*THE EFFECTS OF INSTRUCTIONS AND CALCULATION
PROCEDURES ON OBSERVERS' ACCURACY, AGREEMENT,
AND CALCULATION CORRECTNESS*

RONALD A. BOYKIN AND ROSEMERY O. NELSON

UNIVERSITY OF NORTH CAROLINA AT GREENSBORO

Although the quality of observational data is generally evaluated by observer agreement, measures of both observer agreement and accuracy were available in the present study. Videotapes with a criterion protocol were coded by 16 observers. All observers calculated agreement scores both on their own and their partner's data and on a contrived data set misrepresented as data collected by other observers. Compared with agreement scores calculated by the experimenter, observers erroneously inflated their own agreement scores and deflated the agreement scores on the contrived data. Half of the observers ($n = 8$) had been given instructions emphasizing the importance of accuracy during observation while the other half had been given instructions emphasizing interobserver agreement. Accuracy exceeded agreement for the former group, whereas agreement exceeded accuracy for the latter group. The implications are that agreement should be calculated by the experimenter and that the accuracy-agreement relationship can be altered by differential observer instructions.

DESCRIPTORS: observational data, observer bias, observer agreement, observer accuracy, observer instructions

Applied behavior analysis frequently relies on data collected by human observers (Kelly, 1977). Many variables affecting the quality of observational data have been investigated, including knowledge that interobserver agreement is being assessed (Romanczyk, Kent, Diamant, & O'Leary, 1973) and observer training procedures (Wildman, Erickson, & Kent, 1975).

Of particular interest to the present study are experimenters' verbal interactions with observers. On one hand, experimenters' disclosures of experimental hypotheses (Kent, O'Leary, Dia-

ment, & Dietz, 1974) or of labels describing children's behavior problems (Shuller & MacNamara, 1976) have not influenced observational data. On the other hand, experimenters' feedback on the content of observational recordings has influenced the subsequent frequency with which certain behavioral categories were recorded (O'Leary, Kent, & Kanowitz, 1975). A variable investigated in the present study was the instructions given by the experimenter to the observers. Half of the observers received instructions emphasizing interobserver agreement as the hallmark of the quality of observational data; the other half of the observers received instructions emphasizing accuracy as the hallmark.

The quality of observational data is usually judged from interobserver agreement scores because of the difficulty in procuring a criterion against which to measure the observers' actual accuracy. Possible accuracy criteria, however, include mechanical measurements of behavior (e.g., Bechtel, 1967), mechanically generated re-

The authors thank K. Daniel O'Leary and Ronald N. Kent of the State University of New York at Stony Brook for supplying copies of the videotapes and protocols used in this study. The authors also thank Skip Beck for his contributions during the inception of this study, and the Research Council of the University of North Carolina at Greensboro for grant 977. Requests for reprints should be sent to Rosemery O. Nelson, Psychology Department, University of North Carolina, Greensboro, North Carolina 27412.

sponses (e.g., Repp, Roberts, Slack, Repp, & Berkler, 1976), recorded behaviors orchestrated by a predetermined script (e.g., Mash & McElwee, 1974), and consensually validated criterion protocols produced by the observation of multiple observers (e.g., Kent et al., 1974; Foster & Cone, 1980). Although agreement is generally used to evaluate the quality of observational data, agreement and accuracy are not the same (Foster & Cone, 1980; Johnson & Bolstad, 1973; Kazdin, 1977). The relationship between agreement and accuracy has seldom been investigated. It is not known, for example, if agreement and accuracy are related in a linear fashion where, given particular independent variables, accuracy exceeds or falls short of agreement by X points, or if agreement and accuracy are related in an interactive fashion, where particular independent variables affect accuracy and agreement in different ways. In the present study, both accuracy and agreement measures were obtained. The effects of accuracy and agreement instructions were examined not only on accuracy and agreement separately, but also on the relationship between accuracy and agreement.

Another variable that influences interobserver agreement scores is the calculation procedures that follow data collection. O'Leary and Kent (1973) reported that agreement scores were calculated to be 11 points higher when the experimenter supervising the observers was absent during the calculations than when he was present. They also reported that agreement was eight points higher when observers calculated their own agreement as opposed to calculations performed by the experimenter (see also Kent & Foster, 1977). Similarly, Kent et al. (1974) found that agreement calculated by the observers exceeded agreement calculated by the experimenters by six points. The present study investigated the effects of another variable on calculations of observer agreement: whether the data used in calculating agreement were the subject's own and partner's data, or data that the subject was told were collected by two other observers.

METHOD

Participants

Nineteen undergraduate students from an introductory psychology course received extra credits for participating in the experiment. To assure that arithmetic skill was generally equivalent in the two experimental groups, these students completed 10 problems from the arithmetic portion of the *Wide Range Achievement Test* (Jastak, Bijou, & Jastak, 1965). Training in coding children's behavior from videotapes continued until 16 students had obtained a preestablished level of observational accuracy, as described below. These 16 students (14 female, 2 male) became the experimental participants.

Participants were paired according to the following factors: arithmetic scores (four students who scored below 80% were assigned partners who scored 80% or above); number of hours of training (students who were trained the minimum of one hour were assigned partners who were trained the maximum of three hours); training session attendance (students were assigned partners with whom they had little or no contact during training); and students' schedules. Pairs of students were then randomly assigned to the two experimental groups, with the constraint that the four student pairs that included a student who had scored below 80% on the arithmetic test were equally divided between the two groups. The mean arithmetic scores for the agreement and accuracy instructional groups (explained below) were 86% and 80%, respectively. The mean hours of training per student for the agreement and accuracy instructional groups were 2.13 and 2.38, respectively.

Experimental Design

The experimental design was a 2 (sets of instructions) $\times$ 2 (sets of data) $\times$ 3 (target behaviors) $\times$ 8 (trials) mixed design. Instructions, the between-subjects factor, were either to try to obtain a high level of interobserver agreement (.85), or to try to perform carefully the observa-

tional recording and calculation tasks. Sets of data were either the pair's own data for a given trial, or were contrived by the experimenter to appear as data collected by other students in the experiment. Target behaviors were playing, vocalizing, and orienting. A trial was completed when a student pair had recorded the occurrence of the three target behaviors from a 10-min segment of videotape and had calculated interobserver agreement for each target behavior for each of the two sets of data.

Apparatus

Videotapes were selected from a collection of tapes developed by researchers at the State University of New York at Stony Brook (K. Daniel O'Leary, Ronald N. Kent, and their colleagues). Each tape depicts the behavior of two young males in a classroom. The Stony Brook researchers also developed concensually determined protocols for each tape. These protocols were assumed to reflect closely the actual occurrence of the target behaviors and were used as the standard for assessing students' observational accuracy.

The tapes were presented by a Panasonic videotape recorder (Model NV 3020) and 19-inch (48.26 cm.) diagonal screen monitor (Model AN69V). A Sharp cassette tape recorder presented an audio signal to cue students to correct observing and recording intervals.

The blank data sheets used in the experiment were equipped with carbon paper so that students could exchange their data for each trial and calculate their own agreement figures simultaneously.

The experimenter contrived eight sets of data to appear as data collected by students in the experiment. Of each set, one data sheet represented the protocol for the videotape for that trial; the other data sheet was constructed so that correct agreement levels for the data set varied around .85.

Each student completed a questionnaire at the end of the last trial to determine whether the

student could correctly recognize the instructions he or she had been given.

Observational Procedures and Calculations of Interobserver Agreement and of Accuracy

Students coded the occurrence of three target behaviors (playing, vocalizing, and orienting) from the videotapes. These behaviors comprise part of the Classroom Behavior Code (O'Leary, Kaufman, Kass, & Drabman, 1970) and are defined as follows: Playing is coded when a child uses his or her hands to manipulate his or her own or community property; vocalizing is coded when a non-permitted audible behavior emanates from the child's mouth; orienting is coded when a child, while seated, turns more than 90 degrees from the point of reference. These three behaviors were selected from the Classroom Behavior Code because of their previously reported high frequency (7.22, 4.32, and 4.30 per child per 12.5 min for playing, vocalizing, and orienting, respectively) and respectable reliabilities (.73, .62, and .86 for playing, vocalizing, and orienting, respectively) (Kent, O'Leary, Coletti, & Drabman, Note 1).

Students used a time-sampling coding procedure identical to that used in developing the protocols. Intervals of 30 sec were used; the first 20 sec were designated for observation of all three target behaviors simultaneously, and the final 10 sec for recording the observations.

Students used the exact agreements formula (Repp, Dietz, Boles, Dietz, & Repp, 1976) to calculate interobserver agreement: number of agreement intervals divided by the total of agreement and disagreement intervals. Intervals in which neither observer recorded the occurrence of a behavior were omitted from the calculations. The same formula was used to calculate observer accuracy by comparing the student's recording with the observational protocol.

Training in Observations

Students received a training manual which contained a rationale for the experiment, the behavioral code, coding procedures, an explana-

tion of the time-sampling procedure, procedures for calculating interobserver agreement, and a model set of data correctly coded and with correctly calculated agreement figures. The experimenter reviewed the manual orally and answered questions about it at the first training session.

Five-minute segments from two training tapes were used in training students. The first 5-min segment exposed students to the format of the tapes, allowed the experimenter to point out examples of each behavior, and permitted students to discuss orally their observations interval by interval.

Subsequent 5-min segments allowed students to practice observational recording. At the end of the segment, the experimenter used the protocol (which students were told were the experimenter's own observations) to provide feedback to students regarding their observational accuracy. Students also practiced calculating interobserver agreement at this time; the experimenter moved around the room and provided individual feedback.

A student was considered to be trained when his or her accuracy level for that 5-min segment reached .80 or higher for two of the three target behaviors and when the interobserver agreement calculations were correctly completed. Sixteen students achieved these training criteria by the end of the third 1-hr training session.

Experimental Procedures

The experimenter was the first author. Upon entering the laboratory room, the student pair was seated about 4 ft. (1.2 m) apart, and about 7 ft. (2.1 m) from the video monitor. The behavior code and formula for calculating agreement were printed on the blackboard behind the monitor. On each odd-numbered trial, different instructions were read to the student pair, depending on their experimental group. Unique instructions for students in the agreement group were: "Let me emphasize one thing about your own observations. One way researchers judge the quality of the data collected by observers is

by how closely the records of the observers compare. Minimal interobserver agreement levels of .85 indicate that the data collected are of sufficient quality to be used in reporting the results of a treatment program. Hopefully, you and your partner will achieve a .85 level of agreement on most of your observations." Unique instructions for students in the accuracy group were: "Let me emphasize one thing about your observations. One way researchers judge the quality of the data collected by observers is by how carefully the experimenter feels each observer is collecting his or her own data. Observers are usually asked to consider themselves as independent recording devices in order for the data they collect to be of high quality. The frequency data you and your partner collect are of primary interest to me. These data are averaged, so please try to perform your observational recordings and calculations as carefully as possible." The intent of the phrase "these data are averaged" was that students should be concerned about the accuracy of their own data rather than about interobserver agreement because averaged frequencies would be used in any future hypothetical uses of the data.

After hearing the instructions, the observer pair then made observational recordings on a child specified by the experimenter from a 10-min segment of videotape. The experimenter remained in the room during this time and then collected the observers' data sheets. The experimenter then gave each observer a data set and asked him or her to calculate agreement figures for the three behaviors. After completing this task, the observer was given a second data set on which to perform the same task. The data sets were either the observer pair's own data or were contrived by the experimenter to look like data collected by other observers. The order of presentation of the two data sets was counter-balanced to control for sequence effects. The trial ended when each observer produced six agreement figures (two data sets times three target behaviors). The students did not receive feedback from the experimenter about these agreement

figures. Each observer pair completed four trials on each of two days.

Check on Instructional Manipulation

At the end of the final trial, each student completed a multiple-choice questionnaire designed to ascertain if he or she could correctly recall the experimental instructions. In the accuracy group, 100% (8 of 8 students) correctly recognized the instructions given them; 63% (5 of 8) in the agreement group correctly recognized the instructions given them. The other 37% (3 of 8) in this group selected both options: that they were instructed to try to reach .85 agreement levels and also to perform the observational recording and calculation tasks as carefully as possible. Thus, all students correctly recognized the instructions that had been given to them. Three students in the agreement group also selected the additional response option of being careful, but no students in the accuracy group selected the additional response option of reaching .85 agreement levels. A possible explanation is that an experimental demand to be careful is probably more widespread than an experimental demand to agree with another student.

RESULTS

Errors in the Calculation of InterObserver Agreement

All students calculated agreement on their own and partner's data, and on a data set contrived by the experimenter disguised as data collected by two other students. The experimenter also calculated agreement scores for these two data sets.

Comparisons of student-calculated and experimenter-calculated agreement for each trial.

A 2 (instructional groups) $\times$ 3 (behaviors) $\times$ 8 (trials) $\times$ 2 (student versus experimenter) analysis of variance was conducted on agreement scores for each data set. For the students' own data, the main effect for student versus experimenter calculations was significant, $F(1, 14) =$

3.47, $p < .08$. (In this and other analyses of variance reported in this article, although other main effects and interactions are statistically significant, only those effects directly related to the experimental hypotheses are reported. Copies of the complete summary tables for the analyses of variance are available from the second author.) The mean agreement calculated by the students on their own data per trial per behavior was .605; the related mean agreement calculated by the experimenter was .599. For the contrived data set, the main effect for student versus experimenter calculations was not significant, $F(1, 14) = 1.27$, $p > .10$. The mean agreement calculated by the subjects on the contrived data per trial per behavior was .819; the related mean agreement calculated by the experimenter was .826.

These analyses seem somewhat problematic from the perspective of applied research in two respects. First, applied researchers are aware that means may sometimes be deceiving. Thus, while the above analyses revealed that the mean experimenter calculations and mean students calculations were quite similar, the analyses fail to reveal how the student and experimenter calculations differed for individual observers. To demonstrate, if Student 1's calculations (e.g., .86) exceeded the experimenter's calculations (e.g., .80), but Student 2's calculations (e.g., .80) underestimated the experimenter's calculations (e.g., .86), the means for the student-calculated agreement and the experimenter-calculated agreement could nonetheless be similar (e.g., .83). The following analysis, therefore, uses signed difference scores between student and experimenter calculations on a per trial and per behavior basis. Related arguments favoring individual subjects' difference scores over group means are presented by Huck and McLean (1975). Second, applied researchers are seldom interested in only 10 min of observational data. If data are averaged across the eight trials or 80 observational minutes of the present study, the same problem of obscuring signed differences occurs. Therefore, the following analysis uses a

Fig. 1. The difference between student-calculated and experimenter-calculated agreement scores for the accuracy group for the target behaviors, playing (P), vocalizing (V), and orienting (O), for their own data and for contrived "other" data. Bars above the zero line indicate that the student-calculated agreement scores were higher than the experimenter's; bars below the zero line indicate that the student-calculated agreement scores were lower than the experimenter's.

sum of signed difference scores across eight trials.

Comparisons of student-calculated and experimenter-calculated agreement summed across trials. For this analysis, difference scores were calculated by subtracting the experimenter-calculated agreement score from the student-calculated agreement score for each behavior and each trial. For example, in the contrived data set, the correct agreement for the sixth trial of orienting was .86. If a student calculated .83 for this trial and this behavior, a difference score of $-.03$ was obtained. These difference scores were then summed across the eight trials to examine accumulated rather than average errors or deviations from the experimenter's calculations. A 2 (instructional groups) $\times$ 3 (behaviors) $\times$ 2 (data sets: own versus other) analysis of variance revealed a significant main effect for data

sets, $F(1, 14) = 9.64, p < .01$. The interaction of groups by data sets was not significant, $F(1, 14) = .019, p > .10$. In comparison to the experimenter calculations, students inflated their own agreement scores an average of four points (mean difference score = $+.04$) and deflated agreement scores for the "other" or contrived data set by seven points (mean difference score = $-.07$). Data for individual students are shown in Figures 1 and 2, where it can be seen that (summed across eight trials), students generally inflated their own agreement scores and deflated the agreement scores of others in comparison to the experimenter's calculations.

Whereas the former analyses showed that student-calculated and experimenter-calculated mean agreement scores for individual trials were quite similar, the differences between student and experimenter calculations become rather

Fig. 2. The differences between student-calculated and experimenter-calculated agreement scores for the *agreement* group for the target behaviors, playing (P), vocalizing (V), and orienting (O), for their own data and for contrived "other" data. Bars above the zero line indicate that the student-calculated agreement scores were higher than the experimenter's; bars below the zero line indicate that the student-calculated agreement scores were lower than the experimenter's.

substantive when summed across trials. Moreover, the accumulated differences in calculations are systematic with student calculations being higher than the experimenter's for their own agreement scores, and lower than the experimenter's for the agreement scores of others.

Effects of Instructions on Agreement and Accuracy

In this portion of the analyses, only experimenter calculations are used. Interobserver agreement was determined for each pair of students. Also, each student's accuracy was determined by comparing his or her data with the criterion protocol. Finally, because agreement is frequently used as an estimate of accuracy (when accuracy scores are not available), the relationship between accuracy and agreement was examined. This relationship was examined for each student for each trial for each behavior by calculating difference scores to determine if agree-

ment underestimated or overestimated accuracy. Each student's accuracy score was subtracted from his or her agreement score. For example, if on Trial 2 for vocalizing, a student's agreement score with her or his partner was .75, but her or his accuracy score obtained from the criterion protocol was .50, her or his difference score was +.25.

Agreement. A 2 (instructional group) $\times$ 3 (behaviors) $\times$ 8 (trials) analysis of variance was performed on the experimenter-calculated agreement scores. The main effect for groups was not significant, $F(1, 14) = .57, p > .10$. Thus, regardless of the instructions, the agreement group (mean agreement = .616) and the accuracy group (mean agreement = .584) achieved comparable levels of interobserver agreement.

Accuracy. A 2 (instructional groups) $\times$ 3 (behaviors) $\times$ 8 (trials) analysis of variance was performed on the experimenter-calculated accuracy scores. The main effect for groups was

Table 1

Experimenter-calculated accuracy and agreement scores for individual students averaged across trials.

Group		Accuracy			Agreement		
		Playing	Vocalizing	Orienting	Playing	Vocalizing	Orienting
Accuracy Instructions							
Student	1	.61	.57	.59	.52 ^a	.62	.66
	2	.47	.63	.44			
	3	.55	.62	.61			
	4	.52	.58	.60		.42	.58
	5	.72	.60	.63	.59	.61	.54
	6	.64	.58	.70			
	7	.52	.64	.74			
	8	.60	.65	.76		.49	.72
Agreement Instructions							
Student	9	.65	.64	.65	.75	.69	.82
	10	.63	.48	.68			
	11	.49	.64	.58			
	12	.50	.70	.60		.58	.72
	13	.35	.59	.65	.28	.52	.67
	14	.40	.59	.64			
	15	.41	.63	.54			
	16	.39	.44	.47		.53	.49

^aThere is one accuracy score for each student (per behavior averaged across eight trials) because each student's observational data were compared against the criterion protocol. There is one agreement score per *pair* of students (per behavior averaged across eight trials) because each student's data were compared against his or her partner's. Students 1 and 2 formed a pair, students 3 and 4 formed a pair, etc.

significant, $F(1, 14) = 3.26$, $p < .09$. The group receiving accuracy instructions was somewhat more accurate (mean accuracy = .607) than the group receiving agreement instructions (mean accuracy = .556).

Difference between agreement and accuracy. A 2 (instructional groups) $\times$ 3 (behaviors) $\times$ 8 (trials) analysis of variance was performed on the difference scores between each student's accuracy and agreement. The main effect for groups was significant, $F(1, 14) = 5.25$, $p < .05$. For students given instructions emphasizing agreement, agreement exceeded accuracy scores by five points (mean difference score = +.05). For students given instructions emphasizing accuracy, accuracy exceeded agreement scores by two points (mean difference score = -.02). Thus, the relationship between agreement and accuracy differed, depending on the instructions given to the students.

Accuracy and agreement scores for individual

students and pairs of students are presented in Table 1. In the 24 possible comparisons between accuracy and agreement scores, accuracy exceeded agreement in 15 comparisons for the groups receiving accuracy instructions; agreement exceeded accuracy in 19 comparisons for the group receiving agreement instructions.

The results of the analysis also showed that the relationship between agreement and accuracy is complex and cannot be determined by any one data point. For example, there was a significant behavior by groups interaction, $F(2, 28) = 3.59$, $p < .05$. In a Newman-Keuls post hoc analysis, the mean difference scores between accuracy and agreement were significantly different for the two instructional groups for the target behavior orienting, $p < .01$ (for the agreement group, agreement exceeded accuracy by 11 points; for the accuracy group, accuracy exceeded agreement by two points). These mean difference scores were also significantly different

for the target behavior playing, $p < .01$ (for the agreement group, agreement exceeded accuracy by three points; for the accuracy group, accuracy exceeded agreement by eight points). For the target behavior vocalizing, the mean differences did not differ significantly.

DISCUSSION

Two central findings highlight the present study. First, students erroneously inflated agreement levels for their own data, and deflated agreement levels for data contrived to appear collected by other students. Second, instructions affected the relationship between observer accuracy and interobserver agreement. The group of students to whom instructions emphasized the importance of obtaining high interobserver agreement obtained higher agreement scores than accuracy scores. The group receiving instructions emphasizing careful independent performance obtained higher accuracy scores than agreement scores.

Although the student-calculated and experimenter-calculated mean agreement scores for either their own or others' data were quite similar, it must be recalled that these means represent the agreement scores of 16 students (eight pairs) and of eight trials. In applied settings, data are generally obtained from only one pair of observers. Therefore, of greater relevance are the discrepancies found between student and experimenter calculations when data from individual students are examined in the analysis of variance on difference scores and in Figures 1 and 2. When these difference scores are used, it can be seen that students generally inflated their own agreement scores and deflated their agreement scores of others in comparison to the experimenter's calculations. Similarly, in applied settings, judgments of treatment effectiveness are generally made by evaluating trends in baseline and in intervention data over time. The differences between student and experimenter calculations, when summed across trials, not only vary systematically for their own data versus the

data of others, but also become rather substantive.

The finding that students erroneously inflated agreement levels for their own data and deflated agreement levels for others' data adds evidence to previous studies that calculations of observer agreement can be affected by specific environmental variables. O'Leary and Kent (1973) and Kent et al. (1974) reported erroneously inflated agreement scores by subjects on their own data of eight and six points, respectively, compared with experimenter calculations on the same data. Likewise, O'Leary and Kent (1973) reported that experimenter absence resulted in inflated agreement scores of 11 points. These results, and those of the present study, are generally consistent with the hypothesis that observers attempt to make themselves appear positive in the eyes of the experimenter (Kent & Foster, 1977). They do so not only by the direction of calculation errors, but also by producing greater agreement when agreement is known to be checked than when it is not (Reid, 1970; Taplin & Reid, 1973) and by producing data consistent with experimenter feedback (O'Leary et al., 1975). In any of these situations, conscious cheating is not necessarily implicated. Rather the observer most likely responds to the demands of the situation and alters his or her behavior so that it is most likely to produce positive consequences, including social consequences from the experimenter. In the present study, calculation errors, especially when accumulated across time, favored the observers when they calculated agreement on their own data and detracted from the "other" observers when the student calculated agreement on "others'" observational data. The implication of these findings is that it is preferable for experimenters rather than observers to calculate agreement scores.

The second finding of this study is that the relationship between agreement and accuracy is differentially affected by instructions given to observers. Although accuracy and agreement per se were only minimally affected by instructions,

observers achieved higher agreement than accuracy when given agreement instructions and higher accuracy than agreement when given accuracy instructions. Thus, the relationship between accuracy and agreement is not linear, but instead depends on the instructions given to the observers and on the specific target behavior being observed. The implication of this finding is that experimenters should emphasize accuracy rather than agreement in their instructions to observers. The resultant data may be more accurate even if agreement may be less high than with agreement instructions.

It should be noted that the observers in the accuracy instructional condition were not informed that a criterion protocol was available in this study, which makes the procedure and results more easily generalizable to typical observational settings. The results should be generalized with caution, however, to more experienced observers because the overall levels of agreement reached by these observers after an average of 2 h of training (.52, .62, and .66 for playing, vocalizing, and orienting, respectively) were somewhat lower than those reached by Kent et al.'s observers (Note 1) after a year of experience (.73, .62, and .86 for playing, vocalizing, and orienting, respectively).

It is interesting to speculate on the mechanism by which observers in the group given agreement instructions were able to increase their agreement levels over their accuracy levels. Although neither the accuracy nor agreement group received feedback from the experimenter during the eight experimental trials, student calculations of agreement with their partner provided some feedback on agreement. Because the experimenter remained in the room during both data collection and agreement calculations, there was little opportunity for observer collusion. The observers in the agreement group, however, may have used subtle clues like the movement of their partner's hand to improve their agreement scores.

In conclusion, the results of this study suggest that interobserver agreement scores should be

calculated by the experimenter rather than by observers and that observers should be given instructions emphasizing accuracy rather than interobserver agreement.

REFERENCE NOTE

1. Kent, R. N., O'Leary, K. D., Coletti, G., & Drabman, R. S. *Increasing the validity of an observational code: An empirical methodology*. Unpublished manuscript. Available from K. David O'Leary, Psychology Department, State University of New York, Stony Brook, New York 11794.

REFERENCES

- Bechtel, R. B. Human movement and architecture. *Trans-Action*, 1967, 4, 53-56.
- Foster, S. L., & Cone, J. D. Current issues in direct observation. *Behavioral Assessment*, 1980, 2, 313-338.
- Huck, S. W., & McLean, R. A. Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin*, 1975, 82, 511-518.
- Jastak, J. F., Bijou, S. W., & Jastak, S. R. *Wide Range Achievement Test: Reading, spelling, arithmetic from pre-school to college*. Wilmington: Guidance Associates of Delaware, Inc., 1965.
- Johnson, S. M., & Bolstad, O. D. Methodological issues in naturalistic observation: Some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts and practice*. Champaign, Ill.: Research Press, 1973.
- Kazdin, A. E. Artifact, bias, and complexity of assessment: The ABC's of reliability. *Journal of Applied Behavior Analysis*, 1977, 10, 141-150.
- Kelly, M. B. A review of the observational data collection and reliability procedures reported in the *Journal of Applied Behavior Analysis*. *Journal of Applied Behavior Analysis*, 1977, 10, 97-101.
- Kent, R. N., & Foster, S. L. Direct observational procedures: Methodological issues in natural settings. In A. R. Ciminero, K. W. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment*. New York: Wiley, 1977.
- Kent, R. N., O'Leary, K. D., Diament, C., & Dietz, A. Expectation biases in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology*, 1974, 42, 774-780.
- Mash, E. J., & McElwee, J. D. Situational effects on observer accuracy: Behavioral predictability, prior experience, and complexity of coding categories. *Child Development*, 1974, 45, 367-377.
- O'Leary, K. D., Kaufman, K. F., Kass, R. E., & Drabman, R. S. The effects of loud and soft reprimands on the behavior of children with emotional and conduct disorders. *Journal of Applied Behavior Analysis*, 1977, 10, 151-159.

- mands on the behavior of disruptive students. *Exceptional Children*, 1970, **37**, 145-155.
- O'Leary, K. D., & Kent, R. N. Behavior modification for social action: Research tactics and problems. In L. A. Hammerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts, and practice*. Champaign, Ill.: Research Press, 1973.
- O'Leary, K. D., Kent, R. N., & Kanowitz, J. Shaping data congruent with experimental hypotheses. *Journal of Applied Behavior Analysis*, 1975, **8**, 43-51.
- Reid, J. B. Reliability assessment of observation data: A possible methodological problem. *Child Development*, 1970, **41**, 1143-1150.
- Repp, A. C., Dietz, D. E. D., Boles, S. M., Dietz, S. M., & Repp, C. F. Differences among common methods for calculating interobserver agreement. *Journal of Applied Behavior Analysis*, 1976, **9**, 109-113.
- Repp, A. C., Roberts, D. M., Slack, D. J., Repp, C. F., & Berkler, M. S. A comparison of frequency, interval, and time-sampling methods of data collection. *Journal of Applied Behavior Analysis*, 1976, **9**, 501-508.
- Romanczyk, R. G., Kent, R. N., Diament, C., & O'Leary, K. D. Measuring the reliability of observational data: A reactive process. *Journal of Applied Behavior Analysis*, 1973, **6**, 175-184.
- Shuller, D. Y., & McNamara, J. R. Expectancy factors in behavioral observation. *Behavior Therapy*, 1976, **7**, 519-527.
- Taplin, P. S., & Reid, J. B. Effects of instructional set and experimenter influence on observer reliability. *Child Development*, 1973, **44**, 547-554.
- Wildman, B. G., Erickson, M. T., & Kent, R. N. The effect of two training procedures on observer agreement and variability of behavior ratings. *Child Development*, 1975, **46**, 520-524.

Received March 2, 1979

Final acceptance May 27, 1981