Oregon Research Institute

The Efficacy of the Spot-Check Procedure in Maintaining the Reliability of Data Collected by Observers in Quasi-Natural Settings: Two Pilot Studies

John B. Reid
and
Barbara DeMaster

ORI RESEARCH BULLETIN
Volume 12 Number 8

The Efficacy of the Spot-Check Procedure in Maintaining the Reliability of Data Collected by Observers in Quasi-Natural Settings: Two Pilot Studies

John B. Reid
Oregon Research Institute

and

Barbara DeMaster
University of Wisconsin

Abstract

Preliminary evaluations of two procedures for the assessment and maintenance of observer reliability were conducted. In the first study, one pair of observers was trained to an acceptable level of reliability and was then told that no further assessment of accuracy would be carried out (no-check procedure). In the second study, two pairs of observers were trained to an acceptable level of reliability and were then told that they would be periodically checked for accuracy and informed when the checks were taking place (spot-check procedure). Actually, the reliability of observers in both studies was continuously monitored. The results of the studies were as follows: (a) the accuracy of observers in the no-check condition dropped dramatically immediately after the termination of overt assessment; (b) the accuracy of observers in the spot-check condition dropped after initial training and overt assessment but exceeded the training level of reliability during each of the spot-check sessions; and (c) the magnitude of the drop in accuracy was smaller for observers in the spot-check study than for those in the no-check study. On the basis of these results, it was argued that the spot-check procedure for checking and maintaining observer reliability is of questionable validity.

The Efficacy of the Spot-Check Procedure in Maintaining the Reliability of Data Collected by Observers in Quasi-Natural Settings: Two Pilot Studies[1]

John B. Reid
Oregon Research Institute

and

Barbara DeMaster
University of Wisconsin

The usefulness of most reported research having to do with the application of laboratory-based operant principles to the clinical modification of disturbed behavior hinges in large part upon the ability of observers to collect veridical data in quasi-naturalistic settings. The empirical measure of observer accuracy is typically some statistic (e.g., the percent agreement or a coefficient of correlation) which describes the level at which the observers agree when they are simultaneously but independently recording the behavior of the same subject(s). Because of such considerations as expense and the probable reactivity or obtrusiveness of the observational process, observers often work singly rather than in pairs after their initial training, thereby preventing the continuous monitoring of observer accuracy.[2]

In lieu of the continuous monitoring of observer accuracy throughout data collection, attempts have been made to estimate the veridicality of observation data by carrying out observer-reliability-assessment either before the colelction of usable data (e.g., Arrington, 1943; Merrill, 1946; Patterson & Reid, 1970) or at some point(s) during data collection (e.g., Hartup & Coates, 1967; Buell, Stoddard, Harris, & Baer, 1968; Hall, Lund, & Jackson, 1968; Hart, Reynolds, Baer, Brawley, & Harris, 1968; Walker & Buckley, 1968). Henceforth, the former procedure will be termed a "no check" and the latter will be termed a "spot check." Two features are common to both procedures: First, only a minor portion

of the usable data is monitored for accuracy; second, the observers are typically aware of the reliability assessment when it is taking place (indeed it would be impossible to conduct covert reliability checks in many observation settings).

In order to attribute meaning to the reliability statistics reported in studies using the no-check or spot-check procedure, it must be assumed that observer performance is similar during overt reliability assessment and during unmonitored observations. In the case of the no-check procedure, it must be assumed that once an observer demonstrates reliable performance during overt assessment, his unmonitored performance remains reliable thereafter; in the case of the spot-check procedure, it must be assumed that unmonitored observations between overt reliability assessments are accurate.

Studies recently reported by Reid (1970) and Romanczyk, Kent, Diament, and O'Leary (1971) have brought the no-check procedure into serious question. In the Reid study, observers demonstrated an average drop of 25 percentage points in observer-criterion agreement from the end of training and overt reliability assessment to the very first day of covert assessment. The code used in this study was rather complex (i.e., 33 categories) and the observers were relatively naive (i.e., the median length of training was six hours). In the Romanczyk et al. study, an even greater drop in reliability was evidenced from overt to covert assessment. In this latter experiment, a simpler code (i.e., nine categories) and more experienced observers (i.e., at least three months' prior training and experience) were studied. These data suggest strongly that the reliability of observation data collected via the no-check procedure is significantly overestimated in the research literature.

Two studies are reported below: The first is a replication study of Reid (1970) in a naturalistic setting; the second represents an initial attempt to evaluate the usefulness of the spot-check procedure for insuring and assessing observer reliability.

Study I

Two student assistants were trained in the use of an observation code (Reid, 1967) for the collection of social interaction data on a group of 15 preschool children in a free-play setting. The observational code was designed to yield a running account of the behavior of an individual in terms of 21 categories and the reactions of others to him in terms of 12 categories. Examples of the behavior categories were Hug, Cry, Talk, Physical Aggression, Play, and Yell. Examples of the reactions were Interest, Praise, Disapproval, and Physical Restraint. All categories were assigned symbols to allow for rapid recording. Prior experience with the code (Reid, 1967; 1970) had shown that it was rather easily learned by naive observers.

Observers were trained over a three-week period to a criterion of 80% agreement (entry-by-entry analysis of protocols made while the observers were simultaneously, but independently, observing the same children) over three consecutive days. Each observation session lasted for one hour, during which the observers watched ten children for five minutes each. The observers were instructed not to talk to one another regarding the data collection as observer bias would result. During training and post-training observations, the observers were separated so that they could not see one another, and they were given lists of the children (and their order) to be observed before a session began. Reliability was calculated at the end of each session by taking the last three pairs of five-minute observations and comparing them. At the end of this training

and assessment phase, observers were told that they would be observing different subjects at any given time and that they should be very careful to observe accurately as there would be no way to make an objective assessment of their reliability from that point onward.

Upon completion of training and assessment, the observers were again assigned individual subjects to observe. As during training, ten observations were made each day. The observers used the same room, at the same times, but they were told that they were recording the behavior of different children during any given observational period. Unknown to the observers during the last three five-minute observations per day they were in fact watching the same child at the same time. In effect, this produced a covert check on the interrater reliability of the observers after training and overt assessment. This covert assessment phase was continued for ten days. Inspection and comparison of the data for overt and covert assessment phases (see Figure 1) shows that the level of agreement between the observers fell from 82% on the last day of training to 27% on the first day of covert assessment. On the nine succeeding days, inter-observer agreement varied from 23% to 42%. None of the data taken during the covert assessment phase came close to meeting the predetermined criterion level of 80% agreement.

- - - - - - - - - - - - - - - - -

Insert Figure 1 about here

- - - - - - - - - - - - - - - - -

The results of this modest study replicate the findings of Reid (1970) and Romanczyk, et al. (1971) and indicate the magnitude of error possible in the estimation of observer accuracy via the no-check procedure.

Study II

Using the same behavior code and a lower training criterion (70% inter-observer agreement), two pairs of observers were initially trained and assessed as in the previous study. Pairs were trained and assessed separately. At the completion of training, the pairs were given a pep talk about the importance of continuing their high level of care and accuracy. They were also told that they would be assessed periodically for reliability but that they would know the time of each assessment. After training, the covert assessment procedure used in the first study was instituted. Over the eleven succeeding observation sessions, two "spot checks" were carried out. These spot checks consisted of telling the observers that they were observing the same children simultaneously and that their level of agreement would be calculated and checked at a later date. (To make sure they were cognizant of the manipulation, they wrote the letters T.B.C.--To Be Checked--at the top of each rating sheet.)

The results of this study are presented in Figure 2. Inspection of the data reveals an immediate mean drop of four percentage points in agreement for the two pairs from the last day of training to the first day of covert assessment, and a further decrease of 10% on the second day of covert assessment. While the mean agreement on covert assessment days varied from 57% to 66%, on spot-check days mean inter-observer agreement exceeded training criterion levels (i.e., 76% and 72%).

- - - - - - - - - - - - - - - -

Insert Figure 2 about here

- - - - - - - - - - - - - - - -

This study was successful in reproducing the findings of the previous one, namely that significant observer decrement took place after the terimination of overt assessment. Moreover, this study revealed serious

flaws in the spot-check procedure, both as a device for maintaining the accuracy of observations after training and as a device for estimating the accuracy of the post-training, unmonitored observation data. That is, the spot-check procedure may well serve only to raise reliability on spot-check days alone.

Comparing the results of the two studies, it is evident that the spot-check procedure produced less observer drift and provided a better estimate of "unmonitored" observer performance than did the no-check procedure. It is possible that the expectation of intermittent assessment held by observers in the spot-check procedure kept them better motivated and more alert between checks.

An interesting accidental finding from this study is worth reporting. On the seventh day of covert assessment, the four observers, all enrolled in a psychology course at the University of Wisconsin, were given a lecture which included a lengthy reference to a study on observer drift using a deception format previously conducted by the first author (Reid, 1970). The teaching assistant for the course (the second author) noted that the observers looked quite surprised by that information. Later that day, when the observers did their scheduled observations, their inter-observer agreement was near criterion level, but the next day their reliability dropped again (see Figure 2). It seemed that even suspicion of the manipulation was unsuccessful in keeping them at the predetermined acceptable level of accuracy!

Discussion

The present investigation adds support to the contention that the accuracy of observation data is over-estimated by the no-check procedure, and brings into question the efficacy of the spot check as a strategy

for maintaining or estimating the continued reliability of observers after training. The results suggest that the spot-check method is somewhat superior in that less observer drift was observed, compared with the no-check procedure. It should be kept in mind that the present investigation utilized only a small number of observers and a rather complex code, making the implications of this study for observational research using simpler codes somewhat unclear. It is quite clear, however, that more research in this area is warranted.

References

Arrington, R. E. Time sampling studies of social behavior: A critical review of technique and results with research suggestions. Psychological Bulletin, 1943, 40, 81-124.

Buell, J., Stoddard, P., Harris, F. R., & Baer, D. M. Collateral social development accompanying reinforcement of outdoor play in a preschool child. Journal of Applied Behavior Analysis, 1968, 1, 167-173.

Hall, R. V., Lund, D., & Jackson D. Effects of teacher attention on study behavior. Journal of Applied Behavior Analysis, 1968, 1, 1-12.

Hart, B. M., Reynolds, N. J., Baer, D. M, Brawley, E. R., & Harris, F. R. Effect of contingent and non-contingent social reinforcement on the cooperative play of a preschool child. Journal of Applied Behavior Analysis, 1968, 1, 73-76.

Hartup, W. W., & Coates, B. Imitation of a peer as a function of reinforcement from the peer group and rewardingness of the model. Child Development, 1967, 38, 1003-1016.

Merrill, B. A measure of mother-child interaction. Journal of Abnormal and Social Psychology, 1946, 41, 37-49.

Patterson, G. R., & Reid, J. B. Reciprocity and coercion: Two facets of social systems. In C. Neuringer & J. Michael (Eds.), Behavior modification in clinical psychology. New York: Appleton-Century-Crofts, 1970. Pp. 133-177.

Reid, J. B. Reciprocity in family interaction. Unpublished doctoral dissertation, University of Oregon, 1967.

Reid, J. B. Reliability assessment of observation data: A possible methodological problem. Child Development, 1970, 41, 1143-1150.

Romanczyk, R. G., Kent, R. N., Diament, C., & O'Leary, K. D. Measuring the reliability of observational data: A reactive process. Paper presented at the Second Annual Symposium on Behavior Analysis, Lawrence, Kansas, 1971.

Walker, H. M., & Buckley, N. K. The use of positive reinforcement in conditioning attending behavior. Journal of Applied Behavior Analysis, 1968, 1, 245-250.

Footnotes

1. The studies reported here were conducted at the University of Wisconsin and were supported in part by a grant to the first author from the Wisconsin Alumni Research Foundation. The preparation of this paper was supported by grants MH 15985 and MH 12972 from the National Institute of Mental Health, U.S. Public Health Service. Reprints may be obtained from John B. Reid, Oregon Research Institute, P.O. Box 3196, Eugene, Oregon 97403.

2. In fact, it is not uncommon to find no data at all describing inter-observer reliability in behavior modification studies.

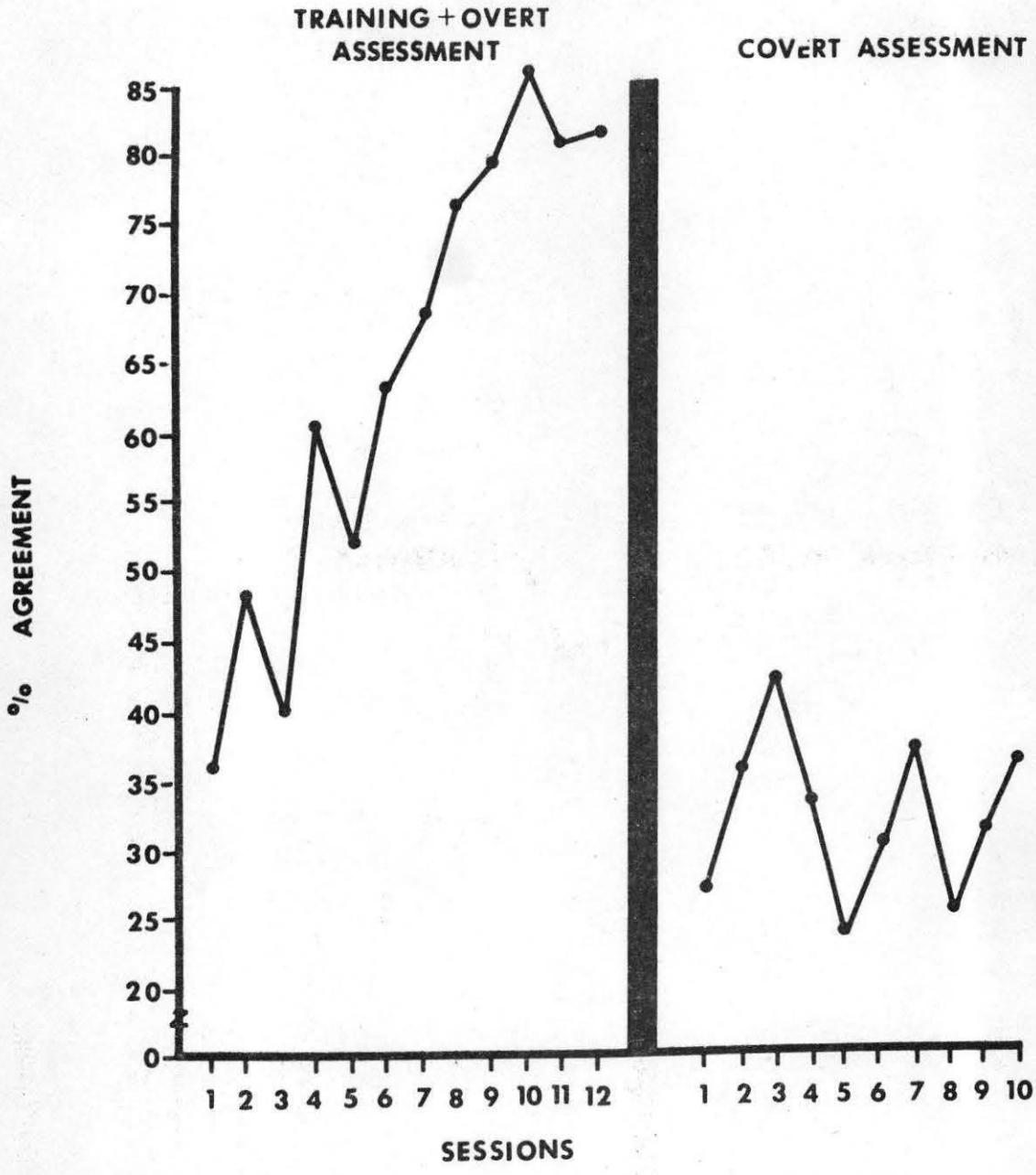
Figure Captions

Figure 1: Percent agreement between two observers.

Figure 2: Mean percent agreement for two pairs of observers.

FIGURE 1

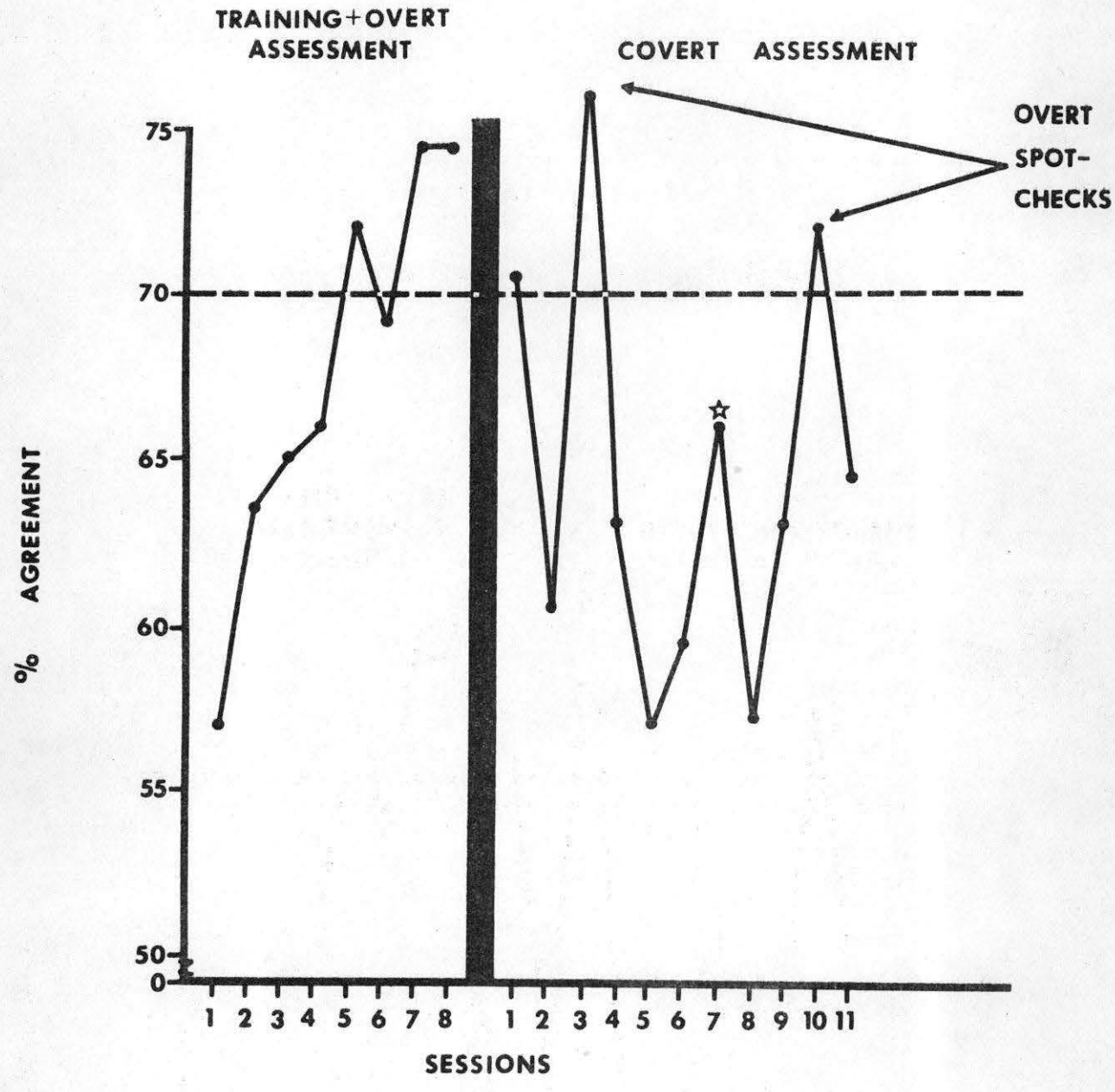
AGREEMENT BETWEEN TWO OBSERVERS*



*EACH DATA POINT REPRESENTS A COMPARISON OF FIFTEEN MINUTES OBSERVATION DATA

FIGURE 2

AGREEMENT FOR TWO PAIRS OF OBSERVERS



☆Os HEARD LECTURE ON RELIABILITY PRECEEDING THIS SESSION