*SHAPING DATA COLLECTION CONGRUENT WITH
EXPERIMENTAL HYPOTHESES¹*K. DANIEL O'LEARY, RONALD N. KENT, AND JAY KANOWITZ²

STATE UNIVERSITY OF NEW YORK AT STONY BROOK

The influence of instructions and feedback from an experimenter on observational recordings of disruptive behavior was evaluated. Four subject-observers recorded four categories of disruptive behavior from videotapes of children in a classroom setting. Two sets of videotapes, labelled "baseline" and "treatment", were matched for rates of disruptive behavior in each category. The observers were told that two target behaviors were expected to decrease and the other two control behaviors were not expected to change during the treatment condition. During observational recording of treatment videotapes, the observers were given positive feedback when they reported decreases in the two target behaviors, and negative feedback when they reported either no change or increases in the two control behaviors. The target behaviors were recorded less frequently by observers, while recordings of control behaviors were unaffected during the treatment phase. These results suggest that contingent feedback to observers should be avoided as a possible source of bias in behavioral recordings.

DESCRIPTORS: feedback instructions, expectations, observer bias reliability, classroom, observer(s)

Bias in observational recordings of behavior has been reported in a variety of studies, reviewed by Rosenthal (1966, 1969), that have involved tasks such as rating the amount of success or failure shown in photos of faces (Rosenthal, Friedman, Johnson, Fode, Shill, White, and Vikan-Kline, 1964), or counting the number of turns or contractions by planaria (Cordaro and Ison, 1963). However, as the specificity of judgemental criteria in these studies was substantially less than that for recording of behavior in many treatment studies, it is uncertain whether the substantial bias due to expectation documented by Rosenthal is also characteristic of other behavioral recordings.

Scott, Burton, and Yarrow (1967) reported data suggestive of observer bias in a field-experimental treatment program. Observations

made by the principal investigator were compared with those of several uniformed observers, revealing that data of the principal investigator more strongly confirmed the experimental hypothesis. Unfortunately, the use of only one informed observer made it impossible to determine the source of bias in this investigation.

Kent, O'Leary, Diamant, and Dietz (1974) found that differential predictions given to observers influenced subjective global evaluations of children's disruptive behavior. In contrast, objective and reliable recordings of specific behavioral categories were unaffected by the differential predictions. These results suggest that current observational recording technology may produce unbiased measures, regardless of experimental predictions. It may be, however, that other factors in field-experimental settings combine with prediction of results to produce biased behavioral recordings. An experimenter may, upon receiving data from individual observers, immediately examine the data and openly make evaluative comments, including surprise, approval, or disapproval regarding the degree to which the data fulfill the experimental predictions. Such feedback from an experimenter may

¹This research was supported by NIMH Grant MH21813 and Office of Education Grant OEG 0-71-2872-607. The authors are indebted to Joan Fisher for her statistical and editorial assistance, and also to Sharon Foster for her editorial assistance. Reprints may be obtained from K. Daniel O'Leary, Department of Psychology, State University of New York, Stony Brook, New York 11794.

²Now at the University of Texas, Austin, Texas.

affect the nature of subsequent recordings. The present investigation sought to determine whether observer recordings can be shaped by experimenter comments regarding the data.

METHOD

Design

Recordings of four categories of disruptive behavior were obtained from four observers who viewed four videotapes of each of two children during both baseline and treatment conditions. The observers were told that a decrease in the frequency of two target behaviors, playing and noise, was expected from baseline to treatment conditions as a function of a token intervention program. The observers were also informed that no change in the frequency of the two control categories, vocalization and orienting (turning around), was expected between baseline and treatment phases. Differential experimenter feedback to individual observers was provided during the treatment phase, contingent on the decreased frequencies of recorded behaviors predicted to decrease. A within-subject design allowed comparison of behavioral recordings of the two feedback and the two nonfeedback categories during baseline and treatment phases of the study.

Preparation of the Videotapes

Videotapes were made of two children with academic and behavioral problems in a laboratory classroom at Stony Brook. It was not apparent from the videotapes whether a treatment program was in effect. The two target children were positioned next to one another during filming and videotaped simultaneously at all times. However, the behavior of only one child at a time was recorded by observers during experimental sessions.

Three individuals who had over a year of experience with the behavioral code served as criterion-observers and recorded the presence of four categories of disruptive behavior: playing, noise, vocalization, and orienting. Intermittent

assessment yielded occurrence reliabilities consistently over 0.80 for these three observers. On the basis of these recordings, two sets of videotapes matched for average rate of occurrence of each of the four categories of behavior were selected to represent baseline and treatment conditions.

Observer Training

The four subject-observers were undergraduate freshmen (females) who were paid for participating in the study. The experimenter and observer-trainer was a sophomore psychology major (male) who had 1 yr of experience utilizing the behavioral code and training observers. The four categories of behavior were:

- (1) Vocalization: child emits a nonpermitted sound from his mouth;
- (2) Playing: child uses his hands to play with his own or community property;
- (3) Orienting: child turns more than 90 degrees from his desk while seated; and
- (4) Noise: child creates a nonpermitted sound other than vocalization.

The occurrence of disruptive behaviors was recorded on a 20-sec observe, 10-sec record basis (O'Leary, Kaufman, Kass, and Drabman, 1970). Reliability was computed by dividing number of agreements on the occurrence of a particular category of behavior during a 20-sec interval by the number of agreements plus disagreements. During the last two training sessions, reliability of observers with the experimenter averaged 0.74 (range 0.70 to 0.79).

Presentation of Research Hypotheses

After training, the present study was described to observers as a comparison of the levels of two target behaviors, playing and noise, that were the targets of a token reinforcement program, with two control behaviors, vocalization and orienting. The basic procedures and terminology of token programs were discussed and samples of the research literature were presented to substantiate the predictions that the frequencies of

playing and noise would decrease sharply during the treatment phase, and that the other two categories would remain unchanged. The purpose of the present study was explained as a replication of the effects of token reinforcement utilizing natural incentives, such as access to a playroom at the laboratory school (Osborne, 1969; O'Leary, Drabman, and Kass, 1973; Salzberg, Wheeler, Devine, and Hopkins, 1971). Observers were told that the videotape recordings of the children during baseline and treatment would be divided among them and viewed individually.

Experimental Sessions

During experimental sessions, the four observers viewed the videotapes separately. To minimize the likelihood of discovery by observers that all were, in fact, viewing the same videotapes, the four observers began the experimental sessions on consecutive days. During each session, each observer viewed a 12.5-min videotape twice, recording the behavior of one of the two target children during the first viewing and the behavior of the other during the replay. Before the second observation of the videotape during the fourth session of baseline and treatment, observers were informed that reliability would be computed with the experimenter's recordings, although reliability was in fact computed with the recordings of the criterion-observers for these respective videotapes. It is important to note that this procedure was different than during training, when reliability was computed with the actual recordings of the experimenter.

During the four baseline observation periods, the experimenter collected data from each subject-observer with no comment regarding any of the observations. During the treatment phase, the experimenter provided positive or negative feedback regarding the recordings of the two categories predicted to decrease. The average difference between the frequencies recorded by criterion-observers and subject-observers during baseline was employed as the criterion for

positive or negative comments by the experimenter on Day 1 of treatment. For example, if during baseline, Observer A recorded an average of two fewer occurrences of playing than the criterion-observer, she would have to exceed this difference (*i.e.*, record at least three fewer than the criterion-observer) on Day 1 of treatment to be commended for her recordings of that category. If not, she would receive a negative experimenter comment. In a similar manner, the average difference between frequency of each target category recorded by the subject-observer and criterion-observer on Day 1 of treatment was utilized as a criterion on Day 2 of treatment. It is important to emphasize that on each day of treatment, separate criteria for playing and noise were determined for each subject-observer. The actual formula utilized in determining criteria for positive *versus* negative feedback for each target category is presented below:

- (a) for Day 1 of treatment:

$$(\bar{S}_B - \bar{C}_B) - (S_{T_1} - C_{T_1}) = F$$

where

$\bar{S}_B$ = the average frequency of a target behavior recorded by a subject-observer during baseline;

$\bar{C}_B$ = the average frequency of a target behavior recorded by a criterion-observer for the videotapes viewed during baseline;

S_{T_1} = the frequency of a target behavior recorded by a subject-observer during Day 1 of treatment; and

C_{T_1} = the frequency of a target behavior recorded by a criterion-observer for the videotapes viewed during Day 1 of treatment;

if $F > 0$, positive feedback was given;
if $F \leq 0$, negative feedback was given;

- (b) for Days 2, 3, and 4 of treatment, the average frequency of a target behavior recorded by a subject-observer and a criterion-observer for the previous day's videotapes was substituted for $\bar{S}_B$ and $\bar{C}_B$, respectively, in the above formula.

Positive experimenter reactions included comments such as: "Those tokens are really having an effect on (category)"; "The reduction on (category) is a great example of how children modify only what is specifically rewarded"; and, "Dr. O'Leary will be pleased to see the drop in the level of (category)". Examples of negative reactions delivered by the experimenter included: "You don't seem to be picking up the treatment effect on (category)"; "It's strange that you have so many (category)—this treatment always works"; and, "I don't think a graph of this data will reflect the decrease in the number of (category)".

Following the final experimental session, all observers were given a questionnaire to assess the experimental manipulation.

RESULTS

During the four days of treatment, the observers reached the criteria for positive experimenter comments on an average of five of eight occasions for their recordings of playing (range 3 to 6) and on an average of 5.3 of eight occasions for their recordings of noise (range 5 to 6). The average frequency of each of the four behavioral categories recorded by subject-observers on each of the four days of baseline and treatment is presented in Figures 1 to 4. The maximum frequency for each behavior each session was 50. Figures 1 to 4 also present the frequency of behaviors recorded by criterion-observers for the same videotapes. Table 1 summarizes the average frequency recorded for each

Fig. 1. Frequency of playing recorded by subject-observers and criterion-observers during baseline and treatment. Subject-observers received contingent experimenter feedback during treatment.

Fig. 2. Frequency of noise recorded by subject-observers and criterion-observers during baseline and treatment. Subject-observers received contingent experimenter feedback during treatment.

category during baseline and treatment by criterion-observers and subject-observers. Both target categories were recorded less, relative to criterion observations, during treatment than during baseline. Playing was recorded 17% more by subject-observers than criterion-observers during baseline and 10% less during treatment. Noise

was recorded 6% more during baseline and 32% less during treatment. In contrast, control categories were recorded at approximately the same rate relative to criterion-observations during baseline and during treatment. Vocalization was recorded 33% less by subject-observers than criterion-observers during baseline and 34% less

Table 1
Frequency of the Observational Categories during Baseline and Treatment

Observational Categories	Baseline			Treatment		
	Average Frequency of Criterion-Observers	Average Frequency of Subject-Observers	% Difference*	Average Frequency of Criterion-Observers	Average Frequency of Subject-Observers	% Difference*
<i>Target Categories</i>						
Playing	14.9	17.4	+17%	15.4	13.9	-10%
Noise	4.8	5.1	+6%	5.0	3.4	-32%
<i>Control Categories</i>						
Vocalization	9.0	6.0	-33%	7.4	4.9	-34%
Orienting	4.9	5.5	+12%	4.8	5.7	+19%

*([frequency of subject-observer] - [frequency of criterion-observer]) ÷ frequency of criterion-observer.

Fig. 3. Frequency of vocalization recorded by subject-observers and criterion-observers during baseline and treatment. Subject-observers received no contingent experimenter feedback.

during treatment. Orienting was recorded 12% more during baseline and 19% more during treatment. The statistical significance of the differential reduction in frequency of playing and noise as a function of predicted results and experimenter feedback was substantiated by analysis-of-variance procedures.³

³The frequency of each category of behavior recorded by criterion observers was subtracted, for each videotape, from the frequency recorded by the respective subject-observers. These difference scores were analyzed within a factorial analysis-of-variance design (category of behavior within feedback/no feedback by observer by baseline/treatment by day). This analysis revealed a significant interaction of the feedback/no feedback with baseline/treatment ($F = 13.83$, $df = 1/3$, $p < 0.05$). A Newman-Keuls

Table 2 presents the average reliabilities obtained during training and the experimental conditions. Training reliabilities are based on three probes during the two days immediately before baseline. Experimental reliabilities are

a posteriori test revealed a significant decrease in the frequency of recording of feedback categories from baseline to treatment ($W = 2.24$, $p < 0.05$) and no significant change in the frequency of recording of nonfeedback categories. In addition, the interaction of the two categories within feedback/nonfeedback with baseline/treatment did not approach significance ($F = 1.79$, $df = 3/9$, $p > 0.20$), indicating that the overall differences between feedback and nonfeedback categories in their interaction with baseline *versus* treatment were consistent for playing and noise *versus* vocalization and orienting.

Fig. 4. Frequency of orienting recorded by subject-observers and criterion-observers during baseline and treatment. Subject-observers received no contingent experimenter feedback.

based on a single probe the last day of each condition. For noise and vocalization, reliabilities decreased from 0.75 and 0.51 during training to unacceptable levels of 0.45 and 0.00, respectively, during baseline. The reliabilities of these two auditory categories remained low (0.26 and 0.40, respectively) during the treatment phase. In contrast, reliabilities of playing and orienting remained consistently high during training and experimental conditions. It seems likely that

the utilization of auditory categories by the experimenter during training differed from that of the criterion-observers, whose ratings were employed in assessing reliability during the experimental conditions.

Questionnaire Responses

As a manipulation check, the observers were asked what sort of change from baseline the experimenter had anticipated in playing and

Table 2
Reliability of the Observational Categories during Training, Baseline, and Treatment

Observational Categories	Training Reliability	Baseline Reliability*	Treatment Reliability*
<i>Target Categories</i>			
Playing	0.75	0.83	0.83
Noise	0.75	0.45	0.26
<i>Control Categories</i>			
Vocalization	0.51	0.00**	0.40
Orienting	0.80	0.85	0.80

*One reliability probe was obtained on the last day of each phase.

**No occurrence recorded by subject-observers; criterion-observer recorded frequency of one on this tape.

noise during the treatment condition. All four observers responded correctly that a decrease had been predicted. In addition, all observers were asked if they felt that they had been misinformed about any aspects of the study. Of the four observers, one indicated that she felt she had been misinformed, and she stated that she suspected that her observations were being compared with those of the other observers who might be viewing the same videotapes. None of the observers indicated that they felt that the videotapes were not actual recordings of a token-economy experiment; nor did they feel that the experimenter was actively attempting to influence their ratings.

DISCUSSION

The present study demonstrates that a combination of observer knowledge of predicted effects and experimenter feedback can produce significant biases in observational data. In this study, experimenter response to observational data was based on criteria that were more systematic and explicit than in field-experimental studies, and they probably increased data bias. However, other aspects of the present investigation may well have lessened the impact of experimental factors. The treatment phase in the present study consisted of only four days; in a typical intervention program it would have lasted 10 to 15 days, allowing considerably more opportunity for shaping of recordings. Further, the limited number of observations and the use of recordings of videotape, rather than live children, eliminated the possibility of prolonged contact of observers with the experimenter, the teacher, and the children. Such contact might have produced a greater commitment and an *esprit de corps* among observers, resulting in greater motivation to produce "acceptable" data. In addition, the experimenter in the present study purposely did not comment on the level of control behaviors recorded, even on days when each of these behavioral categories showed clear drops relative to criterion ratings.

Had experimenter feedback been given for categories for which no change was predicted, as well as for categories predicted to decrease, a more dramatic contrast between experimental and control categories might have resulted. Finally, a greater potential for bias may exist when moderate changes in behavior occur as predicted, thus partially substantiating experimental hypotheses. In the present study, no overall change in the frequency of the four categories of behavior occurred from baseline to treatment.

Theoretically, biases in behavioral observations may occur regardless of level of reliability among observers. Substantial bias in the context of high reliabilities would simply indicate similar recording modifications by different observers. Intuitively, however, it seems that less bias would occur with very reliable recordings. Unfortunately, an investigator can rarely be certain of the actual reliability of his data, as reliability probes are likely to yield artificially inflated coefficients (Reid, 1970; Romanczyk, Kent, Diamant, and O'Leary, 1973). Practically then, the present investigation seems to indicate that biased recordings as a function of knowledge of experimental hypotheses combined with experimenter feedback may represent a potential problem for the majority of investigators. In the present study, widely differing reliabilities produced no significant differences in susceptibility to bias among the target categories or the control categories.⁴ Thus, even the investigator collecting highly reliable observational recordings may risk significant bias in the predicted direction by giving experimental expectations and differential feedback to observers.

It is apparent that circumstances conducive to shaping of observational recordings may occur in many applied settings. For example, after a frequently effective procedure is instituted for a child displaying severely maladaptive behavior, the hospital administrator might convey his

⁴($F = 1.28$; $df = 2/6$, ns).

enthusiasm with the treatment to observers. Especially if the treatment were at least partially effective, and the rates of maladaptive behavior were in fact decreasing, the hospital administrator's comments could shape and reinforce nonveridical recordings of behavior. Alternatively, in a typical university research setting, an undergraduate observer may watch subjects in a hospital token program, an off-campus day-care center, or an alcoholic treatment clinic, and return data to research assistants and graduate students. The latter individuals are generally committed to their investigations and through inadvertent comments such as: "It's great to see that kid improving", or "It looks like my idea was correct", they may shape recordings consistent with their hypotheses.

It is indeterminant from the present results whether knowledge of predicted results, experimenter feedback, or an interaction of the two factors accounted for the obtained bias. However, an earlier study of the effects of predicted results alone produced no significant biasing of observational data (Kent *et al.*, 1974). On the basis of these two studies, it is apparent that *evaluative* feedback to observers should be eliminated. While the expectation of change itself appears to have no significant effect on observational data obtained using well-defined behavioral categories, one might question any study in which experimenter feedback to the observers is not explicitly avoided. The most conservative stance would be to eliminate all review of incoming data that takes place in the presence of observers. Unfortunately, such review provides an educational and motivational function in research settings. It is possible that nonevaluative review of data—even when the experimental hypotheses are apparent—may not result in biased data. At a minimum, however, it is suggested that evaluative review of data must be eliminated if such data are not to be held suspect. While the present results may be idiosyncratic to the particular experimenter, or to the behavioral categories utilized, the potential

effects of a combination of knowledge of predicted results and feedback from an experimenter should be seriously considered when planning and evaluating outcome research.

REFERENCES

- Cordaro, L. and Ison, J. R. Observer bias in the classical conditioning of the planaria. *Psychological Reports*, 1963, 13, 787-789.
- Kent, R. N., O'Leary, K. D., Diamant, C., and Dietz, A. Expectation biases in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology*, 1972, 42, 774-780.
- O'Leary, K. D., Drabman, R. S., and Kass, R. E. Maintenance of appropriate behavior in a token program. *Journal of Abnormal Child Psychology*, 1973, 1, 127-138.
- O'Leary, K. D., Kaufman, K. F., Kass, R. E., and Drabman, R. The effects of loud and soft reprimands on the behavior of disruptive students. *Exceptional Children*, 1970, 37, 145-155.
- Osborne, J. G. Free time as a reinforcer in the management of classroom behavior. *Journal of Applied Behavior Analysis*, 1969, 2, 113-118.
- Reid, J. B. Reliability assessment of observation data: a possible methodological problem. *Child Development*, 1970, 41, 1143-1150.
- Romanczyk, R. G., Kent, R. N., Diamant, C., and O'Leary, K. D. Measuring the reliability of observational data: a reactive process. *Journal of Applied Behavior Analysis*, 1973, 6, 175-184.
- Rosenthal, R. *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts, 1966.
- Rosenthal, R. Interpersonal expectations: effects of the experimenter's hypothesis. In R. Rosenthal and R. L. Rosnow (Eds.), *Artifact in behavioral research*. New York: Academic Press, 1969. Pp. 182-279.
- Rosenthal, R., Friedman, C. J., Johnson, C. A., Fode, K. L., Shill, T. R., White, C. R., and Vikan-Kline, L. L. Variables affecting experimenter bias in a group situation. *Genetic Psychological Monographs*, 1964, 70, 271-296.
- Salzberg, B. H., Wheeler, A. A., Devar, L. T., and Hopkins, B. L. The effect of intermittent feedback and intermittent contingent access to play on printing of kindergarten children. *Journal of Applied Behavior Analysis*, 1971, 4, 163-172.
- Scott, P., Burton, R. V., and Yarrow, M. Social reinforcement under natural conditions. *Child Development*, 1967, 38, 53-63.

Received 10 August 1973.

(Final acceptance 4 November 1974.)