Pharmacology Biochemistry & Behavior, Vol. 13, pp. 243–246. Printed in the U.S.A.

The Use of Human Observers in Psychopharmacological Research

ALAN POLING,[1] JAMES CLEARY AND MICHAEL MONAGHAN

Department of Psychology, Western Michigan University, Kalamazoo, MI 49008

Received 30 January 1980

POLING, A., J. CLEARY AND M. MONAGHAN. *The use of human observers in psychopharmacological research.* PHARMAC. BIOCHEM. BEHAV. 13(2) 243–246, 1980.—During 1974–1978, over 40% of the nonhuman drug studies that appeared in *Psychopharmacology, Pharmacology Biochemistry and Behavior,* and *Journal of Pharmacology and Experimental Therapeutics* involved human observers; far fewer studies published in *Journal of the Experimental Analysis of Behavior* did so. In all of these journals, measures of interobserver agreement seldom were provided. The great majority of studies also failed to utilize one or more "blind" observers, unaware of experimental conditions. These findings are of interest in light of reports that observational data are affected by a wide range of factors and often provide an inaccurate index of behavior. The believability of observational data seemingly is enhanced by careful descriptions of recording procedures coupled with the use of two or more blind observers whose concordance in rating behavior has been determined. These procedures characteristically are followed in some behavioral sciences, such as applied behavior analysis, but not to the same degree in psychopharmacology.

Data-recording procedures Interobserver agreement Blind observer Observational dependent variable Pharmacological independent variable

THE behavioral sciences systematically attempt to disclose functional relations between events, at least one of which involves overt behavior. Empirical demonstration of functionality is often difficult and depends upon the utilization of analytical procedures known collectively as research methodology. Accepted research tactics vary greatly, depending upon the kinds of events under consideration and the theoretical orientation of the experimenter. Research methodology is not consistent across areas of investigation, nor across time. Continual refinement of analytical procedures is characteristic of a healthy discipline, and advances in such procedures have been instrumental in the development of many of the natural sciences [14,23].

During the past century, one refinement of widespread import has been an increased reliance on automated devices for manipulating and monitoring events. Machines allow for the control of variables not directly manipulable by humans, and measure forms and amounts of energy that do not perceptibly affect our unaided senses. Further, machines are objective, tireless, and typically more reliable than their human counterparts. Nonetheless, at the present time certain areas of behavioral research necessitate the use of humans as "transducers" who convert directly observed actions into numeric form. Human observers are required when complex behaviors not easily detected in meaningful ways by machines are of interest, or where the costs of automated recording are too great.

In recent years, the use of human observers has become commonplace in that area of psychology known as applied behavior analysis, which manipulates environmental contingencies in an attempt to produce desired alterations in socially significant human behavior [2]. However, researchers in this area have strongly emphasized that the use of humans to quantify behavior, although often unavoidable, entails procedural difficulties [3, 5, 6, 9, 10, 12].

As transducers, humans invariably are suspect. Folklore suggests that lay observations are an imperfect reflection of actual happenings, and a large and growing body of data indicates that allegedly scientific observations sometimes provide an inaccurate index of the variables being considered [3, 9, 10]. Among the factors demonstrated to influence reported observations are the observer's motivation and expectations (e.g. [21]), the specifics of the observational situation (e.g. [9]), the observational and data recording techniques that are used (e.g. [19]), and the characteristics of the behavior being monitored (e.g. [10]).

In view of these considerations, researchers in applied behavior analysis have gone to great lengths to ensure the believability of their observations. Beyond defining in detail the behavior(s) under consideration and carefully describing the observational procedures used, these investigators nearly always provide some measurement of interobserver agreement, which specifies the degree of correspondence obtained between the data recorded by each of two (or more)

[1]To whom offprint requests should be addressed. We thank Carol Scarberry, Carol Parker and Rhonda Mullen for their help in preparing the manuscript, and the members of the Behavioral Pharmacology Laboratory for their comments on an earlier version. Preparation of this manuscript was supported by a Faculty Research Fellowship awarded to Alan Poling by Western Michigan University.

independent observers [11]. Where feasible, "blind" observers—individuals not aware of the experimental conditions in effect—are employed, and video tapes of the subject's behavior are made and subsequently used to check the accuracy of reported data [3,10]. Despite these precautions, the reliability and validity of observational data in applied behavior analysis continues to be questioned [5, 6, 7, 9].

Interestingly, psychopharmacology and other behavioral sciences regularly use humans to directly collect data, but researchers in these areas seem to be less cautious in this practice than applied behavior analysts. Several authors have criticized the observational procedures used in clinical drug studies [1, 15, 16, 25, 26, 27], but little has appeared concerning data collection in basic research. The present manuscript is concerned with observational techniques in nonhuman drug studies during the last five years (1974–1978).

PROCEDURE

Each empirical article that appeared in *Psychopharmacology*, *Pharmacology Biochemistry and Behavior (PBB)*, *Journal of Pharmacology and Experimental Therapeutics (JPET)*, and *Journal of the Experimental Analysis of Behavior (JEAB)* from 1974–1978 was rated on a standard scale by one or two individuals. Technical reports, review articles, and theoretical papers were not scored. Dichotomous ratings (yes or no) were made on the basis of the following questions: (1) Was at least one independent variable a pharmacological (drug) manipulation? (2) Was at least one dependent variable a behavioral response assessed directly by human observers? A dependent variable was rated as observational only if someone actually watched the subject, and no lasting record of behavior beyond that person's rating was obtained. (3) If the dependent variable was observational, was a measure of interobserver agreement provided? (4) If the dependent variable was observational, was at least one blind observer (a person unaware of experimental conditions) used?

Articles typically were rated by a single person. However, in one volume of each of the journals every article was independently rated by both observers. Two raters scored Volume 55 of *Psychopharmacology* (41 articles), Volume 7 of *PBB* (94 articles), Volume 190 of *JPET* (68 articles), and Volume 24 of *JEAB* (35 articles). Independent ratings were used to calculate a percentage measure of interobserver agreement, according to the formula $(A/A + D) \times 100$, where A is the number of articles where all of the ratings of the observers agreed and D is the number of articles where one or more of the ratings of the two observers disagreed. Interobserver agreement was 93% for *Psychopharmacology*, 89% for *PBB*, 94% for *JPET*, and 100% for *JEAB*. These percentages indicate that the evaluative dimensions under consideration were sufficiently clear and well-defined to allow independent observers to consistently and characteristically agree in their ratings.

RESULTS

Over 4000 studies were evaluated in the present study; Fig. 1 provides a summary of the yearly data collected for each journal. Approximately 40% of the studies published in *Psychopharmacology* and *PBB* that employed a pharmacological independent variable and a behavioral dependent variable used humans to collect data; slightly more than half of the studies that appeared in *JPET* did so. Proportion of observational studies published in these journals was relatively constant across the five years considered, although an upward trend was evident in the *PBB* data. In *JEAB*, only one psychopharmacological experiment involved an observational dependent variable. That experiment was reported in 1976.

Across all years and journals, fewer than 10% of observational drug studies reportedly employed a blind observer, while less than 5% of such studies provided a measure of interobserver agreement. Neither of these measures differed greatly across years or journals.

DISCUSSION

The prevalence of observational dependent variables in nonhuman drug research is of interest, for considerable effort has been expended in developing techniques for automatically recording behaviors that once were of necessity directly measured by humans. For example, assessment of motor activity historically involved observers who recorded the number of grids that animals traversed in an open field, or some similar measure. Today, movements commonly are detected by highly sensitive devices that operationalize activity as disruptions of photocell beams or ultrasonic waves [20]. Comparable technological developments have occurred in the devices used to measure many other behaviors.

This notwithstanding, instrumentation may be prohibitively expensive, and certain behaviors simply cannot be easily monitored by machines. Topographically complex actions are particularly troublesome in this regard. Unfortunately, they also are problematic for human monitors. Although humans sometimes can reliably monitor quite complex actions (e.g. [8]), the accuracy of observation assumedly reflects the complexity of the behavior considered and the concomitant clarity with which it is defined [3]. We did not attempt to evaluate the complexity of the behaviors monitored in nonhuman drug studies, or the apparent likelihood that they could be unambiguously scored. Such data would be meaningful only if our ratings actually predicted observers' performance, a relation difficult to demonstrate. It was clear, however, that in many studies observers were required to make difficult discriminations—for example, to decide whether cats had engaged in "hallucinatory-like" behavior.

Even when the observer's task is a simple one, it cannot be assumed a priori that his or her observations are veridical, or would correspond to those of another trained person, for a number of factors are known to affect observational data [9, 10, 11]. In certain situations, their impact may be sufficient to render observational data inaccurate [3, 10, 12, 21].

In a well known series of studies, Rosenthal and coworkers found that an observer's expectations consistently influenced reported findings [21]. An experiment by Rosenthal and Fode [22] exemplifies studies conducted by this research group, and their findings. Rosenthal and Fode had undergraduates train rats in a maze-learning task. One group of students was told that their rats were "maze-bright," another that their rats were "maze-dull." In actuality, all students were assigned experimentally-naive rats randomly chosen from a homogeneous population. Nevertheless, those students whose rats were labelled "maze-bright" reported faster learning than those students given "maze-dull" subjects. Although certain findings by Rosenthal's group have not been replicated [9,24], the confounding effects of

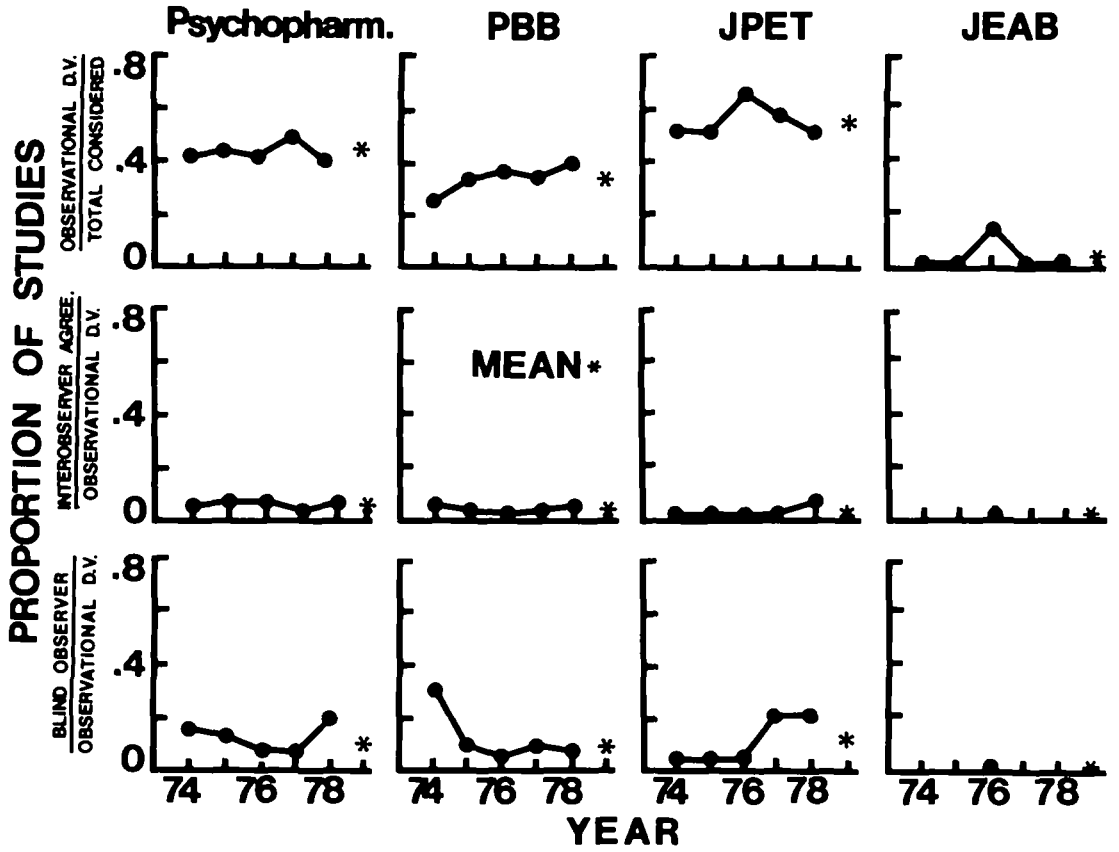

FIG. 1. Proportion of total studies considered that used humans to directly observe behavior, proportion of observational studies that employed a blind observer, and proportion of observational studies that provided a measure of interobserver agreement. Only articles with a pharmacological independent variable and a behavioral dependent variable were considered; rating procedures are described in text.

observer expectation, or bias, are well documented [9]. Simply put, in many situations observers report data consistent with their expectations. Thus it is imperative that observers be unaware of experimental conditions and their expected outcomes [9].

In nonhuman drug studies, observers unaware of experimental conditions and hypotheses rarely are employed. Why this is so is unclear. Perhaps some researchers are unaware of the potential problems associated with observer expectations. This is unlikely, for the confounding effects of this extraneous variable in clinical and preclinical drug studies has been repeatedly emphasized [1, 15, 16, 17, 20, 25, 26, 27]. It is more probable that the conspicuousness of many pharmacological manipulations has seemed to limit the value of uninformed observers (cf. [4]): If the condition in effect is readily apparent, as when an animal has received certain drugs, giving this information prior to observation might appear inconsequential. This is not unreasonable, but it does not follow that observers should be given specific expectations about experimental outcomes.

In any case, having a single informed observer who serves as the primary data source is difficult to defend, especially when that person has a vested interest in the outcome of the experiment [10]. Perhaps the use of such an observer is marginally defensible in those situations where objective, nonobservational data form the analytical basis of the study and the observational data serve an ancilliary function, strengthening or clarifying conclusions based on other information. In several of the articles we rated, observational data were in fact of secondary import; conclusions did not rest exclusively upon them. However, even when data are not of crucial importance, it seems advisable to take reasonable steps to demonstrate that they are unbiased and believable.

Methodologically, the use of multiple observers whose agreement in rating behavior is calculated is as important as demonstrating that experimental conditions apart from observer expectations are responsible for changes in the dependent variable. Interobserver agreement may be calculated in several ways, which are considered elsewhere [3, 5, 7, 10]. Although techniques for calculating interobserver agreement differ in many significant aspects, each ultimately involves consensual validation.

Consensual validation is well accepted in science [14]: If two (or more) independent observers consistently can agree as to whether or not a phenomenon has occurred, that phenomenon is consensually validated, and others have reason to assume that phenomenon is a real event, adequately defined and measured under conditions allowing for reasonably accurate assessment [3, 6, 8]. For those who use measures of interobserver agreement, as the concomitance between the observations of two raters increases, faith in the data grows apace.

However, scientists conducting nonhuman drug studies

have not commonly used such measures, nor gone to great lengths to ensure in other ways the methodological adequacy of their observational procedures. The consequences of this, if any, are unknown. Methodological aspects of clinical drugs studies have been strongly criticized, in part on the grounds that observers have been aware of experimental conditions and data collected via observational procedures without demonstrated repeatibility or objectivity [1, 15, 16, 17, 25, 26, 27]. The saliency of these criticisms is supported by data as well as logic, for Sulzbacher [26] has reported that the likelihood of a beneficial drug effect being reported in a given study is inversely related to the methodological adequacy of the study. Although similar effects have not been reported for nonhuman drug studies, there is little reason to believe that methodological variables are more influential in clinical than in preclinical research. Surely nothing beyond a bit of time and effort is saved by failing to use blind observers and assess interobserver agreement in nonhuman drug studies. But, if the arguments and data of applied behavior analysts [3, 5, 6, 7, 9, 10, 13] and critics of clinical drug studies [1, 15, 16, 25, 26, 27] are generalizable, considerable harm could result from continuing to ignore these methodological conventions.

REFERENCES

1. Aman, M. G. and N. N. Singh. The usefulness of thioridazine for treating childhood disorders: Fact or folklore. *Am. J. ment. Defic.* **4:** 331–338, 1980.
2. Baer, D. M., M. M. Wolf and T. R. Risley. Some current dimensions of applied behavior analysis. *J. appl. Behav. Analysis* **1:** 91–97, 1968.
3. Bailey, J. *A Handbook of Research Methods in Applied Behavior Analysis*. New York: Plenum Press, 1977.
4. Beatty, W. W. and G. A. Holzer. Sex differences in stereotyped behavior in the rat. *Pharmac. Biochem. Behav.* **9:** 777–783, 1978.
5. Birkimer, J. C. and J. H. Brown. A graphical judgemental aid which summarizes obtained and chance reliability data and helps assess the believability of experimental effects. *J. appl. Behav. Analysis* **12:** 523–533, 1979.
6. Hawkins, R. P. and B. D. Fabry. Applied behavior analysis and interobserver reliability: A commentary on two articles by Birkimer and Brown. *J. appl. Behav. Analysis* **12:** 545–552, 1979.
7. Hawkins, R. P. and V. A. Dotson. Reliability scores that delude: An Alice in Wonderland trip through misleading characteristics of interobserver agreement scores in interval recording. In: *Behavior Analysis Areas of Research and Application*, edited by E. Ramp and G. Semb. Englewood Cliffs, New Jersey: Prentice-Hall, 1975.
8. Irwin, S. Comprehensive observational assessment: A systematic, quantitative procedure for assessing the behavioral and physiologic state of the mouse. *Psychopharmacologia* **13:** 222–257, 1968.
9. Johnson, S. M. and O. D. Bolstad. Methodological issues in naturalistic observation: some problems and solutions for field research. In: *Behavior Changes: Methodology, Concepts, and Practice*, edited by L. A. Hammerlynck, L. D. Handy and E. J. Mash. Champaign, Illinois: Research Press, 1973.
10. Johnson, J. M. and H. S. Pennypacker. *Strategies and Tactics of Human Behavioral Research*. New York: Houghton-Mifflin, in press.
11. Kelly, M. B. A review of observational data-collection and reliability procedures reported in the Journal of Applied Behavior Analysis. *J. appl. Behav. Analysis* **10:** 97–101, 1977.
12. Kent, R. N. and S. L. Foster. Direct observational procedures: Methodological issues in naturalistic settings. In: *Handbook of Behavioral Assessment*, edited by A. R. Ciminero, K. S. Calhoun and H. E. Adams. New York: Wiley, 1977.
13. Kratochwill, T. R. and R. J. Wetzel. Observer agreement, credibility, and judgement: Some considerations in presenting observer agreement data. *J. appl. Behav. Analysis* **10:** 133–139, 1977.
14. Kuhn, T. S. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1970.
15. MacDonald, M. L. and L. L. Tobias. Withdrawal causes relapse? Our response. *Psychol. Bull.* **83:** 448–451, 1976.
16. Marholin, D. and D. Phillips. Methodological issues in psychopharmacological research. *Am. J. Orthopsychiat.* **46:** 477–495, 1976.
17. McNair, D. M. Antianxiety drugs and human performance. *Archs Gen. Psychiat.* **29:** 611–619, 1973.
18. Powell, J., A. Martindale and S. Kulp. An evaluation of time sample measures of behavior. *J. appl. Behav. Analysis* **8:** 463–464, 1975.
19. Repp, A. C., D. M. Roberts, D. J. Slack, C. F. Repp and M. S. Beckler. A comparison of frequency, interval, and time-sampling methods of data collection. *J. appl. Behav. Analysis* **9:** 501–508, 1976.
20. Robbins, T. W. A critique of methods available for the measurement of spontaneous motor activity. In: *Handbook of Pharmacology, Volume 7: Principles of Behavioral Pharmacology*, edited by L. L. Iverson, S. D. Iverson and S. H. Snyder. New York: Plenum Press, 1977.
21. Rosenthal, R. *Experimenter Effects in Behavioral Research*. New York: Appleton-Century-Crofts, 1966.
22. Rosenthal, R. and K. L. Fode. The effect of experimenter bias on the performance of the albino rat. *Behav. Sci.* **8:** 183–189, 1963.
23. Sidman, M. *Tactics of Scientific Research*. New York: Basic Books, 1960.
24. Snow, E. E. Unfinished pygmalion. *Cont. Psychol.* **14:** 197–199, 1969.
25. Sprague, R. L. and J. S. Werry. Methodology of psychopharmacological studies with the retarded. In: *International Review of Research in Mental Retardation Volume 5*, edited by N. R. Ellis. New York: Academic Press, 1971.
26. Sulzbacher, S. I. Psychotropic medication with children: An evaluation of procedural biases in results of reported studies. *Pediatrics* **51:** 513–517, 1973.
27. Tobias, L. L. and M. L. MacDonald. Withdrawal of maintenance drugs with long-term hospitalized mental patients: A critical review. *Psychol. Bull.* **81:** 107–125, 1974.