulting in greater and greater statistical "sophistication," which is sometimes hard to discriminate from irrelevance. Those publications will often be written by, reviewed by, and read by behavior analysts who could have been spending their time working on a technology of behavior.

This fear of distraction from our primary subject matter is not unfounded paranoia. Besides the fact that much time is required to learn how to obtain and interpret inferential statistics (Michael, 1974), one has only to open some of the journals in other behavioral sciences to see some unfortunate results of excessive concern with this subject matter. First, statistical sophistication often becomes confused with scientific merit; thus one finds many studies in which there are so many variables and the data are so thoroughly "cooked" on the statistical stove that it is very difficult to detect which of the innumerable results are of importance or even what the results really mean. Second, the statistical analyses sometimes occupy so much of the text of a published article that the material makes difficult, uninteresting reading. Third, researchers become so preoccupied with their statistics that they seem to forget about what use the results might have; thus one sometimes finds an article in which it is reported that "there was a significant difference between the means of the groups," yet there is no mention of what the means were. And, worse yet, occasionally one finds a study reporting a significant difference without mentioning which group performed better than which group! Such preoccupation with complex judgmental aids or tools of science is something behavior analysis has eschewed fairly successfully thus far. Let us continue to avoid it.

A further problem with Birkimer and Brown's statistical approach (and, we are tempted to say, with any other statistical approach, except then someone will publish a paper presenting an exception, thus perpetuating the statistical discussion) is that it is not adequate to deal with reliabilities on very low or high incidence behaviors, and it even deals inadequately with reliabilities on the middle incidence range. Observers could agree perfectly, on repeated sessions, that a particular behavior occurred on every occasion (100%) or failed to occur on every occasion (0%), but their reliability could not be assessed using the Easier Way, even if those same observers showed very high reliability on several other behaviors being recorded at the same time.² If one is interested in documenting, for

² Of course, given the opportunity, some of our statistically sophisticated colleagues will provide us with another statistical test that can be used to evaluate the probability of obtaining, say, five consecutive agreements on the absence of behavior given a par-

example, the absence of praise in certain classrooms (e.g., White, 1975), or a change in performance from 100% errors to zero errors, judgmental aids other than the Easier Way would be necessary. In addition, the Easier Way, like occurrence reliability (S-I) and nonoccurrence reliability (U-I), is increasingly unfair to data as they approach the 100% or 0% levels, because a single disagreement has a drastic effect on interobserver reliability. Most behavior analysts would probably find data very believable if two observers report the following incidences out of, for example, 100 occasions per session:

Session No.	Obs. A	Obs. B
2	1	2
5	0	1
9	4	3
12	3	2
17	5	3

Yet none of these sessions can produce the level of occurrence (S-I) reliability (the presently accepted statistic) we currently demand.³

The Easier Way may even be too lenient with data in the midrange. If two observers were recording data in a session of 50 or more intervals, one observer could report occurrences in as few as 35% of the intervals while the other observer reported occurrences in 65% of the intervals and still have "acceptable" agreement ($p < .01$) according to the Easier Way (see Figure 1, Birkimer and Brown, 1979b). Most behavior analysts would probably consider that unacceptable.

On the other hand, a graphical aid like Birkimer and Brown's or others' (Goldiamond, 1965; Hawkins and Dotson, 1975; Morris, Rosen, and Clinton, Note 2) could largely avoid these problems. The agreement between observers on the low rate behaviors shown above would be seen as adequate, at least in terms of the absolute level of the behavior (function number four presented earlier) and the believability of any major effects shown (function number three). And agreement at all levels of incidence of the behavior could be evaluated by the same yardstick.

Let us consider the adequacy of the Birkimer and Brown Graphical Aid for each of the four functions. First, it appears to offer a better means to evaluate the adequacy of a definition than any other reliability assessment extant. Unlike S-I and U-I reliabilities which suffer from being very unfair at extreme rates, as pointed out above, the Graphical Aid provides an index that appears fair to all data. An example using raw data

ticular level of agreement on the presence of the other behaviors.

³ It is interesting that correlation coefficients will tend to be more fair to such data, provided the data vary sufficiently from session to session.

will help to illustrate this and will offer a context in which we can discuss the particular logic of the Graphical Aid.

Suppose that two observers show the following occurrences (X) of a behavior over 20 trials, momentary time samples (Hawkins, Axelrod, and Hall, 1977), product samples, or intervals:

Occasion	1	2	3	4	5	6	7	8	9	10	11
Observer A	X			X	X	X					X
Observer B						X		X	X		
Occasion	12	13	14	15	16	17	18	19	20		
Observer A	X		X								
Observer B	X	X	X								

Using Birkimer and Brown's method, their data will be graphed at 35% for observer A and 30% for observer B. This would appear to be close agreement until the disagreement range is plotted. It will show a disagreement range of 35% extending from 15% to 50%. In addition, by centering that range around the midpoint between the two observers' data, Birkimer and Brown make it possible to find out on how many occasions observer A reported occurrence while observer B did not. One simply checks the distance between the bottom of the disagreement range—15% or three occurrences in the present example—and observer A's data point. That difference is 20% or four occurrences, which checks with the raw data. Similarly, the distance from the bottom of the disagreement range to observer B's data point, a distance of 15% or three occurrences, shows the percentage or number of occasions on which observer B reported occurrence but observer A did not. Thus not only is the extent of disagreement shown, but also the precise nature of those disagreements. This seems a distinct advantage to both the researcher and the research consumer.⁴

⁴ It should be pointed out that when data are reported in percentage units, the disagreement range is identical to I X I disagreement, which is simply 100% minus I X I agreement. Thus it will be correlated with the rate of behavior just as I X I agreement is (Hawkins and Dotson, 1975; Hopkins and Hermann, 1977). But the Graphical Aid makes the nature and magnitude of agreement and disagreement so clear, and puts that information in such close conjunction with the specific session data from which it comes, that the risks pointed out by Hawkins and Dotson seem largely eliminated. If, in addition, reliability is checked on several occasions during a study and found to be satisfactory, the adequacy of the definition and the competency of its use are even better established.

For the purpose of evaluating the *competency with which the definitions are employed* the Birkimer and Brown Graphical Aid again seems superior to currently used methods. Not only does it give greater information about the nature of agreements and disagreements, it also shows any systematic observer bias. As Hawkins and Dotson (1975) point out, if one observer is unbiased, either a consistent bias or a changing bias (e.g., to show an experimental effect) on the part of the other observer will be evidenced by the direction of the difference between the scores they report. This advantage is absent in all currently used reliability scores but is present in the Birkimer and Brown Graphical Aid.

For the purpose of evaluating observer *reliability on the absolute level* of the behavior, the Graphical Aid seems eminently suited. The reasons include those just listed, plus one other that can best be illustrated by referring again to the example above. Considering observer A as the primary observer for the moment, one might consider observer B's data as an attempt to see how low or high the subject's behavior may really be,⁵ if the observation system were working perfectly. Looking at the raw data we see that four of the occurrences reported by observer A were unconfirmed. Thus, it is possible that the subject's behavior was 20% lower than observer A reported. Similarly, there were three nonoccurrences reported by observer A that were not confirmed by observer B, so it is possible that the subject's behavior was really 15 % higher than observer A reported. It will be noted that these figures define the limits of the disagreement range, as graphed by Birkimer and Brown. This is the logic of the Graphical Aid that most appeals to us and that results in its great advantage as a test of both the reliability of the absolute level and the believability of the experimental effect, which we will discuss next.

Perhaps the greatest value of the Graphical Aid is in evaluating the *believability of an experimental effect*, a function for which other reliability scores have been of little use. It goes beyond the Hawkins and Dotson (1975) and Goldiamond (1965) methods in a very significant way: it shows the magnitude of change needed to convincingly claim that an effect occurred. Because the disagreement range shows how high or low the behavior may really be, the disagreement ranges obtained on repeated reliability checks will form a "no-

⁵ Given the complexity of human behavior we may never be able to assume that any observer or group of observers could tell us the "real" level of behavior, especially in light of recent findings (Kent and Foster, 1977; Wildman and Erickson, 1977) that observer biases result in error. Perhaps the best we can do is to say to what extent a primary observer's data are generalizable to other observers.

change band" across the graph, and only when the bands from any pair of experimental conditions fail to overlap can great confidence be placed in the effect obtained.

The advantages of the Graphical Aid seem even greater than Birkimer and Brown discuss.⁶ First, the plotting of both observers' data is a form of reliability assessment that can be used with frequency, rate, and duration data as well as with the kinds of data dealt with by Birkimer and Brown. Up to the present we have used different methods for assessing the reliability of different kinds of data. A common index would be an asset. In fact, it is even possible to obtain disagreement ranges with frequency, rate, and duration data by simply dividing each session into segments, each of which produces its own level of agreement. Further, data from rating scales (e.g., 5-point scales instead of the binary "yes"- "no" of interval and similar data) can even be plotted with their disagreement ranges.

A second, probably greater, advantage of the Graphical Aid not discussed by Birkimer and Brown is that when researchers find, during a prebaseline (recommended by Hawkins and Dobes, 1977) or baseline evaluation of the adequacy of the measurement system, that the nochange band extends as high (or low) as the level to which they believe they can change the behavior with the intervention planned, they will be alerted to improve the measurement system enough to reduce the no-change band below (or above) that level. Further, if the researchers have exhausted all reasonable resources for narrowing the no-change band and still find it too wide, they will be alerted to change the planned intervention in some way so as to augment its effect. The result of these contingencies on researchers should be exactly the opposite of that which Baer (1977a) feels are wrought by statistical tests of the significance of effects: the behavior analyst would more

⁶ One potential disadvantage of the Graphical Aid should also be noted. Because it is graphical rather than numerical, it would be awkward for a secondary source to describe, as when a reviewer of other scientists' research attempts to relate different studies in which different levels of reliability were obtained. Of course this is partly solved by the reviewers' simply presenting the disagreement range in numerical form; but that then leaves the problem that, like I X I agreement, the disagreement range is highly influenced by the incidence of the behavior. When reading a secondary source, the reader of the disagreement range will have difficulty interpreting it without access to information regarding the incidence of the behavior. The solution to this problem would seem to be simply to include the incidence in the report, as in this example: "Smith reports reliabilities ranging from a disagreement range of 5 % (at incidence of 7%) to 17% (incidence, 79%)."

actively search for robust variables, interventions that reliably produce socially or personally significant effects. That is a result which would preserve and enhance the unique contribution that applied behavior analysis has to offer.

REFERENCE NOTES

1. Owings, R. A. *Assessing interobserver reliability: How much and how likely*. Unpublished manuscript, 1978. (Available from Institute on Mental Retardation, George Peabody College for Teachers).
2. Morris, E. K., Rosen, H. S., and Clinton, L. P. *Reliability considerations in single-subject research*. Paper presented at the meeting of the Midwestern Association for Behavior Analysis, Chicago, May 1975.

REFERENCES

- Baer, D. M. "Perhaps it would be better not to know everything." *Journal of Applied Behavior Analysis*, 1977, **10**, 167-172. (a)
- 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. (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. (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)
- Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. *The dependability of behavioral measures*. New York: Wiley, 1972.
- Goldiamond, I. Stuttering and fluency as manipulatable operant response classes. In L. Krasner and L. P. Ullman (Eds), *Research in behavior modification*. New York: Holt, Rinehart & Winston, 1965.
- Hartmann, D. P. Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, 1977, **10**, 103-116.
- Hawkins, R. P., Axelrod, S., and Hall, R. V. Teachers as behavior analysts: Precisely monitoring student performance. In T. A. Brigham, R. P. Hawkins, J. W. Scott, and T. F. McLaughlin (Eds), *Behavior analysis in education: Self-control and reading*. Dubuque, Iowa: Kendall-Hunt, 1976.
- Hawkins, R. P. and Dobes, R. W. Behavioral definitions in applied behavior analysis: Explicit or implicit. In B. C. Etzel, J. M. LeBlanc, and D. M. Baer (Eds), *New developments in behavioral research: Theory, method and application*. In honor of Sidney W. Bijou. Hillsdale, New Jersey: Lawrence Erlbaum Assoc., 1977.
- 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 reliability of interval data. *Journal of Applied Behavior Analysis*, 1977, **10**, 121-126.
- Johnson, S. M. and Bolstad, O. D. Methodological issues in naturalistic observation: Some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, and E. J. Mash (Eds), *Behavior change: Methodology, concepts, and practice*. Champaign, Illinois: Research Press, 1973.
- Kent, R. N. and Foster, S. L. Direct observational procedures: Methodological issues in naturalistic settings. In A. R. Ciminero, K. S. Calhoun, and H. E. Adams (Eds), *Handbook of behavioral assessment*. New York: Wiley, 1977.
- Michael, J. Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis*, 1974, **7**, 647-653.
- Ryan, T. A. The experiment as the unit for computing rates of error. *Psychological Bulletin*, 1962, **59**, 301-305.
- Walker, H. M. and Hops, H. Use of normative peer data as a standard for evaluating classroom treatment effects. *Journal of Applied Behavior Analysis*, 1976, **9**, 159-168.
- White, M. A. Natural rates of teacher approval and disapproval in the classroom. *Journal of Applied Behavior Analysis*, 1975, **8**, 367-372.
- Wildman, B. A. and Erickson, M. T. Methodological problems in behavioral observation. In J. D. Cone and R. P. Hawkins (Eds), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977.
- 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 21 March 1979.
(Final Acceptance 20 July 1979.)

Reprints may be obtained from Robert P. Hawkins, Psychology Department, West Virginia University, Morgantown, West Virginia 26506.