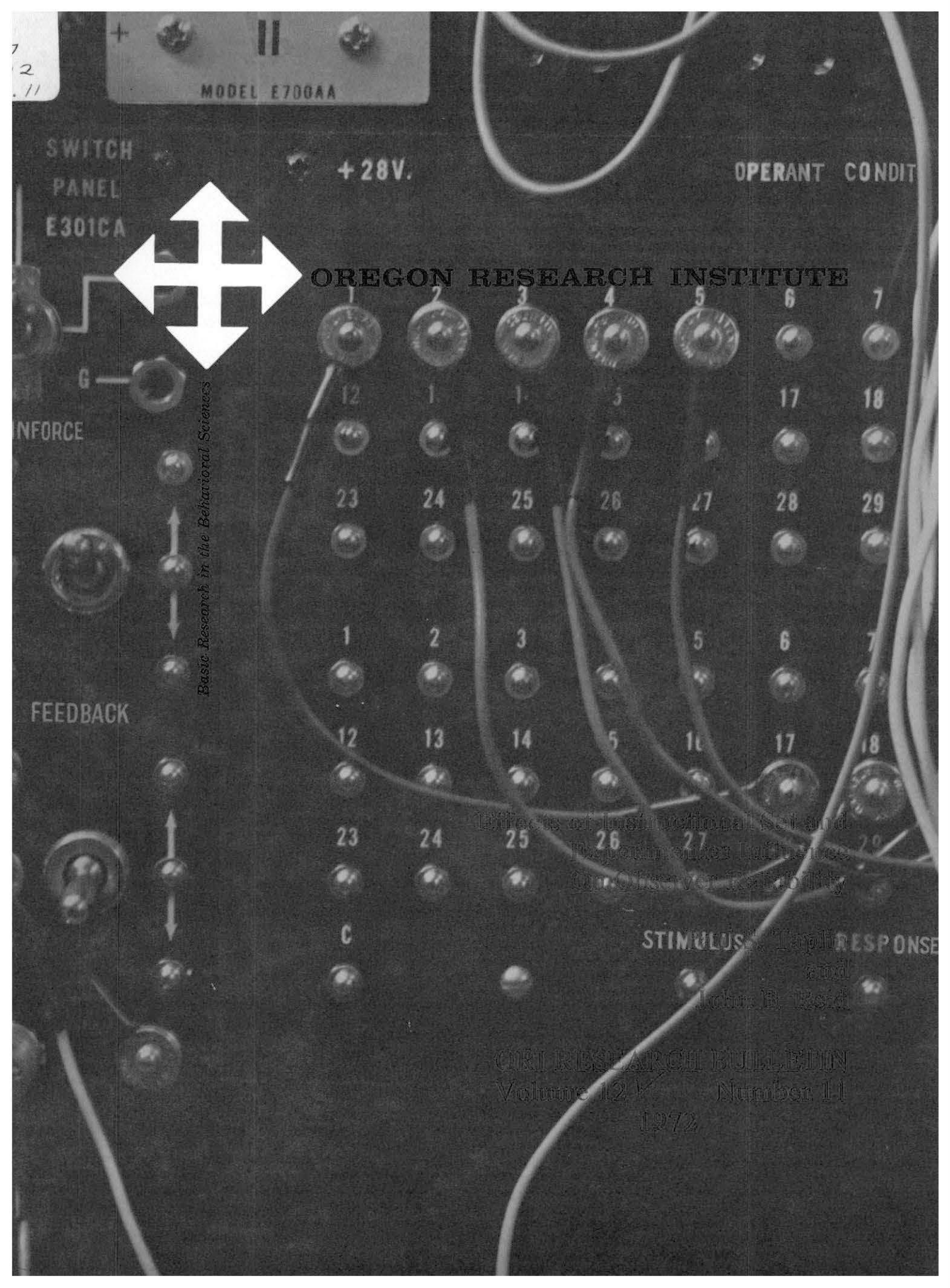


Effects of Instructional Set and Experimenter Influence on Observer Reliability

STIMULUS, Inc. RESPONSE

and

John B. Reid

ORI RESEARCH BULLETIN

Volume 12 Number 11

1972

Effects of Instructional Set and Experimenter Influence On Observer Reliability

Paul S. Taplin and John B. Reid

University of Wisconsin

Abstract

A laboratory analog of naturalistic observation was used to examine the relationship of observer drift to instructional set and experimenter status. Three instructional sets (no check, random check, and spot check) and two levels of experimenter status were studied. Results indicated a highly significant decrease in observer reliability coinciding with the shift from training to data collection. This performance decrement was observed in all three instructional set conditions. Within the spot-check condition, reliability on spot-check days was found to be significantly greater than mean reliability immediately before and after spot checks. Further results revealed that observers trained by the high status experimenter performed less reliably than observers trained by the other two experimenters. The possible implications of these results for future observational research, and suggestions for minimizing observer drift were discussed.

Child Development, 1973, in press.

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 observational 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. (1971) 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 over-estimated in the research literature.

If this observer drift phenomenon is common to observation data collection procedures in general, then the conclusions of a significant body of research are based on data of unknown reliability.

The main purpose of the present experiment was to examine the relationship between instructional set and magnitude of observer drift in the analog situation described by Reid (1970). The term "instructional set" as used here refers to an observer's notion of the conditions under which observational performance will be evaluated. Three different instructional sets were investigated: (a) a no-check set, (b) a random-check set, and (c) a spot-check set.

The no-check instructions were identical to those used by Reid (1970). Observers in this group were instructed that their data collection performance would not be checked by the experimenter. In contrast to Reid's study, observers in this group used a simpler behavior rating code and observed a different stimulus tape. It was expected that the no-check group in this study would show observer drift of a magnitude similar to, or a bit smaller than, that reported by Reid (1970).

The random-check instructional set was examined as a possible means of reducing the magnitude of observer drift. Results of quality control research (e.g., Stok, 1965), as well as descriptions of management supervisory procedures (e.g., Langer, 1970; McGregor, 1960), suggest that work productivity and quality of performance in a variety of job settings are increased when employees are aware that they can be observed by

management at any time. Thus, in the present study, the random-check group performed under the notion that any part of their work could be checked by the experimenter. It was expected that this instructional set would reduce or eliminate observer drift.

The third group of observers was included to provide experimental evidence on the effectiveness of a spot-check instructional set used by some researchers as a means of maintaining observer reliability at criterion level (e.g., Berk, 1971; Buell et al., 1968; Hall et al., 1968; Hart et al., 1968; Hartup & Coates, 1967; Walker & Buckley, 1968). The spot-check procedure, as employed in the above studies, consists of periodic checks of observer reliability made by having a second observer simultaneously rate the same behavior. Usually no attempt is made to conceal the presence of the second observer, so that the spot check is actually an overt reliability assessment. In the present study, observers in the spot-check group were instructed that their data collection performance would be evaluated only during announced spot-check sessions. Since such spot-checks are overt reliability assessments, it was predicted that performance during spot checks would be close to criterion level, but that performance before and after each spot check would be less reliable than spot-check performance.

A secondary purpose of the present study was to explore the possibility of experimenter influence on observer drift. Research on the social psychology of psychological experimentation (e.g., Orne, 1962; Rosenthal, 1964) suggests the possibility that observer drift may be affected by interpersonal characteristics of the experimenter. The present study investigated this possibility in regard to the variable of

experimenter status. Three experimenters, one of high professional status (a faculty member) and two of lower professional status (graduate students), each trained and directed one-third of the observers. Group differences in observer performance as a function of experimenter status were analyzed. Since little was known about the observer drift phenomenon, no predictions were made regarding the effects of experimenter influence on observer drift.

Method

Subjects

Eighteen female undergraduates at the University of Wisconsin served as observers (Os) in this study. Six other students who failed to master the observational code were dropped from the study. All Os were paid volunteers.

Three Es, experienced both in collecting observational data and in training Os, participated in the present study. The Es differed in terms of status. A university professor (E_1) was identified as the head of the research project. Two graduate students (E_2 and E_3) were identified as project assistants.

Apparatus

A video tape apparatus (Reid, 1970) was used (1) to film the interactions of a group of three mothers and their four children and (2) to present this film to Os in this study.

The stimulus tape was filmed in a suburban living room following a warm-up period of about twenty minutes to allow the video Ss to adapt to

the video equipment and to interact naturally. The completed sixty-minute tape was marked off into twelve five-minute segments. Each five-minute segment was further divided into ten thirty-second intervals, by means of audio signals recorded on the tape.

The behavior code and behavior coding sheets used in this study were the same as those used by Reid (1970) except that they were simplified to yield a running account of the behavior of an S in terms of 19 rather than 21 categories, and the reactions of others toward him in terms of 7 rather than 12 categories.

Prior to the commencement of the experiment, E_2 and E_3 used the behavior code to construct criterion protocols which would serve as standards for the measurement of observer reliability. The protocols were prepared for five of the seven video Ss, identified as target Ss.[1] The criterion protocols consisted of individual behavior codings of the five target Ss in each five-minute segment of the tape. Thus, five criterion protocols were prepared for each five-minute segment (one for each target S), with a total of sixty criterion protocols in all.

These protocols were prepared according to the following procedure. E_2 and E_3 made simultaneous, independent codings of a particular target S in a particular segment. Points of agreement and disagreement were identified and then the segment was replayed with each E re-examining his codings and stopping the tape at points of disagreement or uncertainty. This process was continued until a mutually agreeable criterion had been produced. This procedure was followed in preparing each of the sixty criterion protocols.

Procedure

Os were recruited by means of an announcement which explained that part-time research assistants were needed to work as Os in a study on mother-child interaction for twelve consecutive days at $1.50 per hour. Each E worked exclusively with the Os whom he hired.

Before any observational work began, each O attended a thirty-minute interview with one of the Es, during which the following description of the project was presented:

> We are conducting an observational study concerning mother-child interaction. We have a series of video-tape recordings of the behavior of mothers and their children. We have hired you to make observational codings of the behavior on these tapes, by means of a behavior coding system which we will train you to use. It is necessary that we train naive observers to code the tapes rather than coding them ourselves because we might easily bias the codings so that they would support the hypotheses which we are predicting. For this reason, also, we cannot reveal to you at this time the exact hypotheses which we are investigating.
>
> We will train you to use the behavior code reliably. We have prepared accurate criterion codings of some video-tape segments, which we will use to assess the accuracy of your codings throughout the training period. We will continue to train you until your performance reaches a minimum level of 70% agreement with the criterion codings for two consecutive days. After completing training you will be observing

new segments of tape which no one else will be coding. Your codings of these segments will comprise the basic data for our experiment, so you must continue to be very careful and very accurate.

The interview was concluded with a brief description of the behavior coding system. Each O received a copy of the coding manual and was instructed to memorize the code abbreviations before returning for the first training session.

Instructional sets. Prior to an O's first observation session she was assigned to one of three experimental conditions (i.e., instructional sets). Assignment was random, with the restriction that each E run two Os in each of the three conditions. At the first observation session O received instructions appropriate to her experimental group.

No-check group. In the no-check group, O received the basic instructions presented during the initial interview. These Os were told that they would be trained to a criterion level of reliability and then put on their own to code new video-tape segments. It was emphasized that after completion of training observations would no longer be monitored for reliability.

Spot-check group. Spot-check Os were told that at some time after training, when they were coding new segments of video-tape, spot checks of their reliability would be conducted to determine whether their codings were sufficiently accurate. It was explained that the reliability of the spot-check observations would be determined some time after all the data had been collected.

Random-check group. These Os were instructed that a random selection

of about 20% of their behavior coding sheets would be checked for reliability against protocols which the E would make some time after all observations had been completed. It was explained that this would be done to insure that Os had continued to use the behavior code reliably. It was stressed that E's protocols could not be used as experimental data because of the possibility of experimenter bias.

Observation sessions. Observation sessions were held in a small laboratory room which contained a video-tape recorder and monitor set up on a table, two chairs, and a cabinet containing a small library of video-tapes. Os came to this room for one observation session per day. Seated in a chair in front of the monitor, O wrote on behavior coding sheets which were placed on the table in front of the monitor. During training, observation sessions lasted approximately sixty minutes. After training, sessions were completed in approximately thirty minutes.

At each of the twelve observation sessions, the O was shown a new five-minute segment of the stimulus tape. The segments were presented in the order in which they were filmed. During each observation session observers made individual codings of each of the five target Ss. The order in which target Ss were observed was randomly determined.

Training sessions: Overt reliability assessment. The first four or five observation sessions were regarded as training sessions, during which O's task was to master the behavior coding system. Overt reliability measurements, consisting of per cent agreement scores between the observer's protocol and the criterion protocol, were taken during each training session. All Os were instructed that training would be terminated when

an average reliability score of at least 70% had been maintained for two consecutive days.

During training, the behavior codings were made under either training conditions or reliability assessment conditions. The first two of the five randomly ordered target subjects were coded under training conditions. During the training sessions, E guided O, stopping the tape at difficult passages to give immediate feedback, re-playing particularly troublesome passages, and discussing problems in using the coding system. The remaining three target Ss were coded under reliability assessment conditions. These observations were done independently by O: E began running the tape and then sat down in the far corner of the room, giving no feedback whatsoever, until the segment ended. At that point E quickly computed a reliability score by making an entry-by-entry comparison between the just-completed protocol and the appropriate criterion protocol. After the final three target Ss had been coded, the experimenter computed an overall reliability score for the three ratings and gave this information to O.

Data collection sessions: Covert reliability assessment. At the first observation session following the achievement of criterion performance in training, Os were instructed that they were now ready to collect data by observing new, uncoded video-tape segments. Appropriate treatment group instructions were also repeated at this time. It was emphasized that O must remain careful and accurate in her ratings because her protocols would be the only source of data for the particular video-tape segments which she observed.

Under covert reliability assessment conditions all codings were made with no feedback provided by E, who was present in the laboratory only to

operate the video-tape apparatus and to assign target Ss to be coded. For Os in the no-check and random-check groups, all data collection sessions were conducted under covert reliability assessment conditions.

Spot-check reliability assessment. In the spot-check group, Os performed under covert reliability assessment conditions in all but the third and sixth data collection sessions, which were spot-check sessions. In these sessions, Os were informed that behavior codings would be checked for reliability. They were instructed to write the words "to be checked" on the top of each behavior coding sheet completed during a spot-check session.

At the completion of the twelfth and final observation session, each O was given a cover story concerning the nature of the experiment (Reid, 1970) and paid for her work.

Results

Observer performance was assessed by an entry-by-entry comparison of the last three behavior coding sheets in each session with the appropriate criterion protocols, yielding a percent agreement score. Calculation of the percent agreement score was as follows: 100 (numbers of identical entries/number of identical entries + number of errors). The reliability of this scoring procedure, in terms of a product-moment correlation between independent scorings of a random selection of 10% of the behavior rating sheets, was +.97.

Mean percent agreement scores for the three treatment groups during the final two training sessions and the seven subsequent data collection sessions are presented in Figure 1. An immediate drop in performance

---

Insert Figure 1 about here

---

coinciding with the shift from training to data collection conditions is evident in all three groups. Table 1 summarizes the performance of all

---

Insert Table 1 about here

---

18 Os across sessions.

An analysis of variance for the percent agreement scores (Table 2),

---

Insert Table 2 about here

---

using arc sin transformations, revealed two significant sources of variability. First, observer reliability varied significantly across sessions (F = 15.74; 8, 72 d.f.; $p < .01$). Comparison of the final training session with the first covert reliability session showed that the abrupt decrement in observer performance was highly significant, in the predicted direction (t = 6.29; 72 d.f.; $p < .0005$, one-tailed).

Second, the analysis of variance indicated a significant experimenter effect (F = 5.97; 2, 9 d.f.; $p < .025$). Pairwise comparisons of the three experimenter means using Tukey's HSD test revealed that the mean of E_1 (high status) was significantly smaller than the means of E_2 and E_3 (HSD = .0869; 9 d.f.; $p < .05$, two-tailed).

Performance of the spot-check group appears to support the prediction that observer reliability will rise during spot checks and diminish before

and after spot checks. The corrected t-test developed by Cochran and Cox (1957) indicated that observer reliability during spot checks was significantly greater than mean reliability before and after spot checks ($t_1 = 2.22$; $t_2 = 2.05$; $p < .05$, one-tailed).[2] The same analysis for the no-check and random-check groups revealed no significant differences.

Overall mean differences in observer reliability among instructional set groups were not significant.

Discussion

The results of this study point to the presence of significant observer drift: observer reliability decreased by an average of 16% when observers were shifted from training to data collection conditions. The lack of significant differences among treatment groups in terms of magnitude of drift or quality of data collection performance attests to the strength of this observer drift phenomenon. The unique performance of the spot-check group reveals an additional aspect of observer drift, demonstrating that high reliability during spot checks does not insure high reliability during unchecked sessions. All of this evidence favors a rejection of the hypothesis that observer performance during conditions of overt reliability assessment is representative of unmonitored observer performance.

A logical starting point in the search for determinants of observer drift is an examination of observer training programs which are currently being used. Observers who graduate from these programs simply may not be learning the behavior coding procedures sufficiently well to accurately code new behavior. For example, in observer training programs which have

a criterion requirement of one or two sessions of high reliability, observers may tend to achieve criterion performance on relatively simple stimulus material and thus be sent out to collect data with an inadequate mastery of the behavior coding system.

Data which may offer some suggestive support for this hypothesis were analyzed after the completion of this study. Each of the sixty five-minute criterion protocols was scaled for complexity, which was defined as the ratio of the number of different code items used over the total number of entries. Thus, for each five-minute segment of behavior which was coded by each observer during the final two days of training and the seven subsequent data collection days, an entry-by-entry reliability score (used in previous analyses) and a complexity score were available. These two sets of scores were then correlated, resulting in $r_{xy} = -.52$ ($p < .001$). This negative correlation indicates that there was a moderately strong tendency for reliability to fall as the complexity of the observed interaction increased. Similarly, Skindrud (1972) obtained significant positive correlations of +.53 and +.65 between mean observer reliability on two data scores and the percent of consecutive repeated interactions within an observed behavior segment (the latter score measuring the simplicity of the behavior being observed--the inverse of complexity.) In view of these findings, a more meaningful indication of observer performance during training might be some measure which combines observer accuracy with the complexity of the material being coded.

It should also be noted that the results of this study as well as the Reid (1970) study indicate that observer drift is not a continuous decline in performance, but rather an abrupt decrement followed by a

rather stable level of performance (with the exception of the spot-check group). This suggests that one possible method of avoiding the problem of observer drift might be to over-train observers (particularly on the behavior codes of major interest) up to a level of reliability such that a significant drop in performance would not bring them below an acceptable level of accuracy.

The performance of the spot-check group may offer some additional clues to possible methods of minimizing observer drift. The increases in reliability on spot-check days provide evidence that a seemingly minimal difference in instructional set can have a significant effect on observer performance. Furthermore, the finding that observers increased significantly in reliability on days when they were told that they would be checked for reliability suggests that observers may be capable of maintaining criterion performance if the experimenter is able to specify and implement the factors underlying the spot-check phenomenon.[3]

The problem of training and maintaining reliable observers is clearly a complex one. An additional source of complexity revealed by the results of this study is the role of the experimenter or trainer. In this study, $\underline{E}_1$ trained observers who proved to be less reliable throughout the experiment than observers trained by the other experimenters. Perhaps $\underline{E}_1$, who conducted the initial investigation of observer drift, somehow communicated to observers his expectation that their reliability would diminish and this communication resulted in poorer performance in both the training and data collection phases of the experiment. At any rate, the results of this study indicate that the experimenter must be considered as a relevant variable in dealing with the problem of observer drift.

In conclusion, the major thrust of the findings of this study is to point to the danger of assuming that unmonitored observational data are reliable. Rather than being established a priori, the reliability of observational data ought to be empirically demonstrated. Without such demonstrations, the accuracy of observational data can be only a matter of faith.

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.

Beckwith, L. Relationships between infants' social behavior and their mothers' behavior. Child Development, 1972, 43, 397-411.

Berk, L. E. Effects of variations in the nursery school setting on environmental constraints and children's modes of adaptation. Child Development, 1971, 42, 839-869.

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.

Cochran, W. G., & Cox, G. M. Experimental designs. New York: John Wiley and Sons, 1957.

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.

Langer, E. Inside the New York telephone company. The New York Review of Books, 1970, 14, 16-24.

McGregor, D. The human side of enterprise. New York: McGraw Hill, 1960.

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

Orne, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. American Psychologist, 1962, 17, 776-783.

Osofsky, J. D., & O'Connell, E. J. Parent-child interaction: Daughters' effects upon mothers' and fathers' behaviors. Developmental Psychology, 1972, 7, 157-168.

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

Reid, J., & DeMaster, B. The efficacy of the spot-check procedure in maintaining the reliability of data collected by observers in quasi-natural settings: Two pilot studies. Oregon Research Institute Research Bulletin, 1972, 12, No. 8.

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.

Rosenthal, R. The effect of the experimenter on the results of psychological research. In B. A. Maher (Ed.), Progress in experimental personality research. Vol. 1. New York: Academic Press, 1964.

Skindrud, K. An evaluation of observer bias in experimental-field studies of social interaction. Unpublished doctoral dissertation, University of Oregon, 1972.

Stok, T. L. The worker and quality control. Ann Arbor: University of Michigan, 1965.

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

This study was supported in part by Grants MH 10822 and R01 MH 15985, and by a grant from the Wisconsin Alumni Research Foundation. The authors wish to thank Alan Penn who served as E_3 in this experiment. Both authors are currently at Oregon Research Institute, P.O. Box 3196, Eugene, Oregon 97403.

1. Criterion protocols were not prepared for the two infants who appeared on the video tape because of their relative inactivity.

2. The criticial t for these comparisons, computed according to the procedure developed by Cochran and Cox (1957) is 1.684.

3. This pattern of performance by observers in the spot-check group has also been found, and in more exaggerated form, for spot-check observers in a small field study by Reid and DeMaster (1972).

Table 1

Performance of Observers (% Agreement with Criterion)
during Training and Data Collection Sessions

End of training			Data collection						
Session:	1	2	1	2	3*	4	5	6*	7
Mean	.78	.81	.65	.67	.71	.68	.57	.65	.61
Range	.72-.87	.71-.94	.42-.75	.47-.79	.54-.89	.47-.81	.41-.68	.46-.86	.46-.74

* Includes spot checks

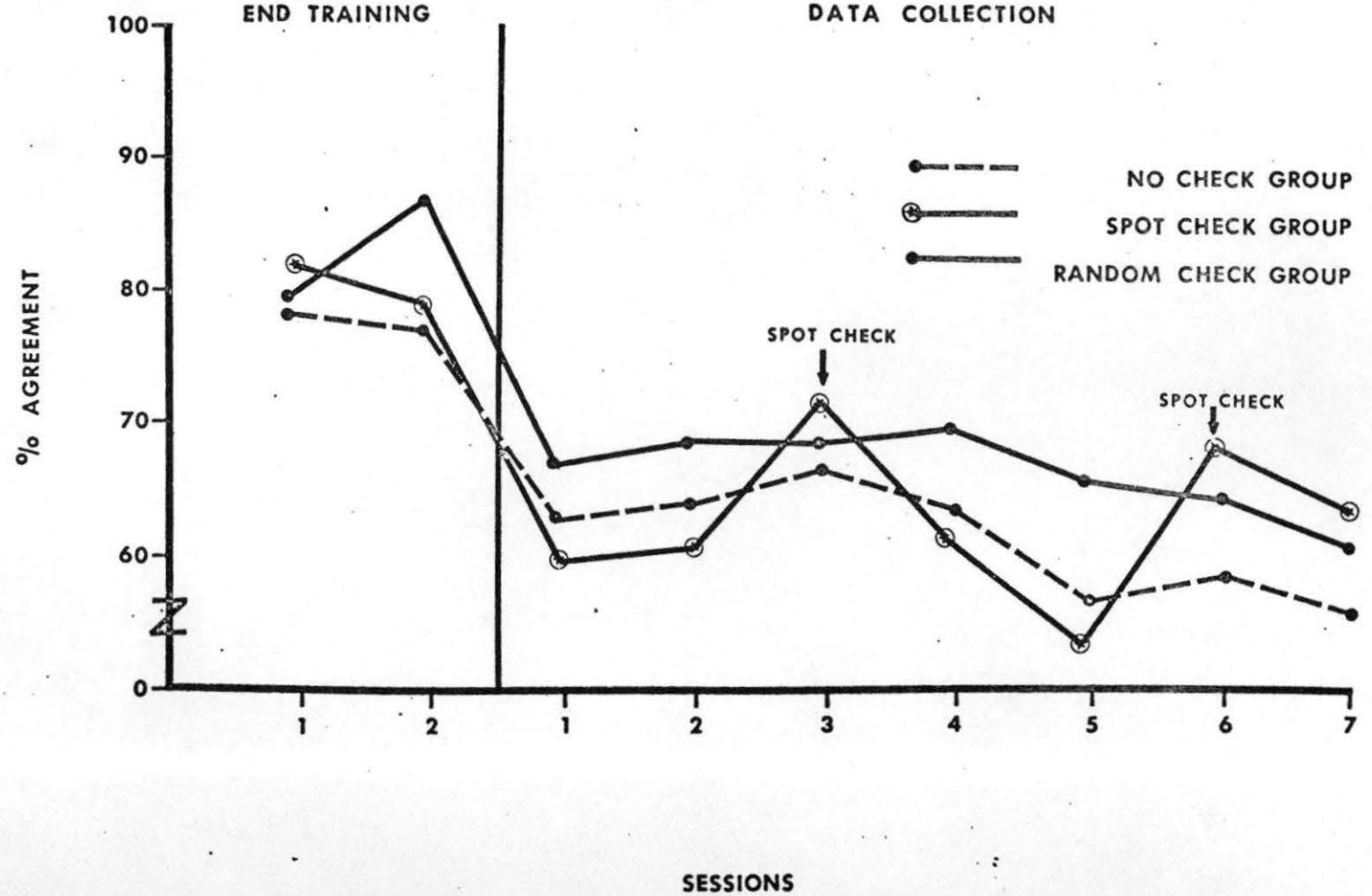
Table 2

Analysis of Variance for Percent Agreement Scores
(Arc Sin Transformations)

Source	SS	DF	MS	F	
A (Experimenter)	.3130	2	.1565	5.973	.025
B (Instructional set)	.1378	2	.0689	2.630	ns
C (Sessions)	4.4702	8	.5588	15.741	.01
AB	.1727	4	.0432	1.649	ns
AC	.7537	16	.0471	1.327	ns
BC	.4646	16	.0290	.817	ns
ABC	1.1351	32	.0355	1.000	ns
S(AB)	.2359	9	.0262		
S(AB)C	2.5537	72	.0355		
Total	10.2367	161			

Figure Caption

Fig. 1. Mean agreement with criterion for three instructional groups over sessions.


END TRAINING
DATA COLLECTION
100
90
80
70
60
0
% AGREEMENT
NO CHECK GROUP
SPOT CHECK GROUP
RANDOM CHECK GROUP
SPOT CHECK
SPOT CHECK
1
2
1
2
3
4
5
6
7
SESSIONS