REVIEWER'S COMMENT: JUST BECAUSE IT'S RELIABLE DOESN'T MEAN THAT YOU CAN USE IT

DONALD M. BAER[1]

THE UNIVERSITY OF KANSAS

Hartmann's (1977) article is scholarly and competent—perhaps completely so, technically. However, it is not a useful one for the *Journal of Applied Behavior Analysis,* and there are two reasons for this lack of utility.

First, the majority of the argument reflects a tradition from which applied behavior analysis is a deliberate departure. It is a tradition in which the detection of *all* functional variables and their interactions is the prime goal, rather than the evaluation of the magnitude, effectiveness, durability, or generality of the variables and interactions detected. By contrast, applied behavior analysis is a discipline deliberately turning away from the detection of weak variables: it systematically filters from its discovery methods the ability to discover variables of less-than-powerful effect. In its attention to systematic replication rather than direct replication (Sidman, 1960), it also eventually filters itself of nongeneral variables, through replication across studies rather than replication within studies. Were the field to become sensitive to the arguments of this article, and adopt them, it would be doing so for only one major reason: to begin to detect weak variables. In this reviewer's opinion, the distinctiveness of the field, its health, and its success, both as a basic theory of behavior and as a useful technology of behavior change and problem solving, would then be at risk. To be only a little too dramatic, everything the field had done correctly in its flight from the morass of traditional psychology would be in danger of abandonment or swamping out, if we were to apply most of the arguments of this article. Michael (1974) had made essentially the same evaluation, discussing JABA's symposium of arguments on the appropriateness and usefulness of statistical analysis of single-subject-design-derived data.

Perhaps a specific case will exemplify the meaning of this point further: Hartmann argues (p. 111, under the heading, *Accuracy of Reliability Estimate*) that "The primary concern with any estimate of reliability is that it reflects accurately and with minimum ambiguity the degree of reliability of the data assessed." Here is a very clear statement of belief in the existence of a true reliability, which has actual existence and is merely estimated by scientists. The reviewer doubts that this conception is viable for applied behavior analysis: reliability *is* what the estimate prcduces; thus, the author's excellent demonstration of the fact that different means of estimating reliability produce different numbers, may well reflect not the inadequacy of one *versus* another, but the arbitrariness of all.

For applied behavior analysis, choice among estimates can be guided by (1) the avoidance of allowing the reliability of occurrence from influencing the reliability of nonoccurrence, and *vice-versa;* and (2) by the apparent, face meaning of the estimate's calculation technique. For applied behavior analysis, the technique is the primary reality. Thus, techniques that cook numbers to produce highly abstract outcomes may be just as good as techniques that do not, *in the abstract,* but they are of correspondingly little use in application to real-life settings. Percentage of agreement, in the interval-recording paradigm, does have a direct and useful meaning:

[1]Reprints may be obtained from the author, Department of Human Development, The University of Kansas, Lawrence, Kansas 66045.

how often do two observers watching one subject, and equipped with the same definitions of behavior, see it occurring or not occurring at the same standard times? The two answers, "They agree about its occurrence X% of the relevant intervals, and about its nonoccurrence Y% of the relevant intervals", are superbly useful. By contrast, the fact that one observer saw four occurrences and the other saw five, yielding a supposed percentage of agreement of 80%, has very little usefulness: no one knows whether the first observer's four were seen at the same times as any of the second observer's five. Thus, their seeing the same behaviors in the same way at the same times could easily be nil. Similarly, phi, kappa, Pearson's r, and the analysis-of-variance proportion of variance are terribly subject to phenomena other than two observers seeing the same behavior in the same way at the same times; their characteristics are interesting, in the abstract, and just as true as any other estimate, in the abstract, but not useful to the simple problem of the behavior analyst: would any two observers using the same behavior code see the same behaviors in the same way at the same time? Thus, this reviewer, taking a deliberately simplistic view as the essence of maintaining the applied approach as what it is, sees no value in this very competent article. If the field were to turn about and begin collecting all variables and interactions, rather than only the ever-powerful ones, then it might need this article. But perhaps not even then, because of the second argument which follows.

Much of this discussion is derived from the long-standing analysis of reliability developed for the evaluation of psychometric instruments. The model for that problem was probably the questionnaire. A questionnaire is a series of items to which the subject responds, either digitally (Yes or No) or along some continuum (very like me, somewhat like me, *etc.*). The central question posed for reliability analysis was whether the instrument as a whole was, in fact, measuring anything—not whether it was measuring what it had been designed to measure, but whether it was measuring *anything.* To be a measuring instrument, it was important that the items in the instrument reflect the same dimension for the majority of subjects. That is, if one item reflected intelligence, the next motivation, the next mood of the moment, the next color preference, the next whether the subject happened to have read the day's newspaper, *etc.,* then the instrument was not going to prove reliable, neither for one subject nor across a sample of subjects. In short, the essence of a score's reliability was the *homogeneity* of the items that were combined to produce that score.

In 10-sec time-sampling observations, by contrast, there is no analogous commitment to homogeneity. The items, so to speak, can only be the successive 10-sec intervals: each of them is like a question put to a subject in a questionnaire. But, whereas a successful questionnaire will ask essentially the same question in a variety of ways and at a variety of intensity levels, 10-sec time sampling is asking a perfectly open-ended question every 10 sec: merely, What are you doing now? It is not asking (for example), Interval 1: Are you aggressive in this way? Interval 2: In that way? Interval 3: Under this much provocation? Interval 4: Under that much provocation? Interval 5: With your body? Interval 6: With your langrage? Interval 7: To this kind of target? Interval 8: To that kind of target? Instead, the observer will have been supplied with a definition of aggression, and each 10-sec interval will ask, Has that sort of aggression happened in the last 10 sec? The answer will be Yes in some intervals, No in others. There is no need to assume that the intervals will be homogeneous. A correlation of any half of them to the other half of them will have no necessary value. Aggression (or any other operant) is not something distributed with some homogeneity across 10-sec intervals; it is a response class discriminated through environmental history to some classes of discriminative stimuli. If there is no reason to believe that these stimuli are distributed homogeneously over time, then no homogenity should be expected in any succession of 10-sec intervals. To put the same point differently: for the usual psychometric instrument, we construct items so that they will appeal with some homogeneity to the same dimension of measurement; in time sampling, we cannot construct the items (the successive 10-sec intervals), but rather must accept them as they come, evoking whatever they will—which may well be anything.

Where homogeneity can be expected—indeed, required—is across *observers*. Any number of observers, equipped with the same definition of aggression, should be able to look at the same event and say Yes or No homogeneously, if looking with that definition is a reliable process. That is where reliability is desired, so that is where homogeneity is meaningful. Hence, the homely measures of observer agreement so widely used in the field are exactly relevant to the problem (properly segregated for occurrence and nonoccurrence, of course). If anything needs improvement, it may well be the extent to which we usually sample across observers: two appears to be considered a generous number, and clearly, if this is generosity, it is the smallest possible version of it. Even so, it is recognized that observers, as a class, generally are homogeneous. There are few, if any, data in the literature to verify this claim, but many workers in the field, especially those who have employed many, many observers over the years, probably would attest to it. (This reviewer, for example, estimates that no more than one reliability "problem" in 20 can be resolved by firing one observer and hiring another, as contrasted to rewriting the definition(s).)

In summary: there is very little wrong with the previous article, and much to admire. However, it is thoroughly inappropriate for this field and this journal. Nevertheless, Hartmann can be respected for demonstrating an unusual degree of scholarship and devotion to furthering the cause of applied behavior analysis. The main problem is that the reviewer does not agree that he has indicated the correct direction.

REFERENCES

Hartmann, D. P. Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis,* 1977, **10**, 103-116.

Michael, J. Statistical inference for individual organism research. Mixed blessing or curse? *Journal of Applied Behavior Analysis,* 1974, **7**, 647-653.

Sidman, M. *Tactics of scientific research.* New York: Basic Books, 1960.

Received 16 January 1975.
(*Final acceptance 15 May 1976.*)