JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1977, 10, 141-150 NUMBER 1 (SPRING) 1977

ARTIFACT, BIAS, AND COMPLEXITY OF ASSESSMENT: THE ABCs OF RELIABILITY

ALAN E. KAZDIN[1]

THE PENNSYLVANIA STATE UNIVERSITY

Interobserver agreement (also referred to here as "reliability") is influenced by diverse sources of artifact, bias, and complexity of the assessment procedures. The literature on reliability assessment frequently has focused on the different methods of computing reliability and the circumstances under which these methods are appropriate. Yet, the credence accorded estimates of interobserver agreement, computed by any method, presupposes eliminating sources of bias that can spuriously affect agreement. The present paper reviews evidence pertaining to various sources of artifact and bias, as well as characteristics of assessment that influence interpretation of interobserver agreement. These include reactivity of reliability assessment, observer drift, complexity of response codes and behavioral observations, observer expectancies and feedback, and others. Recommendations are provided for eliminating or minimizing the influence of these factors from interobserver agreement.

DESCRIPTORS: methodology, observational procedures, observational code, observer bias, expectancies, feedback, reliability, artifact

A major feature of applied behavior analysis is the assessment of a client's overt behavior. The behavioral measures used are not usually standardized in the sense of traditional psychometric assessment; hence, one cannot rely on the consistency with which observations are made based on the assessment device itself, given uniform conditions of administration. Viscissitudes of defining target behaviors, the nature of applied settings, and conditions of observation require demonstration that behaviors are consistently recorded separately in each project. The well-known concern for consistency and accuracy of observations is expressed in the notion of "reliability" in applied behavior analysis. Reliability, as usually employed, refers to agreement between observers who independently score the same behavior of a subject. If the two observers consistently show relatively high agreement, it is assumed that the observations reflect the subject's performance relatively accurately.

Although accuracy of observations often is inferred from interobserver agreement, accuracy and agreement are not the same (*cf.* Bijou, Peterson, and Ault, 1968; Johnson and Bolstad, 1973). Accuracy usually refers to the extent to which observations scored by an observer match those of a predetermined standard for the same data. The standard is determined by other observers who reach a consensus about the data or by constructing observational material, such as videotapes or audiotapes, with predetermined behavioral samples (*e.g.,* Mash and McElwee, 1974). Interobserver agreement reflects the extent to which observers agree on scoring behavior. Usually, there is no firm basis to conclude that the one observer's data should serve as the standard, *i.e.,* is accurate.

As usually discussed, accuracy and interobserver agreement both involve comparing the

[1]The author is grateful for the comments provided by Eric J. Mash and John B. Reid on an earlier version of the manuscript. Reprints may be obtained from the author, Department of Psychology, The Pennsylvania State University, University Park, Pennsylvania 16802.

observer's data with some other source. They differ only in the extent to which the source of comparison can be entrusted to reflect the actual behavior of the subject.[2] Although accuracy and agreement are related, they need not go together. For example, an observer may observe accurately (relative to pre-established standard) but show low interobserver agreement (with another observer whose observations are quite inaccurate), or observe inaccurately (in relation to the standard) but show high interobserver agreement (with another observer who is inaccurate in an identical fashion).

Applied behavior analysis investigations usually assess interobserver agreement and assume that agreement reflects accuracy. Interobserver agreement is often considered adequate if it meets a prespecified level of agreement. Per cent agreement, one of the more commonly used measures, that reaches 70% or 80% often is considered satisfactory.[3] Yet, stressing the quantitative aspects of reliability ignores several assessment characteristics that dictate the meaningfulness of any agreement estimate. These include reactivity of reliability assessment, observer drift, complexity of the responses, information, expectancies, and feedback conveyed to the observers, and others. The present paper describes the characteristics and source of bias of reliability assessment, reviews the relevant research, and provides recommendations to minimize possible sources of bias.

[2]As with the term "reliability", "accuracy" probably is a misnomer in applied behavior analysis. The term implies that the standard of observation used as a criterion is correct and without error. Yet, consensual agreement often is used as the standard and may not reflect "true" performance.

[3]A quantitative criterion is difficult to invoke, in part, because the manner in which agreement should be defined and which of the available descriptive statistics for computing reliability should be used are unresolved (Hartmann, 1976; Hawkins and Dobes, 1975; Hawkins and Dotson, 1975.) The criterion for adequate agreement also depends on such diverse factors as the variability and rate of the observed behaviors, the number of different response codes scored, and the strength of the intervention.

Reactivity of Reliability Assessment

Nature of the problem. Interobserver agreement usually is checked periodically during an investigation. Typically, the observers are aware when their observations are checked, either because another observer enters the situation or because they observe the same client rather than different clients on these occasions. Research has shown that observer awareness of reliability assessment influences the observations.

Reid (1970) evaluated the effects of observer awareness that accuracy was being checked. Observers rated videotapes of parent-child interaction first when they were told that their observations were being compared to a standard and later when they were told that there was no standard. In fact, agreement with the standard was calculated for both sets of observations. Observers showed substantially higher accuracy when they thought that their observations were going to be checked. Indeed, when observers did not believe their observations were to be checked, their agreement with the criterion immediately dropped by approximately 25 percentage points. Although the results could have resulted from the order in which reactivity conditions were presented, subsequent studies have obtained similar findings.

Romanczyk, Kent, Diament, and O'Leary (1973) found that interobserver agreement was higher when observers believed that an individual observing with them was assessing reliability than when they believed the person was not assessing reliability. Differential accuracy of the data as a function of awareness also was suggested. Observers recorded 20% less disruptive student behavior when they were unaware, rather than aware, that interobserver agreement was assessed. Other studies also have demonstrated the influence of observer awareness that observations are being checked on accuracy and interobserver agreement (Kent, Kanowitz, O'Leary, and Cheiken, 1977; Kent, O'Leary, Diament, and Dietz, 1974; Taplin and Reid, 1973).

Awareness of whose observations serve as the standard for comparison also influences agreement. Romanczyk *et al.* (1973) found that an observer's performance was influenced by knowing who the other observer was during a reliability check. Two assessors whose observations were used as the standard were trained to score behaviors differently. For example, in scoring the category "vocalization" for a child, one assessor scored the softest vocalization possible, while the other scored only loud vocalizations. Observers were trained by and communicated with each assessor, thereby learning their idiosyncratic patterns. When the observers were checked with each assessor after training, they markedly shifted their scoring criteria.

Recommendations. The above research suggests that knowledge of reliability assessment and the identity of the reliability assessor affects interobserver agreement. Awareness of assessing agreement as a source of bias can be handled in several ways. The problem of observer awareness stems partially from conducting reliability checks under different conditions (reactive conditions) from those typically used to obtain the data (nonreactive conditions). This problem can be ameliorated in part by standardizing the conditions for reliability and nonreliability assessment. If observers believe that their behavior is not being monitored, these conditions should be maintained during reliability checks. Thus, reliability checks should be unobtrusive or covert. Alternatively, observers could be led to believe that all of their observations are being monitored. Indeed, this approach would appear advantageous because observers tend to be more accurate when they believe their agreement is assessed (Reid, 1970; Taplin and Reid, 1973).

It may be difficult to lead observers to believe that their behavior is always being checked. Covert reliability assessment may be needed. One suggestion for conducting covert reliability assessment is to have individuals score the behavior of different target subjects simultaneously in a group of subjects. In some of the intervals, the same subjects might be observed, although this would not be divulged to the observers. Comparisons of overlapping observations would provide an unobtrusive measure of reliability (O'Leary and Kent, 1973). In practice, these procedures may not be unobtrusive, due to interobserver communication or to events associated with the individual being observed. Observers may realize that they are assessing behavior of the same individuals simultaneously. Another solution is to have an experimenter covertly check reliability throughout the program, as for example, through a one-way mirror, although this may not be feasible in many naturalistic settings.

The problem of observer knowledge of identity of the reliability assessor may be resolved by controlling the communication of the assessor and observer so that they do not learn idiosyncracies of each other's recording. More elaborate solutions are available, such as conducting reliability checks from videotapes of select sections previously recorded by the observer. The assessor never has contact with the observer. Finally, several different assessors could be used so that an observer cannot readily learn the idiosyncratic patterns of any particular assessor.

Observer Drift

Nature of the problem. During training, observers usually receive extensive instruction and feedback regarding accuracy and interobserver agreement. Training is designed to ensure that observers adhere to the definitions of behavior and record behavior at a consistent level of accuracy. Once mastery is achieved, it is assumed that observers continue to apply the same definitions of behavior and record accurately. However, recent evidence suggests that observers "drift" from the original definitions of behavior (*e.g.,* Kent *et al.,* 1974, 1977; O'Leary and Kent, 1973; Reid, 1970; Reid and DeMaster, 1972; Taplin and Reid, 1973; Kent, Note 1). Drift refers to the tendency of observers to change the manner in which they apply the definitions of behavior over time.

Drift may not necessarily be reflected in interobserver agreement. If observers consistently work together and communicate, they may develop similar variations of the original response definitions (O'Leary and Kent, 1973). Thus, high levels of interobserver agreement can be maintained while accuracy has declined. In some reports, drift is revealed by comparing interobserver agreement within a given subgroup of observers who constantly work together with agreement across subgroups of observers who have not worked together (Kent *et al.*, 1974, 1977). Over time, subgroups may modify codes differently, which can be detected as differential within- and between-group interobserver agreement.

Modifications of the codes across observers may make observations from different observers incomparable. If subgroups of observers differ across experimental conditions, as might be the case for observations in a between-group design (*e.g.*, across different classrooms or homes), responses across groups cannot be meaningfully compared because they may not reflect common behavioral definitions. For within-subject designs, the data from a given set of observers or even for a single observer in one phase may not be directly comparable with data in earlier or later phases, due to observer drift.

Recommendations. Drift might be controlled by continually training all observers together as a unit throughout an investigation. Observers could periodically meet as a group, rate behavior, perhaps from videotapes, and receive immediate feedback on the accuracy of their observations relative to a predetermined standard. It is important to control drift by having access to observational data with an agreed upon standard. Otherwise, high levels of agreement might only reflect adjusting observations to meet the criteria of a familiar reliability assessor, rather than correctly applying the codes (Romanczyk *et al.*, 1973). Periodic retraining may limit the overall and differential loss of accuracy among observers. Of course, reactive retraining situations may delimit the generality of training so that behavior in the actual observation situation is not affected.

Drift might be controlled by videotaping the subject's behavior across sessions and by having observers score the tapes in a random order at the end of the study. Drift would not differentially influence data across phases. Unfortunately, taping sessions and observing behavior at the end of a project usually is time consuming and expensive. Also, ongoing data may be needed during the project to determine whether the experimental design or the intervention has to be altered in response to client behavior (Kazdin, 1977). Yet, taped samples of behavior could be compared with actual observations during select sessions partially to assess drift over time.

Drift might also be assessed or controlled by periodically bringing in newly trained observers to assess interobserver agreement during a project (O'Leary and Kent, 1973; Skindrud, 1973). Comparison of newly trained observers with observers who have continuously participated in the project can reveal whether the codes are applied differently over time. Differences between newly trained and experienced observers might simply reflect differences in the training procedures or in proficiency in applying the codes accurately, rather than modifications in applying the codes *per se*. Yet, any systematic alterations over time, including changes in proficiency, reflect observer drift.

Complexity of the Observational Coding System and Behaviors Scored

Nature of the problem. Complexity of the coding systems and behavior can refer to different characteristics of assessment. First, complexity can refer to the number of different response categories of an observational coding system. Systems with more categories are more complex than those with fewer categories. Second, complexity can refer to the number of different behaviors that are scored within a particular observational system on a given occasion. For a given observational system, more complex observations refer to those sessions in which a relatively

high proportion of different codes are scored relative to all of the codes available.[4]

The influence of complexity, defined as the number of response categories of an observational system, has been investigated by Mash and McElwee (1974), who trained observers to score dyadic taped verbal interaction using either four or eight response categories. Because the interactions were preprogrammed and known in advance, observer accuracy could be assessed. Observers using the four-category system showed a higher level of accuracy than did observers using the eight-category system. Thus, the number of response categories in an observational system influences observer accuracy.

Complexity, defined as the proportion of behavioral codes that are actually recorded in a given session, also has been shown to affect agreement. Taplin and Reid (1973) and Reid (Note 4) found that interobserver agreement and the number of different (nonrepeated) codes used were negatively correlated ($r = -0.52$, $r = -0.75$, respectively). Similarly, using two sets of data, Skindrud (Note 6) found that interobserver agreement was positively related with percentage of similar (repeated) behavioral interactions ($r = +0.53$ and $+0.65$). Thus, the greater the diversity of behaviors scored within a given observational system, the lower interobserver agreement.

As discussed earlier, interobserver agreement and accuracy are greater when two observers are aware that their observations are being checked. Interestingly, complexity of the observations within a given observational system seems to vary systematically with observer awareness of reliability assessment. Reid (Note 3) found that complexity of observations was lower during the sessions that reliability was checked than during adjacent unchecked observation sessions (*cf.* Jones, Reid, and Patterson, 1975). Apparently, when observers overtly participate in a reliability check, they may score behaviors more conservatively than when their observations are unmonitored. Given the negative correlation between complexity of observations and interobserver agreement or accuracy, the reduced complexity of observations during monitored reliability sessions may account for the increased interobserver agreement during these sessions.

Related research has examined the influence of the predictability of behavior on accuracy. Behaviors may be performed in sequence so that they are highly predictable (one behavior tends to be followed by other behaviors), or in no particular sequence so that they are unpredictable (one behavior is not necessarily followed by other behaviors). Mash and McElwee (1974) constructed different audiotapes with verbal interactions that differed in predictability. Generally, observing predictable *versus* unpredictable behavior did not lead to differential accuracy in scoring behavior during training. However, when observing new stimulus material, a history of observing predictable behavioral sequences led to decrements in observer accuracy, whereas a history of observing unpredictable sequences led to increments in accuracy. Thus, observers trained in a given situation where behavioral codes are scored in a relatively unpredictable sequence more readily adapt to new situations than individuals exposed to predictable behavioral sequences.

Mash and Makohoniuk (1975) replicated and extended the previous study and demonstrated that observers with a history of scoring predict-

[4]Complexity also might be defined as the number of individuals who are observed in a given study (analogous to the number of response codes) or as the proportion of individuals who are observed on a given day relative to all those possible (analogous to the proportion of response categories used). Thus, observing several individuals might be considered more complex than observing fewer individuals. And, scoring the behavior of relatively more individuals on a given day would constitute a more complex assessment. A differential number of individuals might be scored if the observation system focuses on a particular behavior in a group situation (*e.g.*, a class) independently of who performs that behavior. The proportion of individuals observed of the total number available might be a measure of complexity. Measures of complexity based on the number of subjects observed have not been evaluated in the context of reliability assessment.

able rather than unpredictable responses made more of perseverative errors when scoring new observational material. Also, providing subjects with an instructional set to see a pattern in the data, *i.e.,* by noting that certain response categories are likely to follow other response categories, led to lower recording accuracy than a set specifying no pattern in the data. Looking for a pattern increased the frequency of not scoring behaviors that occurred (*i.e.,* omission errors).

There are important implications for the influence of complexity and predictability for interpreting estimates of interobserver agreement. Initially, reliability estimates of a given percentage or correlation level must be viewed in relation to the complexity of the observation system. Agreement estimates for a given category within an observation system might be influenced by the number of other categories that are scored or can be scored.

Second, observations for a given observational system may vary in complexity and predictability over time. Categories for a given observation system may be differentially utilized over time. Indeed, a larger or smaller proportion of different codes may be systematically confounded with experimental conditions. For example, as the intervention begins to affect behavior, the number of different coded entries may decrease (*e.g.,* for disruptive behaviors) or increase (*e.g.,* for prosocial behaviors). In such cases, changes in frequency of several categories and the overall proportion of different categories used would be confounded with the presentation and withdrawal of the intervention. Thus, for a given observational system in a single experiment, interobserver agreement estimates of equal magnitude for a given behavior may not be equally meaningful across phases. Even if the same number of coded entries are used across phases, behaviors may be differentially predictable. The behavior of the subject is likely to become more predictable and, indeed, more homogeneous in general during the intervention when target responses are systematically consequated than during the nonintervention phases when consequences may be allowed to vary unsystematically.

The problems of complexity and predictability may apply to the specific subjects observed. Subjects in a given experiment may vary in the complexity of behavior (*i.e.,* the number of different data entries made). Interobserver agreement based on data from a particular subject can over- or underestimate the agreement obtained from observation of another subject (Reid, Skindrud, Taplin, and Jones, Note 5).

Recommendations. Specific recommendations cannot be made for each form of complexity. Certain assessment characteristics are dictated by the nature of the investigation. For example, the number of codes employed in an observational system usually is controlled by the client's behavior and goals of the project. Similarly, complexity of observations scored within a given system are controlled by the behavior of the client. The influence of complexity within an observational system on interobserver agreement can be controlled by assessing agreement across all phases of an investigation and across all subjects, or at least a large sample of subjects, to ensure that agreement is not confounded with complexity.

Because of the consistent relationship between complexity of the observations (*i.e.,* the proportion of different codes used for a given observational system), some investigators have proposed that interobserver agreement routinely take complexity into account (Reid *et al.,* Note 5). Specifically, these investigators proposed that percentage agreement and complexity (defined as the percentage of nonrepeated code entries) should be multiplied for a given reliability session. The resulting proportion provides a *proficiency score.* Use of this score protects against obtaining high levels of interobserver agreement due to a session of relatively low complexity.

The data on complexity have clear implications for observer training. Occasionally, observers repeatedly score the same stimulus material (*e.g.,* from videotapes) until a criterion level of agreement or accuracy is achieved. Then, they

are permitted to begin observations in the actual situation. During training, observers eventually may be able to predict the sequence of behaviors on the training stimuli. Accuracy or agreement obtained during training may overestimate post-training reliability when the observational samples are less familiar, more complex, and less predictable. The materials used in training observers should vary so that observations are not predictable. Also, because reliability and complexity of observations are related, high levels of interobserver agreement should be established for relatively complex observations for a given observational system. If complex observational stimuli are used during training, interobserver agreement is likely to estimate agreement conservatively during actual data collection, where complexity is allowed to vary.

Observer Expectancies and Feedback

Nature of the problem. Another potential source of bias is the expectancies of the observers regarding the subject's behavior and the feedback observers receive from the experimenter in relation to that behavior. Several studies suggest that observers who look for behavior change are more likely to find it (*e.g.,* Azrin, Holz, Ulrich, and Goldiamond, 1961; Scott, Burton, and Yarrow, 1967).

Recent investigations using behavioral assessment methods commonly employed in applied behavior analysis have examined observer expectancies. Kass and O'Leary (Note 1) told some observers that disruptive child behavior would increase and told others that it would decrease during treatment. All individuals observed the same classroom videotapes, which showed a decrease in disruptive behavior during treatment. In general, observers who expected a decrease recorded a greater reduction in some disruptive behaviors than those who expected an increase. Unfortunately, differential observer drift across groups, evident even in baseline, could have accounted for the results.

Kent *et al.* (1974) told some observers that disruptive behavior would decrease and told others that it would not change from baseline. The data on videotape in fact showed no change in disruptive behavior across phases. Overall, expectancies did not influence observer recordings. But when observers were asked to characterize the effect of the program on a questionnaire, their evaluation reflected the expectancy of the experimenter. Similarly, Skindrud (Note 6) found that informing observers of the experimental treatments did not bias the results of behavioral observations. Also, Redfield and Paul (1976) found that behaviors expected to change by observers were not influenced by these expectations on observational data. Overall, these results suggest that behavioral observations are not readily altered by observer expectancies.

Expectancies combined with feedback from the experimenter can influence observer performance. O'Leary, Kent, and Kanowitz (1975) led observers to believe that a token economy (treatment) would alter disruptive behavior on videotapes of children in a classroom. Actually, tapes of baseline and treatment were matched for disruptive behavior and no treatment was given. The experimenter provided positive comments (approval) of the observers' data if a reduction in the target behaviors was scored during the "treatment" phase, and negative comments (disappointment) if no change or an increase in the target behaviors was scored. Instructions to expect change and feedback for scoring reductions in target behaviors biased the data. Interestingly, child responses that observers were told would not change did not change during the experiment. Thus, expectancies and feedback about the effect of treatment exerted specific effects on the data.

Recommendations. The above research suggests that expectancies alone are not likely to influence behavioral observations unless some feedback also is provided. Presumably, feedback may be given by the experimenter or even be obtained by the observers from the data they are collecting. Thus, controlling expectancies and feedback may be difficult. Observers can readily detect interventions that require change in the

environment (*e.g.*, delivery of tokens, use of timeout) and are alerted to the desired therapeutic effects. Observer expectancies for change might be controlled by periodically bringing in new observers who are unfamiliar with the reinforcement history of the client or behavior change that has been achieved.

Another solution might be to videotape samples of performance throughout phases of the experiment. Ratings of the tapes in random order could provide a standard against which observations used in the study could be compared. Observer accuracy could be assessed to determine whether observers in the actual situation and familiar with the clients and interventions systematically differed in their observations across phases.

The problems of providing observers with explicit feedback are somewhat more easily controlled than expectancies. Obviously, experimenters should not and probably do not usually provide feedback to observers for directional change in client behavior. Any feedback given to observers should be restricted to the accuracy of their observations, rather than for changes in the client's behavior.

Additional Influences on Reliability Assessment

The above factors do not necessarily exhaust the possible procedural influences that need to be considered when interpreting estimates of accuracy and interobserver agreement. Other variables that might influence interobserver agreement have been explored. For example, Taplin and Reid (1973) attempted to show that observer accuracy is partially determined by the status of the experimenter. Observers trained with a high-status experimenter (university professor) showed lower accuracy estimates than did observers trained by lower-status experimenters (graduate students). Regrettably, only one experimenter served in the high-status condition and, thus, individual experimenter characteristics were confounded with status. Yet, research on experimenter characteristics appears to warrant additional scrutiny.

The individuals who compute interobserver agreement may influence reliability estimates. For example, Kent *et al.* (1974) found that observer agreement tended to be higher when computed by observers than by the experimenter. Similarly, O'Leary and Kent (1973) found that higher estimates of interobserver agreement were obtained when observers were allowed to score behavior and calculate reliability without, rather than with, the supervision of an experimenter. Although calculation of data by individuals who participate in a project is not necessarily biased (Rusch, Walker, and Greenwood, 1975), as a precaution, those who compute the data should not be the same individuals as those who calculate and evaluate the data.

CONCLUSION

The above overview described major characteristics and sources of bias that need to be considered when evaluating reliability estimates. Essentially, the characteristics describe only some of the major conditions that may influence the interpretation of reliability. Interobserver agreement and accuracy can be viewed as target behaviors in their own right that are a function of a variety of variables. These include characteristics of the observational system, characteristics of the experimenter, observer, and client, methods of scoring behavior, the nature and duration of observing training, situational and instructional variables during assessment of reliability, the pattern of client behavior, concurrent observation of stimulus and consequent events, and so on. Generally, current research has only begun to evaluate these variables and supports the contention that agreement is multiply-determined.

Because the data obtained in a given investigation depend on diverse factors in addition to the specific responses of the client, some investigators (Jones *et al.*, 1974; Mash and McElwee, 1974) have advocated that the reliability of behavioral observations be conceptualized from the standpoint of generalizability theory (Cron-

bach, Gleser, Nanda, and Rajaratnam, 1972). Generalizability theory extends the notion of reliability so that generalizability of observations across different conditions within an investigation can be assessed. The extent to which observations in a study vary across facets or dimensions (*e.g.,* observers, occasions, phases, *etc.*) can be studied, and the generalizability of the data across different levels of these facets can be evaluated directly. An advantage of studying generalizability is that it simultaneously examines the contribution of diverse characteristics of assessment to the data. Also, the theory of generalizability emphasizes the relative nature of reliability, *viz.,* that there is no reliability for a given assessment method, but rather an infinite number of reliabilities that are a function of the range of assessment conditions.

While the research reviewed in the present paper strongly suggests that diverse sources of bias and characteristics of assessment influence reliability estimates, the generality of many of the specific conclusions must be made cautiously. Many of the investigations were laboratory analog studies and approach only some of the conditions present in naturalistic settings. For example, in some studies the duration of observer training was brief relative to the training used for many applied studies (Mash and McElwee, 1974; Reid, 1970); the observers were paid volunteers or subjects fulfilling experimental credits for a course and were not necessarily screened for their competence (Mash and McElwee, 1974; Taplin and Reid, 1973); also, the codes included multiple behaviors (*e.g.,* over 30 categories), rather than the few that are more commonly studied (Reid, 1970). Also, in a few studies, conditions are designed to maximize bias and artifact such as intentionally giving reliability assessors different behavioral definitions, permitting observers to calculate their agreement, and encouraging interobserver communication (Kent *et al.,* 1974; Romanczyk *et al.,* 1973). Yet, this area of research cannot be discounted on the grounds of frequent reliance upon analog studies for at least two reasons. First, some studies have employed observers trained for extended periods and have used observational codes evaluated in many applied investigations (*e.g.,* Romanczyk *et al.,* 1973). Second, while analog studies always raise questions about the generality of the findings, the consistency of the sources of bias revealed by the studies reviewed in the present paper presents a convincing demonstration of the importance of bias. In light of the specific characteristics and sources of bias associated with assessing interobserver agreement, any estimate of agreement must be qualified by the specific conditions of assessment. Research needs to establish the ideal conditions under which agreement can be assessed and the effects of deviation from these conditions in applied settings.

REFERENCE NOTES

1. Kass, R. E. and O'Leary, K. D. *The effects of observer bias in field-experimental settings.* Paper presented at symposium, Behavior Analysis in Education, University of Kansas, Lawrence, April 1970.
2. Kent, R. N. *Expectation bias in behavioral observation.* Unpublished doctoral dissertation. State University of New York at Stony Brook, 1972.
3. Reid, J. B. *Differences in the complexity of reliability assessment vs. adjacent non-reliability assessment observation sessions: A technical note.* Unpublished manuscript, University of Oregon, 1973. (*a*)
4. Reid, J. B. *The relationship between complexity of observer proctocol and inter-observer agreement for twenty-five reliability assessment sessions: A technical note.* Unpublished manuscript, University of Oregon, 1973. (*b*)
5. Reid, J. B., Skindrud, K. D., Taplin, P. S., and Jones, R. R. *The role of complexity in the collection and evaluation of observation data.* Paper presented at meeting of the American Psychological Association, Montreal, Quebec, September 1973.
6. Skindrud, K. *An evaluation of observer bias in experimental-field studies interaction.* Unpublished doctoral dissertation, University of Oregon, 1972.

REFERENCES

Azrin, N. H., Holz, W., Ulrich, R., and Goldiamond, I. The control of the content of conversation through reinforcement. *Journal of the Experimental Analysis of Behavior,* 1961, **4,** 25-30.

Bijou, S. W., Peterson, R. F., and Ault, M. H. A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. *Journal of Applied Behavior Analysis,* 1968, **1,** 175-191.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley, 1972.

Hartmann, D. P. Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis,* 1977, **10,** 103-116.

Hawkins, R. P. and Dobes, R. W. Behavioral definitions in applied behavioral analysis: Explicit or implicit. In B. C. Etzel, J. M. LeBlanc, and D. M. Baer (Eds.), *New developments in behavioral research: theory, methods, and applications. In honor of Sidney W. Bijou.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1975.

Hawkins, R. P. and Dotson, V. A. Reliability scores that delude: an Alice in Wonderland trip through the misleading characteristics of inter-observer agreement scores in interval recording. In E. Ramp and G. Semb (Eds.), *Behavior analysis: Areas of research and application.* Englewood Cliffs, New Jersey: Prentice-Hall, 1975. Pp. 359-376.

Johnson, S. M. and Bolstad, O. D. Methodological issues in naturalistic observation: some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, and E. J. Mash (Eds.), *Behavior change: methodology, concepts, and practice.* Champaign, Illinois: Research Press, 1973. Pp. 7-67.

Jones, R. R., Reid, J. B., and Patterson, G. R. Naturalist observation in clinical assessment. In P. McReynolds (Ed.), *Advances in psychological assessment,* Vol. 3. San Francisco: Jossey-Bass, 1975.

Kazdin, A. E. Methodology of applied behavior analysis. In T. A. Brigham and A. C. Catania (Eds.), *Handbook of applied behavior research: social and instructional processes.* New York: Irvington/Naiburg—Wiley, 1977, (*in press*).

Kent, R. N., Kanowitz, J., O'Leary, K. D., and Cheiken, M. Observer reliability as a function of circumstances of assessment. *Journal of Applied Behavior Analysis,* 1977, (*in press*).

Kent, R. N., O'Leary, K. D., Diament, C., and Dietz, A. Expectation biases in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology,* 1974, **42,** 774-780.

Mash, E. J. and Makohoniuk, G. The effects of prior information and behavioral predictability on observer accuracy. *Child Development,* 1975, **46,** 513-519.

Mash, E. J. and McElwee, J. Situational effects on observer accuracy: behavioral predictability, prior experience, and complexity of coding categories. *Child Development,* 1974, **45,** 367-377.

O'Leary, K. D. and Kent, R. N. Behavior modification for social action: research tactics and problems. In L. A. Hamerlynck, P. O. Davidson, and L. E. Acker (Eds.), *Critical issues in research and practice.* Champaign, Illinois: Research Press, 1973. Pp. 69-96.

O'Leary, K. D., Kent, R. N., and Kanowitz, J. Shaping data collection congruent with experimental hypotheses. *Journal of Applied Behavior Analysis,* 1975, **8,** 43-51.

Redfield, J. and Paul, G. L. Bias in behavioral observation as a function of observer familiarity with subjects and typicality of behavior. *Journal of Consulting and Clinical Psychology,* 1976, **44,** 156.

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

Reid, J. B. and DeMaster, B. The efficacy of the spot-check procedure in maintaining the reliability of data collected by observers in quasi-natural settings: two pilot studies. *Oregon Research Institute Research Bulletin,* 1972, **12.**

Romanczyk, R. G., Kent, R. N., Diament, C., and O'Leary, K. D. Measuring the reliability of observational data: a reactive process. *Journal of Applied Behavior Analysis,* 1973, **6,** 175-184.

Rusch, F. R., Walker, H. M., and Greenwood, C. R. Experimenter calculation errors: a potential factor affecting interpretation of results. *Journal of Applied Behavior Analysis,* 1975, **8,** 460.

Scott, P., Burton, R. V., and Yarrow, M. Social reinforcement under natural conditions. *Child Development,* 1967, **38,** 53-63.

Skindrud, K. Field evaluation of observer bias under overt and covert monitoring. In L. A. Hamerlynck, L. C. Handy, and E. J. Mash (Eds.), *Behavior change: methodology, concepts, and practice.* Champaign, Illinois: Research Press, 1973. Pp. 97-117.

Taplin, P. S. and Reid, J. B. Effects of instructional set and experimenter influence on observer reliability. *Child Development,* 1973, **44,** 547-554.

Received 6 February 1976.
(Final acceptance 15 May 1976.)