JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1973, 6, 175-184 NUMBER 1 (SPRING 1973)

MEASURING THE RELIABILITY OF OBSERVATIONAL DATA: A REACTIVE PROCESS[1]

RAYMOND G. ROMANCZYK[2], RONALD N. KENT, CHARLES DIAMENT, AND K. DANIEL O'LEARY[3]

STATE UNIVERSITY OF NEW YORK AT STONY BROOK

Reliability of observational data was measured simultaneously by two assessors under two experimental conditions. During overt assessment, observers were told that reliability would be measured by one of the two assessors, thus permitting computation of reliability with an identified and an unidentified assessor. During covert assessment, observers were not informed of the reliability measured. Throughout the study, each of the assessors employed a unique version of a standard observational code. In the overt assessment condition, reliability of observers with the identified assessor was consistently higher than reliability with the unidentified assessor, indicating that observers modified their observational criteria to approximate those of the identified assessor. In the covert assessment condition, reliability with the two assessors was substantially lower than during overt assessment. Further, observers consistently recorded lower frequencies of disruptive behavior than the two assessors during covert assessment.

The systematic collection of objective records of human behavior in natural settings is an integral part of behavior modification research. Unfortunately, mechanized observational techniques have not been developed that are adequate to monitor complex social behavior seen in a classroom or on a hospital ward. In the absence of mechanical recording devices, investigators have formulated operational definitions of the specific behaviors under investigation, and have trained observers, typically college undergraduate or graduate students, to observe and record behavior.

The training of observers is a relatively straightforward procedure. For example, if the behavior to be observed is "out-of-seat", the experimenter first operationally defines what constitutes "out-of-seat" behavior. His observers then learn that definition and record whether or not "out-of-seat" occurs during a specific time interval, *e.g.,* during each minute. Pairs of observers will also occasionally compute a reliability coefficient to determine their level of agreement with each other. The level of reliability obtained may vary greatly depending upon the type of behavior observed, and the method of reliability calculation. Generally, however, average reliabilities of classroom behaviors are reported above 0.75 (O'Leary and Becker, 1967; Hall, Lund, and Jackson, 1968; Barrish, Saunders, and Wolf, 1969). Demonstration of high reliability is critical for clear interpretation of experimental data. To the degree that different observers record different behaviors, it is uncertain what behavior is "actually" being emitted by an experimental subject. Depending upon the specific experimental design employed, low reliability of behavioral recordings may result in behavioral measures that include a high proportion of error and, therefore, are insensitive to the effects of an experimental manipulation. More seriously, behavioral recordings of low reliability may result

[1]Portions of this paper were read at the Second Annual Symposium on Behavior Analysis, Lawrence, Kansas, May 11, 1971. This study was supported in part by Regional Office of Education Grant OEG-2710017 and in part by a NSF Institutional Grant from President Toll, SUNY, Stony Brook. The authors wish to thank David Santogrossi for his editorial comments and his suggestions.

[2]Now at the Psychological Clinic, Rutgers—State University, New Brunswick, New Jersey.

[3]Reprints may be obtained from Ronald N. Kent or K. Daniel O'Leary, Psychology Department, SUNY, Stony Brook, New York 11790.

in data that are confounded with and biased by the idiosyncratic observational criteria employed by each observer. This would be the case, for example, if two observers who consistently utilized different observational criteria, were assigned to observe two different experimental conditions.

It is generally assumed that the process of reliability assessment is non-reactive; that obtained reliability coefficients reflect the general adequacy of the observational process at times when reliability is not being measured. This would be the case if observer ratings were unaffected by (a) knowledge that reliability is being assessed, and (b) knowledge of which observer is assessing reliability. Empirical support for these assumptions is particularly critical, since neither of these factors is routinely concealed from observers.

Reid (1970) reported that observers obtaining median reliabilities of 75% when they were aware that reliability was being assessed, dropped to a median reliability of 51% when they were told that reliability would not be assessed. Unfortunately, Reid employed observers trained for a median of only 6 hr to apply a 33 item behavioral code to 5-min segments of video tapes of parent-child interactions. After computation of reliability measures for each observational sequence, observers were instructed that no further reliability measures would be taken. As Reid noted, these circumstances represent only a laboratory analogue of naturalistic observation, and other more typical strategies for the collection and assessment of observational data may be immune to such effects.

From Reid's data, it is impossible to evaluate the effects of telling the observer which person is assessing his reliability. However, anecdotal evidence from observers in studies at Stony Brook casts doubt on the assumption that ratings are unaffected by knowledge of who is assessing reliability. Their comments indicate that observers may adjust their rating criteria to match those of any particular reliability assessor. In order to test the two assumptions, the design employed in the present study permitted assessment of the reliability of each observer with each of two assessors under two different conditions:

1. *Overt assessment.* Observers were told *either* that Assessor I was performing the reliability assessment *or,* that Assessor II was performing the reliability assessment, although, in fact, both assessors were performing the assessments simultaneously. This permitted computation of two reliability coefficients: (a) reliability with an *identified assessor,* (b) reliability with an *unidentified assessor.*
2. *Covert assessment.* Observers were not told that a reliability assessment was being performed, although both assessors were performing the assessments simultaneously.

In order to allow a clear evaluation of the extent to which observers' recordings are influenced by knowledge of the identity of the particular assessor performing the reliability assessment, four of the nine behavioral categories were modified such that different observational criteria were established for the two assessors. This strategy assumes that such differences in observational criteria employed by assessors may, in fact, exist in field experimental research and allows a measure of the degree to which observers' recordings are affected by such differences.

METHOD

The present experiment was conducted at a university laboratory school at Stony Brook. Five undergraduate observers with at least three months' experience in recording the disruptive behavior of children served as subjects. Three of these observers were employed during a morning period, and the other two during an afternoon period. Two other undergraduates

with 2 yr of experience in observing served as reliability assessors during both morning and afternoon. Observers in the present study were engaged in recording the level of disruptive behavior of eight first-grade children in an experimental classroom. All observations were made from a room adjoining the classroom which was equipped with an observation mirror and an audio-amplification system.

Standard observation procedures employed throughout this study included the synchronization of stopwatches of all observers with assessors before each assigned observation period began, and the recording of the behavior of any target child observed for a 12.5-min period. Observations of disruptive behavior, as defined by a nine-category behavioral code (O'Leary, Romanczyk, Kass, Dietz, and Santogrossi, *unpublished*) were made on a 20-sec observe, 10-sec record basis, *i.e.,* the observer would watch the child for 20 sec, then take 10 sec to record the disruptive behavior that had occurred during that 20-sec period. A category of behavior was recorded as occurring if a behavior as defined by that category was observed one or more times during a particular 20-sec interval. A brief description of the behavioral code is presented in Table 1.

The measure of reliability employed for each category was number of agreements in coding the occurrence of behavior, interval by interval, divided by the number of agreements plus disagreements. In addition, a measure of total reliability for modified and unmodified categories was obtained by dividing the total number of agreements in recording behavior by the total number of agreements plus disagreements. Reliability was computed on the basis of 12.5-min observation periods. Average reliability for a particular experimental condition on a particular day was obtained by computing the arithmetic mean of reliabilities for 12.5-min periods.

For the purpose of the present study, four of the nine categories of the behavioral rating code were modified to produce stable but differential observational criteria for the two assessors. This

Table 1

Brief Summary of Disruptive Behavior Categories

Title	*Description*
1. Out of chair	Observable movement of the child from his chair when not permitted or requested by teacher. None of the child's weight is to be supported by the chair.
2. Modified out of chair	Movement of child from his chair, with some aspect of the body still touching the chair.
3. Touching others' property	Child comes into contact with another's property without permission to do so.
4. Vocalization	Any non-permitted "audible" behavior emanating from the mouth.
5. Playing	Child uses his hands to play with his own or community property, so that such behavior is incompatible (or would be incompatible) with learning. Also, reading non-task related material.
6. Orienting response	Child is seated and turns more than 90 degrees using the desk as a reference point.
7. Noise	Child creates any audible noise without permission, other than vocalization.
8. Aggression	Child makes an intense movement directed at another person so as to come into contact with him, either directly or by using a material object as an extension of the hand.
9. Time-off-task	Child does not do assigned work for entire 20-sec interval.
10. Absence	No inappropriate behavior as defined by the above categories.

manipulation was intended to increase the detectability of matching by the observers of the different observational criteria employed by each assessor. As a result of these modifications, the code employed by Assessor I produced a higher frequency than the code employed by Assessor II on two categories: vocalization and noise. In employing the modified code, Assessor I would record even the softest vocalizations and also any "mouthings" the child might make as vocalizations, while Assessor II would record only the louder vocalizations and ignore such behavior as humming, whispering, and sighing. Further, the behavioral code was modified so that Assessor II would record a greater frequency than Assessor I on two other behaviors: playing and orienting. It was required that these differential observational criteria be sufficiently well defined that the assessors would be reliable with each other at a moderate level and that this level of reliability between assessors not vary across experimental conditions. In short, an artificial difference was created between the ratings of Assessor I and Assessor II.

For two and one half weeks before the experiment, each assessor employed his respective version of the modified code and on regular but different occasions, computed total reliability (for modified and unmodified categories combined) with each observer. Reliability was computed for the five observers a median of four times (range 2 to 4). These reliability computations provided the only opportunity for observers to note the unique observational criteria being employed by the two assessors. At no time, however, did either reliability assessor make any statement that overtly contrasted his rating criteria with those of the other assessor.

During the experimental conditions, both assessors measured the reliability of particular observers simultaneously for both the overt and covert assessment conditions. Data were collected over a three-week period on eight separate days, and the results are based on a total of 86, 12.5-min observational recordings by the five observers.

The presence of one or both assessors not identified as assessing reliability during a particular period was not questioned because the assessors were understood to be engaged in observational tasks other than reliability assessment, such as developing specialized codes for certain target children. Additionally, the assessors positioned themselves so as to preclude the possibility of an observer viewing an assessor's recording sheet or the assessor viewing observer's sheet. The routine synchronization of stopwatches and prearranged scheduling of observation facilitated collection of covert reliability measures.

During the experimental conditions, the eight children in the class were observed in random order. Throughout the study, absences of both children and observers, as well as scheduling difficulties, precluded the possibility of equal recording of each child by each observer in each experimental condition. In all, the five observers were absent, respectively, during 3, 3, 2, 2, and 1 of the eight experimental sessions. On Day 2, three observers were absent, on Days 1, 3, and 5, two observers were absent, and on Days 6 and 8, one observer was absent.

Both experimental conditions (overt assessment and covert assessment) were implemented during each of the eight days of the study. The order of overt and covert experimental conditions was systematically varied across days, to eliminate a sequence effect. On all but one of the eight days of the study, each of the assessors served as the identified and unidentified assessor during the overt assessment condition. The order in which each assessor was identified was varied across days.

Each experimental day included a morning and an afternoon observation session. Three observers were employed in the morning and two in the afternoon. The two assessors were present during both sessions. During overt reliability assessment, all observers recorded the behavior of the same target child simultaneously with the identified and unidentified assessors. After an overt assessment, each observer computed reli-

ability with the identified assessor. This arrangement provided an efficient method for generating extensive reliability data within the restrictions imposed by the field-experimental setting. During covert reliability assessment, however, only one observer recorded the behavior of a particular target child simultaneously with the two assessors. Covert reliability of each assessor with a particular observer was computed surreptitiously at the end of each experimental day. During each day, an observer typically made four or five consecutive 12.5-min observations of different children, with interruptions for computing reliability as required. As was the case during training, reliability computation included a discussion of disagreements between the observers and the identified assessor. At no time did the assessor make any statement that overtly contrasted his rating criteria with those of the other assessor.

RESULTS

The total reliability between the two assessors averaged over all experimental conditions was 0.52. A one-way analysis of variance revealed no significant differences in reliabilities between the two assessors across the three experimental conditions ($F = 3.19$, $df = 1/5$; $p > 0.05$). As planned, this indicates that a moderate, but consistent level of reliability existed between the two assessors across the experimental conditions.

Knowledge of Assessment

Figure 1 presents a comparison of the average total reliabilities on modified categories obtained by the five observers across the eight days of the study, during covert assessment and during overt assessment by an identified assessor. Figure 2 presents this same comparison for average total reliabilities on unmodified categories. As can be seen, reliability during covert assessment was consistently lower than reliability during overt assessment by an identified assessor on both modified and unmodified behavioral categories.

Figure 3 presents a comparison of average total frequency of disruptive behavior recorded by the five observers during overt and covert reliability assessment. On seven of the eight days of the study, observers recorded a lower frequency of disruptive behavior when they

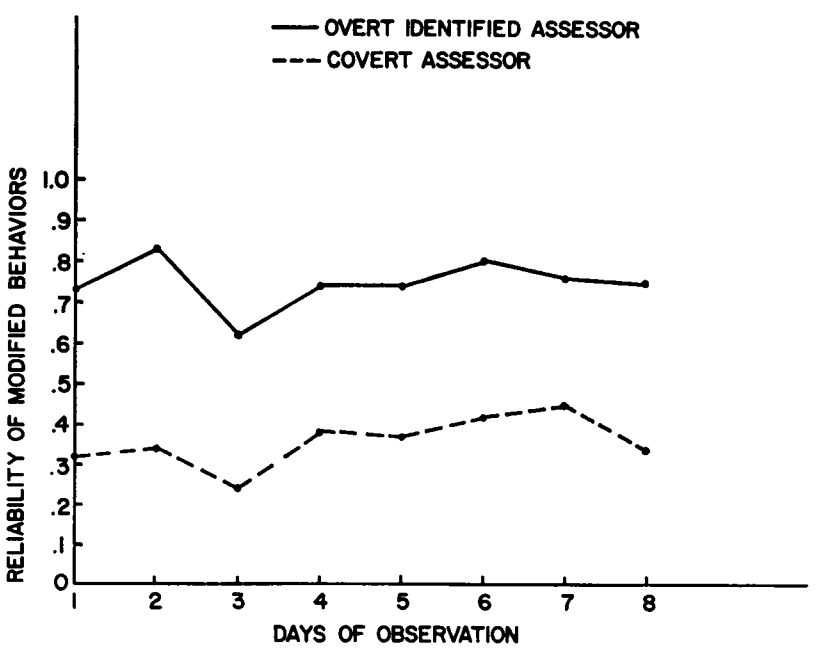

Fig. 1. Average reliability of observers on modified categories during overt and covert assessment conditions.

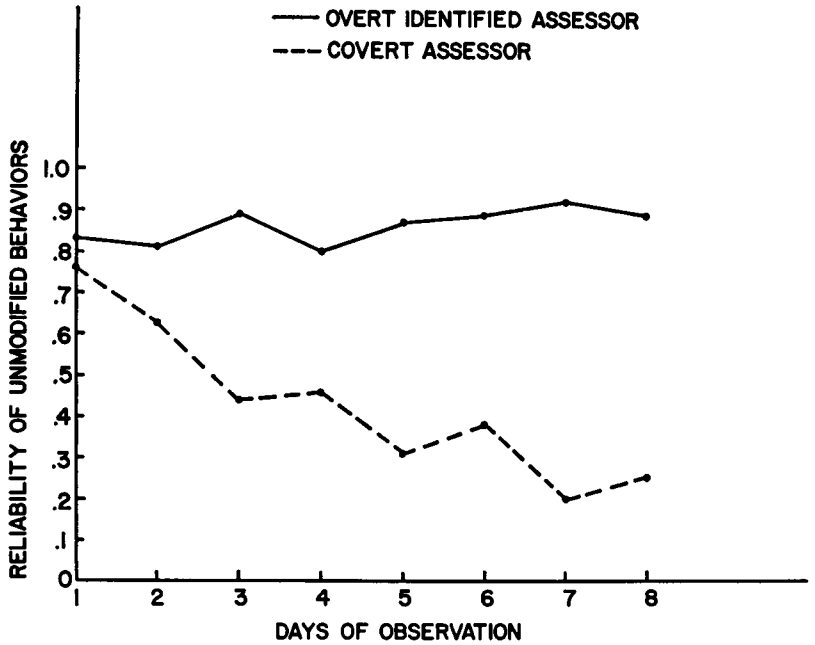

Fig. 2. Average reliability of observers on unmodified categories during overt and covert assessment conditions.

were not instructed that reliability was being assessed. In fact, over the eight days of the study, the level of disruptive behavior recorded during covert assessment was 75% of the level recorded during overt assessment. Thus, both the frequencies and reliabilities of behavior recorded were consistently increased by the knowledge that reliability was being assessed.

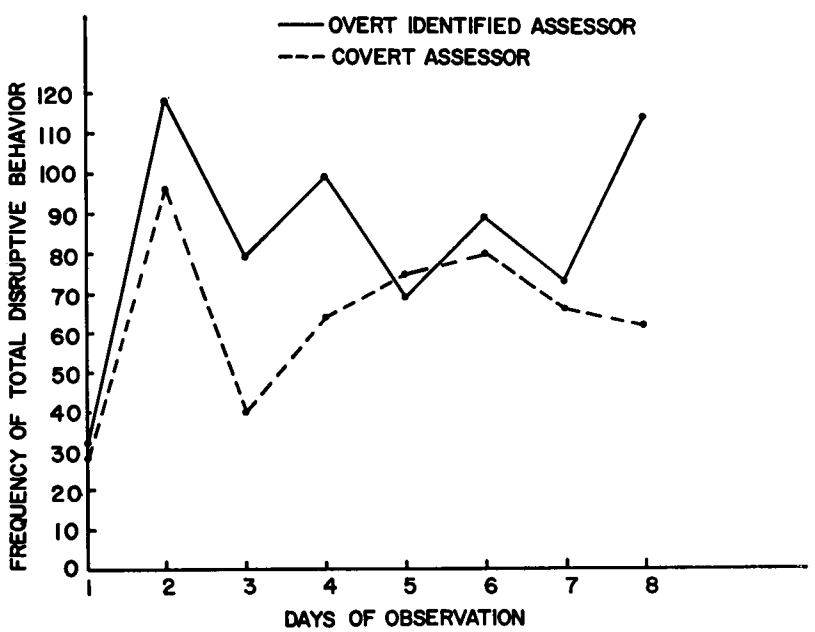

Fig. 3. Average frequency of disruptive behavior recorded during overt and covert assessment conditions.

Knowledge of Assessor

Figure 4 presents a comparison of the average total reliabilities on modified categories during overt assessment by an identified and an unidentified assessor. As predicted, reliabilities with the identified assessor are consistently higher than with an unidentified assessor. This indicates that the observers shifted their observational criteria to match the idiosyncratic criteria employed, respectively, by the two assessors.

Figure 5 presents the same comparison of reliability with identified and unidentified assessors during overt assessment for unmodified categories. This figure demonstrates an unexpected similar, but less substantial, tendency for observers to produce higher reliabilities with an identified assessor on the unmodified categories.

DISCUSSION

The present study indicated that reliability measures were consistently and substantially inflated by knowledge that reliability was being assessed and by knowledge of which assessor was performing the assessment. Further, the frequency of behavior recorded was 25% lower when observers were not instructed that reliability was being assessed, thereby systematically biasing the data generated toward underestimates of disruptive behavior.

Knowledge of which assessor was measuring reliability produced a substantial shift in observational criteria on the modified rating categories and to a lesser extent, on unmodified categories. A shift of observers on unmodified categories was not predicted. High reliability between the assessors for the unmodified categories would have precluded the possibility of such a shift. In fact, average total reliability between the two assessors on the unmodified categories was only 0.72. Thus, natural (noninduced) differences in the observational criteria of the two assessors were sufficient to allow a shift in ratings of observers on unmodified categories.

Several important implications for investigations utilizing observational procedures emerge from the present data. In such studies, reliability assessment serves two separate func-

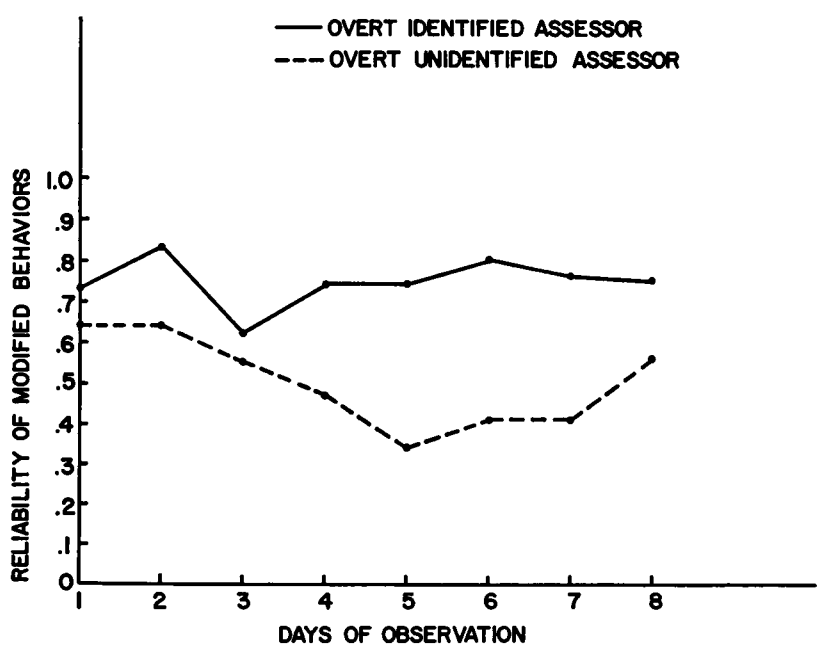

Fig. 4. Average reliability of observers on modified categories with an identified and unidentified assessor.

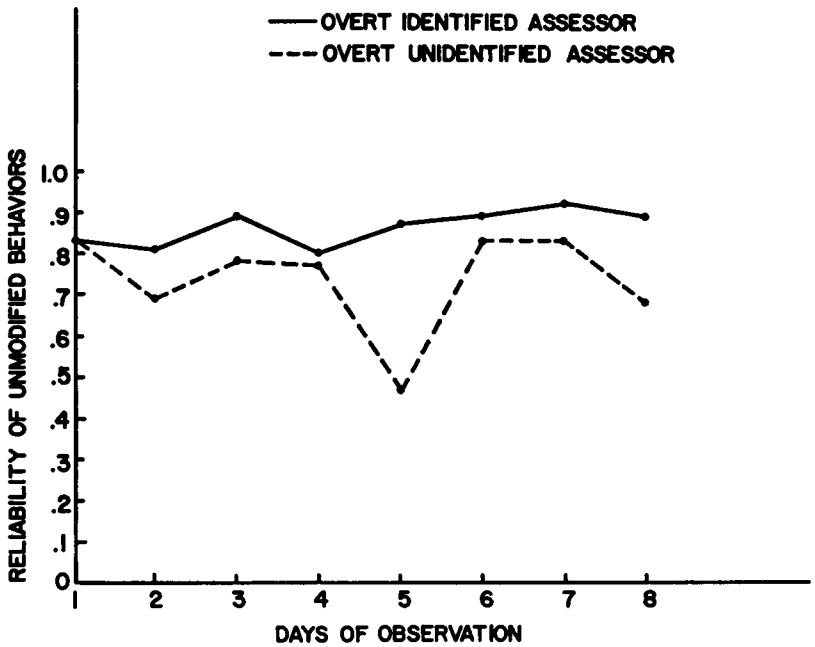

Fig. 5. Average reliability of observers on unmodified categories with an identified and unidentified assessor.

tions. The first of these functions is to provide feedback to observers regarding the accuracy of their recordings. During training of observers the level of reliability serves to provide feedback regarding their progress. These reliabilities are obtained with standard ratings of videotapes, the experimenter, or each other, in conjunction with discussion of the specific disagreements in recording. Similarly, computation of reliability and discussion of disagreements may continue to provide feedback, which modifies the behavioral definitions employed throughout the collection of data. The present study indicated that observers may, in fact, adjust their rating criteria as a function of the feedback they receive. Similarly, O'Leary and Kent (*in press*) have presented data indicating that when observers were divided into separate groups and restricted to computation of reliability with members of their own group, they soon began to "drift" in their application of the behavioral code. The result of this was the development, by different groups of observers, of idiosyncratic definitions of the behaviors to be recorded.

The implication of these findings is that reliability assessment for the purpose of providing feedback to observers, during training or during data collection, must be based on criteria that remain constant. Perhaps the best way of accomplishing this would be to prepare a standard set of videotapes of the behavior of interest and to obtain ratings by the experimenter or experienced observers. These ratings could be employed as the operational definition of the behavioral code and employed as the only source of feedback for observers during training and intermittently during data collection. A second possibility would be for a single individual to provide all feedback to observers. Unfortunately, changes in the recording of this individual, during the different phases of a particular study or between studies, would remain a possibility.

On the basis of the present data it would be predicted that failure to observe these precautions may produce increases in error variance in observational data, as observers adjust their rating criteria to resemble those of a variety of assessors. A more serious consequence is the possibility that observers, assigned to different experimental conditions (*i.e.*, in a between-subjects design), may receive differential feedback from different reliability assessors, thus

confounding the experimental manipulation with the rating criteria employed.

A second function of reliability assessment is to provide, for the experimenter, an estimate of the consistency with which measures are being obtained. High reliability indicates a low proportion of variance in the data due to measurement error, as well as an increased likelihood that experimental data could be replicated by observers other than those employed. The present study, as well as an earlier study by Reid (1969), indicates that observers may record behavior more reliably when they have been informed that reliability is being assessed than at other times. The implication of this finding is that reliability assessment for the purpose of evaluating the consistency of data must be accomplished without the knowledge of the observers. One way of accomplishing this would be to monitor observers via an assessor who is present throughout the study and intermittently performs covert checks of reliability. Alternately, recordings of observers could be compared intermittently with those obtained via a closed circuit television camera. Kent, Diament, Dietz, and O'Leary (*in preparation*) have indicated that observational recordings obtained via closed circuit television are comparable in frequency and reliability with ratings obtained *in vivo*. Under this circumstance, the recordings of observers could be evaluated at any time, without their knowledge. A more practical arrangement would involve two or three observers who would be working together in a particular setting. Each member of this group would be given a schedule specifying the person to be observed during a particular time period, and these schedules would arrange for occasional simultaneous observation of the same person by two observers. When recordings were returned, the experimenter could determine the level of agreement of these simultaneous observations. The potential difficulty with this less costly arrangement is that observers, if motivated to do so, could easily determine from one another which observation intervals would allow reliability assessment and respond as they would to an overt reliability assessment.

There exists the clear possibility that the present results may not generalize to other circumstances. However, in view of the present data, the potentially reactive nature of reliability assessment of observational recordings should be studied utilizing other behavioral codes in a variety of natural settings. The present report represents one of a very few studies that even attempts to define the problems involved in observing behavior. The solutions to such difficulties remain, to date, completely a matter of speculation.

In view of the difficulties associated with observational recordings, one might well consider abandoning measurement of behavior requiring a judgment. In fact, Winett and Winkler (1972) suggested the more critical importance in a classroom setting of one potential alternative, product measurement. However, as O'Leary and Kent (*in press*) have noted, "even in the case of product measures, data regarding social behavior often provide a measure of the degree to which important stimulus factors have remained constant across experimental conditions. For example, when measuring the number of products correct before and during a token program, it is critical to measure the degree to which the teacher instructs, as well as prompts and reinforces problem completion during baseline and treatment conditions. In the absence of such data, it is impossible to conclude that the reward contingency is the critical factor in producing change.

"Use of product measures does not eliminate judgmental factors which seem so problematic in observational recording of behavior. It seems likely that evaluations of handwriting or short answers in the classroom may also suffer from lack of consistent judgment. In other settings of interest to child behavior modifiers, such as the home or the playground, product measures are simply not available. In fact, it seems there is an entire realm of social behaviors, such as cooperation among children, creativity, and fol-

lowing instructions, which are of direct interest and for which there are no tangible products." Measures of social behavior will continue as a dependent variable of major importance, not only for clinical psychologists, but also in the areas of social and developmental psychology. It is necessary to develop our measurement and research technology sufficiently to allow us to interpret such measures unambiguously.

REFERENCES

Ayllon, T., Layman, D., and Burke, S. *The control of disruptive behavior through reinforcement of academic objectives.* Unpublished paper presented at 1971 Symposium, Behavior Analysis in Education, Lawrence, Kansas.

Barrish, H. H., Saunders, M., and Wolf, M. M. Good behavior game: effects of individual contingencies for group consequences on disruptive behavior in a classroom. *Journal of Applied Behavior Analysis,* 1969, **2,** 119-124.

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

Kent, R. N., Diament, C., Dietz, A., and O'Leary, K. D. *Observational recordings of child behavior obtained via observation mirror, closed circuit television, and in vivo.* In preparation.

O'Leary, K. D. Behavior modification in the classroom: a rejoinder to Winett and Winkler. *Journal of Applied Behavior Analysis,* 1972, **5,** 505-511.

O'Leary, K. D. and Becker, W. C. Behavior modification of an adjustment class: a token reinforcement program. *Exceptional Children,* 1967, **33,** 637-642.

O'Leary, K. D. and Kent, R. N. Behavior modification for social action: research tactics and problems. In L. A. Hamerlynk, P. O. Davidson, and L. E. Acker (Eds.), *Critical issues in research and practice.* Champaign, Illinois. Research Press, 1973.

O'Leary, K. D., Romanczyk, R. G., Kass, R. E., Dietz, A., and Santogrossi, D. *Procedures for classroom observation of teachers and children.* Unpublished manuscript, SUNY, Stony Brook, 1969.

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

Winett, R. A. and Winkler, R. C. Current behavior modification in the classroom: Be still, be quiet, be docile. *Journal of Applied Behavior Analysis,* 1972, **5,** 499-504.

Received 20 August 1971.
(Revision requested 19 May 1972.)
(Final acceptance 6 September 1972.)