JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1972, 5, 19-30 NUMBER 1 (SPRING 1972)

ACHIEVEMENT PLACE: THE RELIABILITY OF SELF-REPORTING AND PEER-REPORTING AND THEIR EFFECTS ON BEHAVIOR[1,2,3]

DEAN L. FIXSEN, ELERY L. PHILLIPS, AND MONTROSE M. WOLF

UNIVERSITY OF KANSAS

The reliability of the boys reporting their own behavior and the behavior of their peers was measured in two experiments at Achievement Place, a community based, family style, behavior modification program for delinquents based on a token (point) economy. The results of these experiments indicated that; (a) the boys were not "naturally" reliable observers, (b) the reliability of peer-reporting could be improved by providing training on the behavioral definitions and by making points contingent on agreement between each boy's peer-report and an independent adult observers' report, (c) the reliability of self-reporting could be improved by making points contingent on agreement between the self-report and the trained peer's report, and (d) giving self-reports and peer-reports did not produce a systematic effect on the boys' room-cleaning behavior as measured by an independent observer.

In most current applied research, the behavior of interest is quantified by using human observers, and the reliability of this recording technique is measured by having a second observer simultaneously but independently record the same behavior (Baer, Wolf, and Risley, 1968). Usually, the observers are not the subjects in the experiment. In some cases, however, it is impractical to have observers continually present in the experimental situation. In such cases, it may be feasible to have the subject record his own behavior and have a peer or other observer record the same behavior at times to assess reliability. Azrin and Powell (1969) used this measurement procedure by requiring adult subjects to record the time a given behavior occurred. Reliability was measured by having a "participant observer" (a fellow employee named by the subject) observe the subject for two minutes at specified periods each day. They found that the self-reports from the subjects and the peer-reports from the observers agreed 98% of the time. Surratt, Ulrich, and Hawkins (1969) found that a peer observer (a fifth grader) averaged 95% agreement with the classroom teacher and the experimenters in recording the time four first graders spent studying. With preschool children, Risley and Hart (1968, Exp. 1) found a low degree of correspondence between subjects' verbal reports of their behavior and the observer's reports during baseline conditions, but they demonstrated that it could be trained. When a snack was made contingent on the subjects' verbal reports, there was an immediate increase in the number of verbal reports but only a small and unreliable effect upon the actual behavior (*i.e.*, correspondence between self-report and observer-report decreased). Snacks were then made contingent upon correspondence between the

[1]The authors wish to thank Gary Timbers who served as the primary observer in both experiments.

[2]This investigation was supported by grant MH 16609 from the National Institutes of Mental Health (Center for Studies of Crime and Delinquency) and by grant HD 03144 from the National Institute of Child Health and Human Development to the Bureau of Child Research and the Department of Human Development, University of Kansas. Preparation of the manuscript was partially supported by PHS Training Grant HD 00183 from the National Institute of Child Health and Human Development to the Kansas Center for Research in Mental Retardation, University of Kansas.

[3]Reprints may be obtained from the authors, Bureau of Child Research, University of Kansas, Lawrence, Kansas 66044.

self-report and observer-report, and a greater degree of correspondence ensued.

Broden, Hall, and Mitts (1972) reported two experiments in which the effects of self-recording on classroom behavior was measured. In the first experiment, a junior high school girl had a low percentage of study behavior during baseline. When the self-recording procedures were instituted, the percentage of study behavior doubled. This effect was replicated in the next baseline and self-recording phases. When the observer's daily records of study behavior were compared to the student's daily records, there was a low degree of correspondence between the two. However, the mean percentages for each condition showed a high degree of agreement between the observer and the student.

In the second experiment Broden, *et al.*, (1972) found a high frequency of talk-outs for a fifth-grade boy during baseline. Using a combination of multiple baseline and reversal experimental designs they reported that the number of talk-outs was substantially decreased initially when the student recorded his own behavior. By the third reversal, however, self-recording had no effect on talk-outs. In the second experiment there was no assessment of the reliability of the student's self-recording.

In summary, the results of these studies indicate that agreement between self-reports and observer-reports is slight (Broden, *et al.*, 1972; Risley and Hart, 1968), agreement between self-reports and peer-reports is high (Azrin and Powell, 1969), agreement between peer-reports and observer-reports is high (Surratt, *et al.*, 1969), and that self-reports have an effect on the behavior being reported (Broden, *et al.*, 1972). These conclusions should be considered tenuous because the subjects, behaviors, and procedures varied considerably among the studies reviewed. However, it appears that in some cases, untrained subjects "naturally" give reliable reports of the behavior of their peers and that self-reports have an effect on the subject's behavior.

The purpose of the first experiment was to investigate the reliability of self-reporting and peer-reporting and to assess their effects on the behavior of pre-delinquent boys in a group home.

METHOD

Subjects and Setting

Seven "pre-delinquent" boys who had been committed by the Juvenile Court to Achievement Place, a family style rehabilitation program in Lawrence, Kansas (Phillips, 1968; Phillips, Phillips, Fixsen, and Wolf, 1971), averaged 14.0 yr in age and ranged from 12 to 16 yr. At the end of both experiments, which required six months to complete, the boys' length of stay at Achievement Place averaged 14.7 months and ranged from 4 to 30 months. Five of the seven boys were committed to Achievement Place after they were adjudicated for offenses that would have been felonies if the boys were adults. The other two boys had been declared dependent and neglected.

On a typical school day at Achievement Place, the boys were awakened at 7:00 a.m. then they showered, dressed, and cleaned their bedrooms and bathrooms before breakfast. After breakfast they attended to clean-up chores in the kitchen, reviewed their homework, or watched TV before leaving for the public schools. After school the boys returned to Achievement Place where they usually had a snack (if they had earned that privilege) before they began any of their afternoon activities.

The treatment program at Achievement Place is based on a token economy (point system). The boys earn points (tokens) for engaging in appropriate social, academic, and self-care behaviors and lose points for inappropriate behaviors. At specified times (each day for the boys who have recently entered the program, each week for the other boys) the points can be used to purchase a variety of privileges in the home. The privileges include snacks, TV, allowance, and permission to leave Achievement Place as

well as others. When purchased on a weekly basis, the cost of each privilege varies from 1000 to 3000 points. Each boy must earn approximately 10,000 points each week to buy enough privileges to "live comfortably".

EXPERIMENT I

Reliability of Reporting

Procedures

Subjects. The subjects were six boys from Achievement Place.

Observation. Before the experiment, definitions specifying the degree of cleanliness or placement of objects were composed for 21 areas of each boy's room. These definitions are given in Table 1. Using the 21 room-cleaning definitions, an adult observer recorded whether or not each boy's room met each of the criteria. Room cleaning was recorded every school day while the boys were at school. The boys were not told at any time that their rooms were being checked.

At least once during each condition, another adult observer independently recorded room cleaning. The two adult observers' records were then compared and scored, item by item, for the number of agreements across all the boys' rooms. The total number of agreements $\times$ 100 was divided by the total number of agreements plus disagreements to obtain the per cent of inter-observer agreement. For the seven reliability checks in Experiment I, inter-observer agreement ranged from 83% to 94% with a mean of 89%.

Conditions

Baseline. There were no scheduled consequences for room cleaning.

Self-Report. Each boy was given the list of 21 room-cleaning definitions and a check sheet with spaces numbered 1 through 21. The boys were simply asked to check their rooms each morning before they left for school. They were told to put a check mark by the number on the check sheet if that definition had been met or a zero if the definition had not been met. They were also instructed to pick up a new check sheet each morning, check their rooms, and put the completed check sheet on the teaching-parents' desk before leaving for school. If the teaching-parents noted a missing check sheet, the boy was reminded to fill one out before he left for school. Failure to fill out a check sheet after being reminded resulted in a 100-point fine, which was levied after the boy returned from school. There were no scheduled consequences for room cleaning or for accurate self-reporting.

Peer-Report. Each boy was told to complete a check sheet for a "buddy" that was assigned to him. In all other respects this condition was identical to the self-report condition.

Results

Figure 1 shows the average performance of all six boys under the experimental conditions. As shown in Figure 1, under baseline conditions the cleanliness of the boys' rooms averaged seven or eight items correct out of the 21 items possible. When the self-report condition was initiated, the measured behavior (the adult observer's report) increased to 11 to 13 items correct over the first six days then declined to about 10 items correct over the last few days. The boys' self-reports, however, averaged 19 items correct across all 13 days of this condition. When baseline conditions were reinstated, the measured behavior averaged about 10 items correct, slightly higher than the original baseline. After reinstating the self-report condition there was no change in the measured behavior while the self-report of the behavior again averaged about 19 items correct. Thus, the disparity between the measured behavior and the self-reports of that behavior was replicated while the initial increase in the measured behavior found under the first self-report condition was not replicated. When peer-report conditions were instituted, the measured behavior averaged about nine items correct while the peer-reports of that behavior averaged about 19 items correct out of 21 possible. (The missing peer-report

Table 1

The Definitions Used To Measure the Cleanliness of Each Boy's Room

1. The bed should be made with two sheets, one pillow with case, and the bedspread (NOTE: there should be *no* blankets).
2. None of these objects should be visible except the bedspread while looking at the level of the steel frame that runs around the lower edge of the bed or while looking down at either end of the bed.
3. There should be no objects greater than 0.25 by 0.25 by 0.25 in. on the bedspread (this includes blankets).
4. The cord around the outside of the bedspread should not be more than 1 in. from the cord that runs around the outside edge of the mattress.
5. The pillow must be within 6 in. of the head of the bed and centered within 6 in. of the center of the mattress.
6. The bedspread must be tucked at least 1 in. under the pillow so that it makes a straight line across the bed. The line should not be more than 3 in. from a 90-degree angle with the edge of the bed.
7. There should be no wrinkles greater than 6 in. long, 1 in. wide, and 0.5 in. high below the pillow and between the cords on the bedspread.
8. All the clothes must be in the closet and must be on hangers except for hats, shoes, gloves, ties, belts, and folded clothes that belong in the dresser.
9. Shoes should be on the closet floor with the toes or heels touching a wall and the sides of a pair of shoes touching in one place. All shoes must face the same direction.
10. Objects greater than 0.25 by 0.25 by 0.25 in. should not be on the closet floor without permission.
11. All the hangers should be in the closet and hanging on the cross bar.
12. Hats, gloves, ties, and belts must be on hooks in the closet.
13. All clean clothes that are folded should be put away in the dresser.
14. Clothes should not be visible with the drawers closed.
15. The desk top must be clear of all objects greater than 0.25 by 0.25 by 0.25 in. except a lamp, clock, and other objects that you have permission to leave out.
16. All of these objects must be within 2 in. of the edge of the desk and must not extend over the edge.
17. The seat of the chair must be under the desk with the back of the chair within 0.5 in. of the desk and the drawers should be closed so that the drawer is not open more than 1 in.
18. No object greater than 0.25 by 0.25 by 0.25 in. should be on the floor of the room other than the furniture (bed, dresser, desk, chair, waste basket, electric cords, and other items with permission). There should be no dirt or dust that covers an area greater than 0.5 by 0.5 in.
19. The windows should be closed so that there is no light visible between the window sill and the bottom of the window.
20. The curtains, window ledges, and window glass should be clear of all objects.
21. The back of all the furniture except the bed must be within 6 in. of the wall and one end should not deviate from the other more than 3 in. The head of the bed should be at least 6 in. but not more than 18 in. from the wall.

data on the fourteenth day of this condition was the result of no checksheets being available to the boys that day).

The self-reports and peer-reports given by each boy and the measured behavior for each boy (data not shown) were more variable than the group means presented in Figure 1 but the group average is representative of the difference between the reported behavior and the measured behavior for each boy.

Discussion

The results of Experiment I indicated that both self-reports and peer-reports were unreliable when compared with the observer's report

of room cleaning. When the first self-report condition was initiated, there was a small increase in the measured behavior, but this effect rapidly disappeared. Neither the second self-report nor the peer-report conditions produced any observable effect on the measured behavior.

The results of Experiment I indicate that there is little correspondence between self-reports and observer reports. This conclusion is also supported by the results obtained by Risley and Hart (1968) and Broden *et al.,* (1972). Experiment I also showed that peer-reports were unreliable when compared to observer reports. This result is not consistent with the Surratt *et al.,* (1968) finding that inter-observer agreement between a fifth-grade peer-observer and a teacher averaged about 95%.

When considering the effects of self-reporting procedures on the measured behavior, the results of Experiment I indicated there was no reliable effect. This is in sharp contrast to the impressive results obtained by Broden, *et al.,* (1970), although they did find for one subject that continued exposure to the self-reporting procedure attenuated its effect on the measured behavior.

EXPERIMENT II

Reliability Training

Since the boys were not "naturally" reliable self-reporters or peer-reporters, attention was directed toward the production of reliable reporting. Risley and Hart (1968) produced an increase in the reliability of self-reporting by making food contingent upon accurate self-reports. The degree of reliability obtained by Risley and Hart was comparable to that found by Broden, *et al.,* (1972, Exp. 1). The purpose of Experiment II was to explore several possible methods of training the boys at Achievement Place to give reliable reports of their own behavior and the behavior of their peers.

Procedures

Subjects. The subjects were the same as those in Experiment I except for one boy. One of the boys in Experiment I moved away before Experiment II began and he was replaced by another boy who had recently entered the Achievement Place program. After the new boy (Leonard) had earned his way onto the same point system the other five boys were on, he was then included in Experiment II. Leonard's late arrival accounts for the missing data for Leonard and Lucius in Figure 2.

Observation. Room-cleaning data were collected as described in Experiment I using the same 21 definitions. Inter-observer agreement, computed as described in Experiment I, averaged 91% and ranged from 86% to 95% for the 10 reliability checks in Experiment II.

Throughout this experiment, each boy was required to turn in both a self-report on his own room and a peer-report on a designated peer's room. For purpose of clarity in describing the procedures used, the term self-report is used to refer to the report a boy turned in on his own room and the term peer-report is used to refer to the report he turned in on his peer's room. The term peer's report is used to refer to the report the peer-reporter turned in for the self-reporter's room. Thus, when the agreement between a self-report and the peer's report is discussed, it should be clear that both reports concerned the cleanliness of the same room, even though the two reports were submitted by two different boys.

Because of the experimental design and the fact that each boy was required to submit both a report on his own room and a report on a peer's room, each boy was assigned a "buddy". Thus, Clark and John were "buddies" and, in addition to each boy giving a self-report on his own room, Clark would give a peer-report on John's room and John would give a peer-report on Clark's room. The other "buddy" pairs were Ralph and Charles and Leonard and Lucius. (Each boy had the same "buddy" he had in Experiment I, except Lucius.)

The basic experimental design was a multiple baseline design in which data were recorded simultaneously for each boy, but each

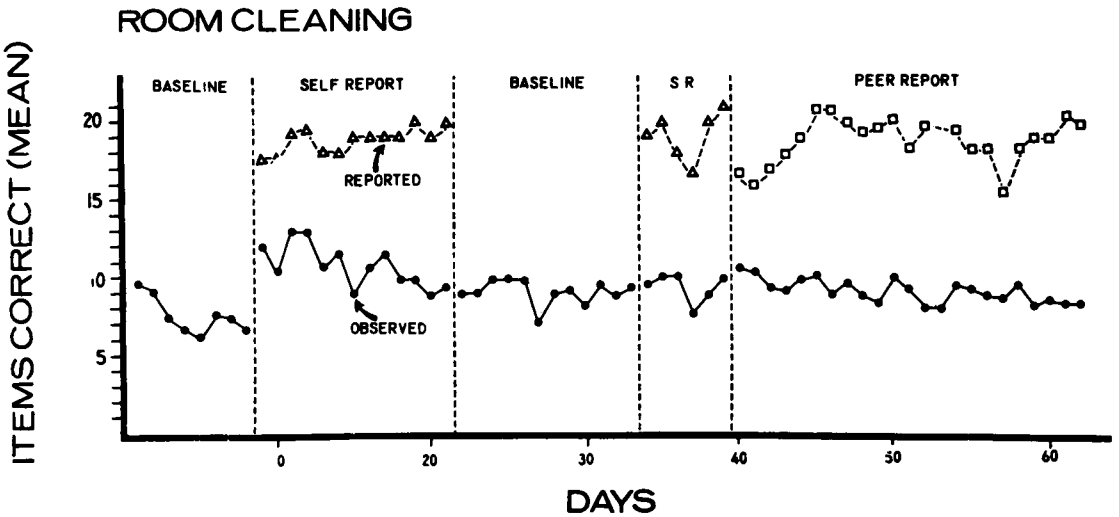

Fig. 1. The mean number of room-cleaning items that were correct as reported by an adult observer ("observed") and by the boys ("reported").

of the "buddy" pairs were subjected to the experimental procedures one pair at a time.

Conditions and Results

17/21 = 500 Report. In this condition, each boy was given 500 points if his self-report on his room agreed with the peer's report on the same room on 17 or more of the 21 items possible. There were no scheduled consequences for agreement less than 17 out of 21. Agreements were counted by comparing the self-report with the peer's report, item-by-item. Thus, if both the self-report and the peer's report had a check mark or if both had a zero for a given item, it counted as an agreement. All other combinations of checks and zeros were counted as disagreements.

Figure 2 shows the item-by-item agreement between each boy's self-report and the adult observer's report and between each boy's peer-report and the adult observer's report. For example, the self-report check sheet Clark completed for his own room was compared, item-by-item, with the adult observer's report for Clark's room. Similarly, the peer-report check sheet Clark completed for John's room was compared with the adult observer's report for John's room. The number of agreements with the adult observer's report in each case is plotted for

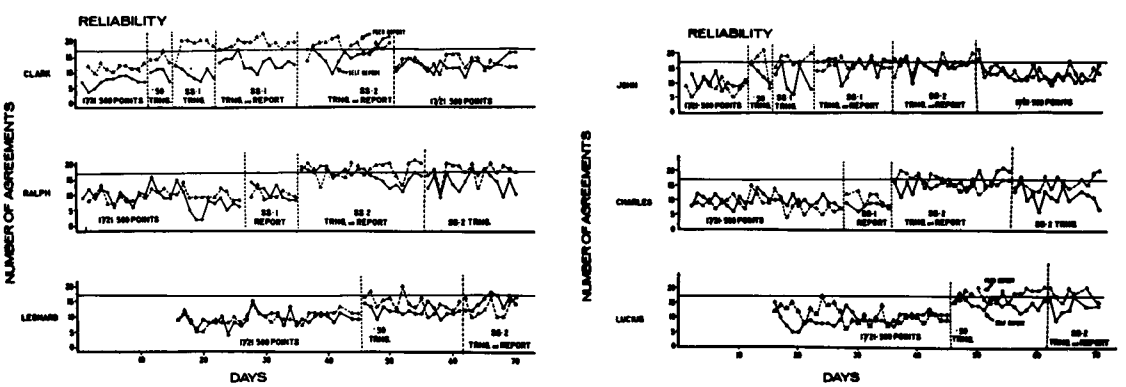

Fig. 2. The item-by-item agreement between each boy's self-report and the adult observer's report (solid line) and between each boy's peer-report and the adult observer's report (broken line).

Clark in Figure 2. The number of agreements given on the ordinate in Figure 2 can be easily transformed into per cent inter-observer agreement or reliability by dividing the number of agreements for either the self-report or peer-report on a given day by 21, the total number of agreements plus disagreements. Thus, the data in Figure 2 represent daily measures of reliability between two independent "observers" where one "observer" was the boy and the other was the adult observer. The line drawn across each graph in Figure 2 at 17 agreements represents about 80% agreement between the boy's report and the adult observer's report. Table 2 presents a summary of the procedures and results of Experiment II.

As shown in Figure 2, the self-report and peer-report for each of the six boys agreed with the adult observer's report about 10 times out of 21 possible under the 17/21 = 500 Report procedure. That is, the boys' reports on their own rooms and on their "buddies'" rooms were only about 50% reliable in this condition. This finding replicates the results of Experiment I.

In the 17/21 = 500 Report procedure, the point consequences were available for 17 or more agreements between the self-report and the peer's report. The agreement between these two measures averaged 16 across all of the boys, or 76% inter-"observer" agreement. Thus, there was fair agreement between the boys' reports (76%) but poor agreement (50%) between the boys' self- and peer-reports and the observer's reports.

±50 Training. In this condition, each boy continued to receive 500 points if his self-report agreed with the peer's report on 17 or more of the 21 items. In addition, each boy could make or lose points when he was given training on the definitions used. The training, which was given each day after school, consisted of the boy going to his "buddy's" room with the observer where the boy read aloud each definition, checked the room, and verbally stated whether the room met that definition or not. Immediate verbal feedback concerning the correctness of the boy's statement was given by the observer. If the boy's statement was incorrect, the observer also pointed out the parts of the room relevant to the definition and briefly discussed the source of the error. After the statement by the boy and the feedback by the observer, the check sheet the boy had filled out that morning before school for his "buddy's" room was checked to see if he had marked it correctly. For each item, if the check sheet had been marked correctly the boy was given 50 points; if it had been marked incorrectly he lost 50 points. If all 21 items had been marked correctly, the boy was given 1050 points (see Table 3). Thus, point consequences during training were for correspondence between the peer-report and the observer's report.

Clark and John and Leonard and Lucius participated in this procedure. All four boys gave more accurate self-reports and peer-reports under this condition relative to the 17/21 = 500 Report procedure. For all four boys, the peer-report agreed more closely with the observer's report (average of 16 agreements or 76% reliability) than did the self-report (average of 13 agreements or 62% reliability). Agreement between the boy's self-reports on their own room and the peers' reports on the same room averaged 13 agreements or 62% reliability even though the 17/21 = 500 points consequence was still in effect. Thus, relative to the 17/21 = 500 Report condition, the ± 50 Training condition produced an increase in the reliability of peer-reporting from about 50% to 76%, produced a smaller increase in the reliability of self-reporting from about 50% to 62%, and decreased the agreement between the self-reports and the peer's reports from 76% to 62%.

Sliding-scale (1) training. This condition was identical to the ± 50 Training condition except that the point consequences were delivered according to a "sliding scale" rather than giving the boy 50 points for each correctly marked item and taking away 50 points for each incorrectly marked item. The sliding scale is shown in Table 3 under SS-1. When the sliding-scale method of delivering points is compared to the ± 50 points

Table 2

A Summary of the Procedures and Results in Experiment II

Condition	*Number of Subjects*	*Consequences During Training on Peer-Report*	*Consequences for Agreement Between Self-Report and Peer's Report*	*Average Number of Agreements*		
				Self-Report/ Observer	*Peer-Report/ Observer*	*Self-Report/ Peer's Report*
17/21=500 Report	6	NO TRAINING	17+/21=500 Points	9.8	10.2	16.0
SS-1 Report	2	NO TRAINING	SS-1	9.6	11.0	18.0
± 50 Training	4	± 50 Points	17+/21=500 Points	13.0	16.0	13.0
SS-1 Training	2	SS-1	17+/21=500 Points	9.8	19.0	10.1
SS-1 Training and Report	2	SS-1	SS-1	15.0	16.5	15.0
SS-2 Training and Report	6	SS-2	SS-2	15.3	16.3	16.3
SS-2 Training	2	SS-2	NONE	12.5	17.0	13.0

Table 3

Number of Points Given for Each Number of Agreements in Each Point Condition.

Number of Correctly Marked Items	*Point Conditions* ± 50	*SS-1*	*SS-2*
21	+ 1050	+ 1050	+ 1050
20	+ 950	+ 1000	+ 1000
19	+ 850	+ 950	+ 950
18	+ 750	+ 900	+ 900
17	+ 650	+ 850	+ 850
16	+ 550	− 300	− 850
15	+ 450	− 350	− 900
14	+ 350	− 400	− 950
13	+ 250	− 450	− 1000
12	+ 150	− 500	− 1050
11	+ 50	− 550	− 1100
10	− 50	− 600	− 1150
9	− 150	− 650	− 1200
8	− 250	− 700	− 1250
7	− 350	− 750	− 1300
6	− 450	− 800	− 1350
5	− 550	− 850	− 1400
4	− 650	− 900	− 1450
3	− 750	− 950	− 1500
2	− 850	− 1000	− 1550
1	− 950	− 1050	− 1600
0	− 1050	− 1100	− 1650

method, three differences are apparent. First, the sliding-scale method allows the specification of a given level of behavior at and above which points are earned and below which points are lost. As shown in Table 3, when an equal number of points are earned or lost for each item marked correctly or incorrectly as in the ± 50 points method, the "specified level of behavior" will always be one-half of the total. Second, the sliding scale allows greater flexibility in assigning point consequences to each level of behavior, as can be seen by comparing SS-1 to SS-2. Third, the ± 50 method has a near zero point (at 10 and 11) where the boy can "break even", whereas the sliding-scale method allows a sizable consequence to be set for each level of behavior.

Only Clark and John were exposed to the Sliding-Scale (1) Training procedure. Under this procedure, the number of agreements between Clark's peer-report and the observer's report increased to a median of 19 agreements. No effect was observed for Clark's self-report or for either report for John. For both boys, the agreement between the self-reports and the peer's reports averaged 10 agreements or about 50% reliability even though the 17/21 = 500 points consequence was still in effect.

Sliding-scale (1) training and report. This condition was identical to the Sliding-Scale (1) Training condition except that points delivered according to the sliding scale (SS-1 in Table 3) were also made contingent on agreement between the self-report and the peer's report. Thus, if the self-report agreed with the peer's report on 17 or more of the 21 items, the boy earned points and he lost points if there were fewer than 17 agreements, according to the sliding scale.

Clark and John participated in the Sliding-Scale (1) Training and Report procedure. Changing the consequences for agreement between the self-report and the peer's report increased the reliability of the self-report for each boy relative to the Sliding-Scale (1) Training condition. The agreement between the self-report and the observer's report increased from a median of nine agreements to 13 agreements for Clark and from a median of 9.5 to 17 agreements for John. Agreement between the self-report and peer's report also increased to a median of 15 agreements for both boys.

Sliding-scale (1) report. Identical to the 17/21 = 500 Report condition except that points earned for agreement between a boy's self-report and the peer's report were delivered according to the first sliding scale (SS-1 in Table 3). There was no training given during this condition.

The change in consequences for agreement between the self-report and the peer's report produced no observable change in the reliability of the self-report or peer-report for Ralph or Charles, the two boys who were exposed to this procedure. However, agreement between the self-report and the peer's report for the two boys increased from 15 to 18 agreements.

Sliding-scale (2) training and report. Identical to the Sliding-Scale (1) Training and Report condition except the second sliding scale (SS-2 in Table 3) was used to deliver points during training and for agreement between the self-report and the peer's report.

All six boys were exposed to the Sliding-Scale (2) Training and Report procedure. By the end of this condition, all six boys' self-reports and peer-reports were near or above the criterion of 17 agreements with the observer's report. Across all the sessions for all the boys in this condition, the self-report averaged 15.3 agreements with the observer report, the peer-report averaged 16.5 agreements with the observer's report, and the self-report averaged 16.3 agreements with the peer's report, or reliabilities of 73%, 79%, and 78%, respectively.

Sliding-scale (2) training. Identical to the Sliding-Scale (2) Training and Report condition except there were no point consequences for agreement between a boy's self-report and the peer's report. Points during training were given according to the second sliding scale (SS-2 in Table 3).

Ralph and Charles participated in this procedure following the Sliding-Scale (2) Training and Report condition. The effect of removing the consequences for agreement between the self-report and peer's report was a reduction in the reliability of the self-report for each boy. The sliding-scale (2) consequences for agreement between the peer-report and the observer's report during the training sessions maintained a high level of agreement between each boy's peer-report and the observer's report.

Clark and John were returned to the 17/21 = 500 Report condition following the Sliding-Scale (2) Training and Report condition and the reliability of each boy's self-report and peer-report decreased to a median of 12 agreements, or 57% agreement with the observer.

Effects on Room Cleaning

None of the procedures used in Experiment II produced a systematic effect on room cleaning (data not shown). The average number of room-cleaning definitions met under each condition varied from 9.3 to 12.8 but there was also considerable variability within each condition. Other research (Phillips, Phillips, Fixsen, and Wolf, 1971) has demonstrated that room-cleaning behavior can be manipulated directly by making points contingent upon performance of the room-cleaning tasks.

Discussion

The results of Experiment II demonstrated that reliability of recording was sensitive to the experimental contingencies and consequences. When no training was given and 500 points were made contingent on 17 or more agreements between the self-report and peer's report in the 17/21 = 500 Report condition, the two reports agreed 16 times on the average, *i.e.*, 76% inter-"observer" reliability. When the consequences were increased in the Sliding-Scale (1) Report condition, agreement between the two reports averaged 18 agreements, *i.e.*, 86% inter-"observer" reliability. In both conditions, however, the self-report and peer-report each averaged only about 10 agreements with the observer's report, *i.e.*, about 50% agreement with the adult observer.

Point consequences during training were for correspondence between the peer-report and the observer's report and that contingency produced a greater number of agreements between the peer-report and the observer's report. In the ± 50 Training condition, agreement between the peer-report and observer's report averaged 16 (76% reliability), in the Sliding-Scale (1) Training condition agreement averaged 19 (90% reliability), and in the Sliding-Scale (2) Training condition agreement averaged 17 (81% reliability). In the ± 50 Training condition, and the Sliding-Scale (1) Training condition, 500 points were also contingent on agreement between the self-report and the peer's report on 17 or more of the 21 items. However, in both conditions, agreement between the self-

report and peer's report averaged fewer than 13 agreements (less than 62% reliability) as did the agreement between the self-report and the observer's report. The 500-point contingency for agreement between the self-report and the peer's report did not include a point loss if fewer than 17 agreements were obtained. Thus, if a boy had fewer than 17 agreements between his self-report and the peer's report he did not earn the 500 points nor did he lose any points. The use of the sliding scale rectified that problem. With the sliding scale (see Table 3), points were either made or lost for each number of agreements. Under the Sliding-Scale (1) Training and Report condition the number of agreements between the peer-report and the observer's report remained high (16.5 agreements or 79% reliability) and the agreement between the self-report and the peer's report increased to 15 agreements (71% reliability), as did the agreement between the self-report and the observer's report. The effect of the second sliding scale in the Sliding-Scale (2) Training and Report condition was further improvement in agreement between the self-report and the peer's report to 16.3 agreements (78% reliability).

The importance of the sliding-scale points to maintaining agreement between the self-report and the peer's report was further indicated by the results of the Sliding-Scale (2) Training condition, where point consequences for agreement were absent. Under this condition, the self-report agreed with the peer's report an average of 13 times (62% reliability) compared to the 16.3 agreements (78% reliability) found under the Sliding-Scale (2) Training and Report condition.

Finally, it should be noted that the agreement between the self-report and the observer's report was a joint function of the contingencies on the peer-report during training and the contingencies for agreement between the self-report and peer's report. Highly reliable peer-reporting did not lead to highly reliable self-reporting, as shown in the Sliding-Scale (1) Training condition. This was true even though the definitions and discriminations required to mark a peer-report accurately were exactly the same as those required to mark a self-report accurately. The results of the Sliding-Scale (1) Report condition showed that a high level of agreement between the self-report and peer's report did not lead to reliable self-reporting. It was only when both contingencies were in effect during the Sliding-Scale (1) Training and Report and Sliding-Scale (2) Training and Report conditions that the reliability of self-reporting exceeded 70%. Thus, with both contingencies in effect the number of agreements between the self-report and the observer's report increased, even though there were no consequences that were directly contingent on agreement between the two reports.

GENERAL DISCUSSION

The results of these two experiments indicate that at least some youths are not "naturally" reliable observers and that the reliability of self-reporting and peer-reporting can be improved with appropriate contingencies and consequences. The practicality of having a subject record his own behavior and having a peer record the same behavior at times to assess reliability must be questioned in view of the results of Experiment II. Under the $17/21 = 500$ Report and the Sliding-Scale (1) Report conditions, the self-report agreed to a considerable extent with the peer's report but the self-report and the peer-report were each only about 50% reliable when compared to the adult observer's report. In addition, the reliability of both the self-report and the peer-report decreased for Clark and John when training was discontinued and 500 points were again made contingent on 17 or more agreements between the self-report and peer's report. Both of these findings suggest that further research is required to develop procedures to generalize and maintain reliable reporting in the absence of daily contingencies and consequences.

REFERENCES

Azrin, N. H., and Powell, J. Behavioral engineering: the use of response priming to improve prescribed self-medication. *Journal of Applied Behavior Analysis,* 1969, **2**, 39-42.

Baer, D. M., Wolf, M. M., and Risley, T. R. Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis,* 1968, **1**, 91-97.

Broden, M., Hall, R. V., and Mitts, B. The effect of self-recording on the classroom behavior of two eighth grade students. *Journal of Applied Behavior Analysis,* 1971, **4**, 191-199.

Phillips, E. L. Achievement Place: token reinforcement procedures in a home-style rehabilitation setting for pre-delinquent boys. *Journal of Applied Behavior Analysis,* 1968, **1**, 213-223.

Phillips, E. L., Phillips, E. A., Fixsen, D. L. and Wolf, M.M. Achievement Place: modification of the behaviors of pre-delinquent boys within a token economy. *Journal of Applied Behavior Analysis,* 1971, **4**, 45-59.

Risley, T. R. and Hart, B. Developing correspondence between the non-verbal and verbal behavior of preschool children. *Journal of Applied Behavior Analysis,* 1968, **1**, 267-281.

Surratt, P. R., Ulrich, R. E., and Hawkins, R. P. An elementary student as a behavioral engineer. *Journal of Applied Behavior Analysis,* 1969, **2**, 85-92.

Received 9 December 1970.
(Revised 1 December 1971.)