OBSERVER RELIABILITY AS A FUNCTION OF CIRCUMSTANCES OF ASSESSMENT¹

RONALD N. KENT, JAY KANOWITZ, K. DANIEL O'LEARY, AND
MICHAEL CHEIKEN

STATE UNIVERSITY OF NEW YORK AT STONY BROOK

Three factors characteristic of experimental settings were hypothesized to inflate artificially the reliability of observational recordings: (a) knowledge by observers of when and by whom their reliability is being assessed, (b) the absence of the experimenter or a monitor to prevent cheating, and (c) computation of reliability within- (*versus* between-) observer group. Three groups of four observers used a standard nine-category observational code for disruptive behavior in recording from videotapes of a classroom for 22 days. Analyses revealed considerable increases in average occurrence reliability as a function of the main effects of each of the experimental factors. The specific increases in reliability associated with each of the 12 combinations of the experimental factors are presented for each category of behavior. The possible role of observer-training procedures and behavioral definitions as determiners of nonartifactual reliability is discussed.

DESCRIPTORS: classroom behavior, disruptive behavior, behavioral recording, observational recording, methodology, reliability, observer bias, overt *versus* covert assessment, students

When observational recording procedures are employed, frequent evaluations of the similarity of simultaneous judgments are obtained from pairs of observers. These samples of observer agreement are referred to as "reliability", and provide the basis for assertions that behavioral recordings are the replicable product of well-specified recording procedures, rather than idiosyncratic judgments. Typically, reliability is assessed under a set of advantageous circumstances. The observer not only knows when (s)he is being evaluated but also has experience from training or from previous reliability assessments in com-

paring recordings with those of the assessor. Recent research indicates that obtained reliabilities may be enhanced both by overt reliability checks (Kent, O'Leary, Diament, and Dietz, 1974; Reid, 1970; Romanczyk, Kent, Diament, and O'Leary, 1973) and by knowledge of the particular reliability assessor (Romanczyk *et al.*, 1973). In addition, when several groups of observers are involved in a project and the observers serve as assessors for each other, reliability is routinely computed between observers in the same group, rather than between observers in different groups. If group differences in recording emerge over time, reliability computed within groups will be systematically higher than measures computed between observer groups (Kent, Note 3; Kent *et al.*, 1974; O'Leary and Kent, 1973, pp. 69-96). Finally, reliability is frequently assessed in the absence of experimenter or another monitor which may enhance reported reliabilities as a function of interobserver influence or cheating (O'Leary and Kent, 1973, pp. 69-96).

These characteristics of reliability assessment in field-experimental settings—knowledge of as-

¹This research was supported by Grant MH 21813 from the National Institute of Mental Health and Grant OEG-0-71-2872-607 from the Office of Education. The authors are indebted to Sharon Foster and Joyce Sprafkin for their editorial comments, and to Ronald Prinz for his assistance with data analysis. Portions of this paper were presented at the Western Psychological Association Conference in San Francisco, May 1974. Reprints may be obtained from K. Daniel O'Leary, Department of Psychology, State University of New York, Stony Brook, New York 11794. Ronald N. Kent is now at Brown University, Providence, Rhode Island, and Jay Kanowitz is now at the University of Texas, Austin, Texas.

assessment and assessor, absence of the experimenter, and calculation of reliability within groups—have been investigated in scattered studies differing in many respects such as the particular code employed and the nature of stimulus material. To date, there has been no investigation of these biases (in reliability assessment) in combination. Biases in reliability estimation associated with these factors in combination may be greater than indicated by evaluations of each singly. Because reliability assessment in many studies is characterized by two or more of these factors, it would be useful to evaluate the magnitude of potential bias associated with these factors separately and in combination. The present research assessed the effects of these factors using well-trained observers in a single study over an extended period of time.

METHOD

As part of an undergraduate course in psychological research, 12 students were trained in the use of a nine-category observation code for disruptive classroom behavior, which has been repeatedly employed by O'Leary and his associates (*cf.* O'Leary, Kaufman, Kass, and Drabman, 1970) and which is similar to a variety of classroom codes utilized by other investigators (*e.g.*, Becker, Madsen, Arnold, and Thomas, 1967; Hall, Lund, and Jackson, 1968; Walker, Mattson, and Buckley, 1971, pp. 236-263). The categories of this code were: out of chair, modified out of chair, touching, vocalization, playing, orienting, noise, time off task (not attending), and aggression.² Observers monitored the occurrence of each category during 20-sec observation intervals, separated by 10-sec recording intervals. At the end of each 20-sec viewing interval, observers tallied either the occurrence of one or more categories of disruptive behavior, or an absence of disruptions. The use of an "absence" recording was intended to mark intervals that

had passed and to reduce cues to other observers that a disruptive behavior had or had not been recorded. Material for observation was provided by videotapes of a class for behavior problem children. During training, observers reached an overall level of reliability with each other of 0.70. An occurrence reliability was employed in which the number of agreements between two observers on the occurrence of a category of disruptive behavior during respective observation intervals was divided by the sum of agreements and disagreements.

After training, the 12 observers were randomly divided into three groups of four members each and told that all future computation of reliability and discussion of disagreements in recording would occur only within the individual groups. It was explained that the purpose of their observations was to obtain a pool of ratings of a large set of videotapes to be used as the basis for future training of observers. The importance of maintaining accurate recordings at all times was stressed. Further, copying of another observer's recordings was emphatically discouraged. However, throughout training and the experimental sessions no pay was involved and only the usual social contingencies for high reliability were in effect.

Data were collected for 22 days during an eight-week period. Within each two-day period, the observers viewed six, 12.5-min videotape segments under six observation conditions (see Table 1). In conditions A and B, observers were not told that their reliability would be assessed. Recordings in these conditions were collected without comment at the end of the videotape segment. In conditions C and D, observers were told before viewing the videotape that their reliability would be assessed. It was not until recordings of that segment had been completed, however, that the experimenter selected one observer from each group to serve as assessor for that group, and the reliability of each group member with the respective assessor was computed. In conditions E and F, observers were told that their reliability would be assessed, and

²The code is a nine-category code; however, after the data collection, "aggression" was excluded as a category for analysis because it yielded frequencies too low for interpretation.

Table 1
Conditions of Reliability Assessment

		<i>Not Told of Assessment Assessor Not Identified</i>	<i>Told of Assessment Assessor Not Identified</i>	<i>Told of Assessment Assessor Identified</i>
		A	C	E
Presence of Experimenter during Observations	Present			
	Absent	B	D	F

an assessor was selected from each group before observation. In conditions A, C, and E, the experimenter remained present during the observation; during conditions B, D, and F, the experimenter left the room after starting the videotape.

An experimental condition was in effect for all three groups of observers simultaneously. During the 22 days of the study, there were 11 presentations of each of the six experimental conditions. During each successive two-day block, all six experimental conditions were presented in an order systematically varied according to a balanced Latin Square design.

The videotape segments selected for observation during the experimental sessions were matched so that the level of disruptive behavior during each successive two-day period was essentially equivalent. These tapes, which had not been viewed by observers previously, were then randomly assigned to the respective two-day periods by the *apriori* Latin Square design, as described above. All 12 observers viewed each videotape simultaneously, each group from its respective row of chairs in the room. When a reliability assessment had been announced, each group of observers adjourned immediately after that segment to one of three rooms provided for the respective groups. Here they computed the reliability of the other three group members with the assessor, and disagreements in recording were discussed. Selection of reliability assessor was systematically varied according to a Latin Square design. To discourage changes in recorded behavior during reliability computation, all recordings were done in ink, and summations of observed frequencies were collected by the ex-

perimenter before observers adjourned to compute reliability.

After all observations had been collected, reliabilities, which served as the basis for all data analyses, were calculated by computer for recordings collected under the six experimental conditions shown in Table 1. Reliabilities were computed between individual observers both within group and between group. The recordings of a particular observer for a single videotape segment were utilized in only one reliability computation. The analyses systematically counterbalanced observer pairs selected for reliability computation so that (a) each observer was paired with each other individual in the group an equal number of times; (b) each individual was paired with each individual in the two other groups an equal number of times, and (c) each observer contributed equally to reliability coefficients representative of each of the experimental conditions. In all, 33 reliability coefficients were generated for each of the 12 experimental conditions (see Table 3).

RESULTS

The data presentation must be limited to descriptive statistics, rather than inferential statistics which are complicated by such factors as the nonnormal distribution of reliability coefficients and complex interdependencies in the data matrix resulting from multiple computations of reliability among the various possible pairs of observers. Table 2 presents the average increase in reliability produced by the main effects of each of the experimental conditions for each of the

Table 2

Inflation of reliability coefficients as a function of the main effects.

Category	Knowledge of Assess- ment and Assessor- No Knowl- edge	Absence of Experi- menter- Presence	Within Group Compu- tation between Group
Out of Chair	0.08	− 0.04	0.11
Modified			
Out of Chair	0.06	− 0.02	0.05
Touching	0.41	0.11	0.11
Vocalization	0.05	0.01	0.22
Playing	0.02	0.02	0.09
Orienting	0.03	0.06	0.04
Noise	0.07	0.05	0.15
Time Off Task	0.14	0.09	0.22
Composite			
(weighted mean)	0.06	0.02	0.11
Category average			
(weighted mean)	0.11	0.04	0.12

eight observational categories, and for a composite representing the total of the other categories, weighted by frequency of occurrence.

Table 3 presents the average reliability coefficients produced by all 12 combinations of the experimental factors. That combination of factors predicted to produce the highest reliability—knowledge of assessment and assessor, absence of experimenter, and computation of reliabilities within observer group—is shown in the first column. These reliabilities are compared with that combination of factors predicted to produce the lowest reliabilities in the last column. Modified out-of-chair and playing were least affected by the experimental factors, showing increases of three and nine percentage points respectively. In contrast, touching was inflated 66 points, increasing a reliability of 0.10 to a reliability of 0.76. Similarly, time off task was increased 57 points, from a reliability of 0.27 without artifacts to a reliability of 0.84. On average, the experimental factors in combination increased reliability 29 points, from 0.50 to a respectable 0.79.

Of the three experimental factors, computation of reliability within-*versus* between-observer

groups was predicted to produce increasing differences in obtained reliability coefficients over time. Figure 1 presents the reliabilities obtained under these two circumstances, averaged across all other experimental factors, for a composite of the eight categories during the 11 consecutive two-day blocks of the study. During the first block, the difference in reliability was 0.05. However, during the 10 subsequent blocks, average within-group reliability increased substantially, from 0.64 to 0.81, while average between-group reliability increased only slightly, from 0.59 to 0.65. Thus, during block 11, there was a 0.16 discrepancy between reliability coefficients obtained within *versus* between groups of observers.

As noted by Kass (Note 2), occurrence reliability produces a numerically lower coefficient for low-frequency behavior than other common measures. For this reason, the effects of the experimental factors in combination were evaluated using three other popular measures of reliability, *i.e.*, phi coefficient, occurrence-nonoccurrence, and relative frequency. Phi, the four-fold correlation coefficient was computed utilizing a formula provided by Hays (1963, pp. 604-606). Occurrence-nonoccurrence reliability was calculated by dividing the number of agreements on the occurrence *and* nonoccurrence of each behavior in each interval by agreements plus disagreements. Relative frequency reliability was computed on the basis of the total frequency of each behavior recorded by dividing the smaller by the larger frequency. These three measures are compared with occurrence reliability for each category of behavior in Table 4.

All four measures, when inflated by the experimental factors in combination, reflect levels of reliability that would be acceptable in the current experimental literature. Three of the four measures, when inflationary factors are removed, provide reliability coefficients that are at best marginal, reflecting decrements of 28, 31, and 26 points, respectively. Only occurrence-nonoccurrence reliability, a measure particularly inflated by chance agreement among observers (Hart-

Table 3
Reliability Coefficients as a Function of the Experimental Factors in Combination

Category	Assessment Known Assessor Known				Assessment Known Assessor Unknown				Assessment Unknown Assessor Unknown				Inflation*
	E-Absent		E-Present		E-Absent		E-Present		E-Absent		E-Present		
	Within Group	Between Group	Within Group	Between Group	Within Group	Between Group	Within Group	Between Group	Within Group	Between Group	Within Group	Between Group	
Out of Chair Modified	0.98	0.69	0.84	0.81	0.77	0.69	0.83	0.82	0.77	0.70	0.84	0.69	0.29
Out of Chair	0.68	0.68	0.81	0.70	0.75	0.69	0.66	0.70	0.71	0.57	0.70	0.65	0.03
Touching	0.76	0.65	0.80	0.75	0.50	0.31	0.29	0.30	0.54	0.42	0.27	0.10	0.66
Vocalization	0.72	0.37	0.63	0.30	0.58	0.35	0.55	0.41	0.53	0.34	0.51	0.46	0.26
Playing	0.71	0.61	0.71	0.58	0.74	0.63	0.69	0.64	0.70	0.59	0.63	0.62	0.09
Orienting	0.88	0.76	0.79	0.73	0.83	0.81	0.78	0.79	0.82	0.80	0.74	0.70	0.18
Noise	0.75	0.53	0.68	0.53	0.75	0.59	0.60	0.50	0.65	0.46	0.58	0.52	0.23
Time off Task	0.84	0.36	0.53	0.44	0.62	0.33	0.61	0.36	0.55	0.43	0.37	0.27	0.57
Composite (weighted mean)	0.79	0.61	0.75	0.62	0.76	0.63	0.72	0.65	0.70	0.59	0.64	0.61	0.18
Category Average (unweighted mean)	0.79	0.58	0.72	0.61	0.69	0.55	0.63	0.57	0.66	0.54	0.58	0.50	0.29

*Inflation was computed by subtracting reliability coefficients in the extreme right column (representing the most rigorous assessment) from those in the extreme left column (representing the presence of all three potential sources of bias).

Fig. 1. Within-group *versus* between-group reliabilities during 11 consecutive blocks of recording.

Table 4

A comparison of the effects of the experimental factors in combination as reflected by four methods of reliability computation.

Observation Categories	Occurrence Reliability		Phi-Coefficient		Occurrence- Nonoccurrence Reliability		Relative Frequency Reliability	
	Inflated	Unin- flated	Inflated	Unin- flated	Inflated	Unin- flated	Inflated	Unin- flated
Out of Chair	0.98	0.69	0.98	0.84	1.00	0.97	0.99	0.73
Modified Out of Chair	0.68	0.65	0.86	0.80	0.98	0.97	0.77	0.68
Touching Others'								
Property	0.76	0.10	0.94	0.48	0.99	0.98	0.76	0.20
Vocalization	0.72	0.46	0.78	0.61	0.94	0.89	0.78	0.53
Playing	0.71	0.62	0.77	0.55	0.91	0.82	0.76	0.80
Orienting	0.88	0.70	0.91	0.77	0.96	0.90	0.89	0.74
Noise	0.75	0.52	0.77	0.55	0.91	0.76	0.83	0.64
Time Off Task	0.84	0.27	0.87	0.57	0.98	0.92	0.90	0.33
Composite Score								
(weighted mean)	0.79	0.61	0.86*	0.71	0.96	0.91	0.91	0.82
Category Average								
(unweighted mean)	0.79	0.50	0.86	0.65	0.96	0.90	0.84	0.58

*Weighted mean
Computed by Fisher r to Z transformation, $r = 0.89$

mann, Note 1), fails to reflect a substantial effect due to the experimental factors. Interestingly, of 48 comparisons with other measures in Table 4, occurrence reliability is consistently lower numerically on all but three occasions.

DISCUSSION

These data indicate that high reliabilities obtained under typical conditions of assessment may be extremely misleading. If reliability estimates are to be considered representative of data collected in the absence of reliability checks, reliability must be assessed under a set of circumstances not common in current research. Assessment must be performed either covertly or continuously and supervised by the experimenter or an agent who is unlikely to allow cheating. Further, if more than one group of observers is employed at different times or in different settings, reliability coefficients must be computed between these groups, preferably in the actual settings and circumstances of interest.

In the absence of high reliabilities obtained under the rigorous conditions outlined above, observational recordings cannot be assumed to reflect replicable applications of the behavioral code. In the absence of high reliability between and within observer groups, the design of a study must not confound observers with the experimental conditions. The systematic rotation of all observers across all experimental conditions will ensure that differences in recording between groups of observers do not exaggerate or mask actual treatment effects. Further, frequent "recalibration" with a single person or set of training tapes may reduce the magnitude of between-group differences in recording. Finally, in the absence of consistent reliability of observers with a stable criterion throughout data collection, recordings at different points in time may not be comparable.

The present findings replicate previous investigations and represent a solid basis for suggesting a thorough review of observer training and observation-code development. Much observer

training relies heavily on the development of consensus with a trainer or among the members or a group of observers as a basis for reliability. As such, it is not surprising that much of what is learned involves idiosyncratic interpretations of the vagaries of category definition. A remedy may well require the development of detailed behavior definitions in the form of manuals and videotapes with stable criterion protocols which would provide less variable feedback during training than an individual trainer or a group of inexperienced colleagues. Focus on the *definition* of a behavior, rather than multiple and variable interpretations, may help resolve the difficulties documented here.

A second set of considerations involves the characteristics of behaviors and behavior definitions. Although clear and substantial increases in reliability as a function of bias in assessment were evident on most categories, others were virtually unaffected. Such differences among the categories may be due, in part, to the characteristics of the behavior included. For example, it is interesting to note that the two categories showing the greatest differences in within-group *versus* between-group computation were the two auditory categories: noise and vocalization. Additionally, it seems likely that characteristics of the category definitions, such as specificity or complexity of the operational definitions, may be important determinants of the nonartifactual reliability of a category. The categories of behavior investigated here are highly similar to those employed by many developmental, educational, and clinical psychologists. These measures have been widely utilized without being studied or understood themselves. It seems necessary to investigate the dimensions along which observational categories vary and to determine empirically what characteristics are associated with data of uniformly acceptable quality.

REFERENCE NOTES

1. Hartmann, D. P. *Chance agreement in computation of inter-observer reliability*. Paper presented at symposium *The quality of observational data*,

- Western Psychological Association Conference, San Francisco, 1974.
2. Kass, R. E. *The effects of observer bias in field-experimental settings*. Unpublished master's thesis, State University of New York at Stony Brook, 1972.
 3. Kent, R. N. *Knowledge of predicted results as a factor in observational recording of behavior*. Unpublished doctoral dissertation, State University of New York at Stony Brook, 1972.
- ### REFERENCES
- Becker, W. C., Madsen, C. H., Arnold, C. R., and Thomas, D. R. The contingent use of teacher attention and praise in reducing classroom behavior problems. *Journal of Special Education*, 1967, 1, 287-307.
- Hall, R., Lund, D., and Jackson, D. Effects of teacher attention on study behavior. *Journal of Applied Behavior Analysis*, 1968, 1, 1-13.
- Hays, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.
- Kent, R. N., O'Leary, K. D., Diamant, C., and Dietz, A. Expectation biases in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology*, 1974, 42, 774-780.
- O'Leary, K. D., Kaufman, K. F., Kass, R., and Drabman, R. The effects of loud and soft reprimands on the behavior of disruptive students. *Exceptional Children*, 1970, 37, 145-155.
- O'Leary, K. D. and Kent, R. N. Behavior modification for social action: Research tactics and problems. In L. A. Hamerlynck, L. C. Handy, and E. J. Mash (Eds), *Behavior change: methodology, concepts, and practice*. Champaign, Illinois: Research Press, 1973. Pp. 69-96.
- Reid, J. B. Reliability assessment of observation data: A possible methodological problem. *Child Development*, 1970, 41, 1143-1150.
- Romanczyk, R. G., Kent, R. N., Diamant, C., and O'Leary, K. D. Measuring the reliability of observational data: a reactive process. *Journal of Applied Behavior Analysis*, 1973, 6, 175-184.
- Walker, H. M., Mattson, R. H., and Buckley, N. K. The functional analysis of behavior within an experimental class setting. In W. C. Becker (Ed), *An empirical basis for change in education*. Chicago: Science Research Associates, 1971. Pp. 236-263.

Received 27 February 1975.

(Final acceptance 16 July 1976.)