SWETLOW, LINDA KATHERINE. A Comparison of the Utility of Five Methods
of Time Sampling. (1976)
Directed by: Dr. Rosemery O. Nelson. Pp. 110.

The present study was concerned with assessing the reliability, representativeness, and utility of a sequential method of intermittent time sampling. This new approach to taking a time sampling of behaviors involved having one observer track two or four behaviors, one category of behavior per interval, with the particular behavior to be tracked during an interval varied systematically between intervals. This sequential approach to observation was compared to other methods of time sampling involving tracking one, two, or four categories of behavior per interval in a continuous fashion. The observation methods were compared in terms of reliability, representativeness, and practicality for use by the clinician.

The reliability of a particular observation method was assessed in terms of how closely two observers recording the same material using that method would agree on what had transpired during the recording session. Representativeness was assessed in terms of how well these observational records represented what had transpired during the observation period as represented by the frequencies of behavior generated by using the method. Practicality or utility was assessed in terms of how useful a particular method would be for the clinician, given its degree of reliability and representativeness.

Results of the present investigation provided substantial evidence that the sequential approach to observation provides a useful research tool for the clinician. This utility is a function of a combination of several advantages the sequential approach offers compared to other

approaches involving tracking behaviors in a continuous manner. Use of the sequential approach resulted in minimal training time required for an observer to be able to take reliable recordings of behavior. After completing training, high levels of reliability in data collection were demonstrated by the observers using the sequential method. Further, adequately representative samples of behavior were generated and reliable information about many categories of behavior during one data collection session were possible.

Given the high practicality offered by the use of the sequential method, this approach to observation should be very appealing for the clinician and researcher interested in collecting reliable, representative data with a minimum of effort and expenditure.

A COMPARISON OF THE UTILITY OF FIVE METHODS OF TIME SAMPLING

by

Linda Katherine Swetlow

A Thesis Submitted to
the Faculty of the Graduate School at
the University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Greensboro
1976

Approved by

Rosemery O. Nelson
Thesis Adviser

APPROVAL PAGE

This thesis has been approved by the following committee of the Faculty of the Graduate School at The University of North Carolina at Greensboro.

Thesis Adviser Rosemary O. Nelson

Committee Members Jacquelyn Gaebelein

Mary Fulcher Geis

May 4, 1976

Date of Acceptance by Committee

ACKNOWLEDGEMENTS

The author wishes to offer her gratitude to Dr. Rosemery O. Nelson for her dedicated guidance, editorial suggestions, and most important, her patience, in the preparation of this work. The benefit of her knowledge of the subtleties in the field of behavioral assessment was invaluable.

Grateful appreciation is extended to Dr. Jacqueline Gaebelein for her help in selecting and simplifying the design of this study and for serving as a member of the thesis committee. Appreciation is also extended to Dr. Mary Geis for serving as a committee member.

The author wishes to offer her thanks to Dr. Ronald N. Kent of the State University of New York at Stony Brook for graciously permitting copies of videotapes he prepared to be used in this study; Dr. William Powers for his help in the statistical analyses; and to the ten undergraduates who served as observers.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

APPENDICES

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Behavioral approaches to psychology view behavior in terms of interactions between the organism and his environment. Normal as well as abnormal behavior patterns are seen as a function of the organism's present environment interacting with his past learning history. Behaviors are elicited by certain antecedent environmental events and maintained or extinguished by other consequating environmental events. This notion forms the basis for the popular S - O - R - C (stimuli - organism variables - responses - consequences) model of behavior analysis (Goldfried & Sprafkin, 1974). Bandura (1969, p. 63) sees psychological functioning as involving "a reciprocal interaction between the behavior and its controlling environment." Behavior modification as a therapeutic orientation concentrates on the remediation of overt behaviors (Jones & Cobb, 1973; O'Leary & Kent, 1973) by applying the principles outlined by the S - O - R - C model to change the organism's behavior (Goldfried & Kent, 1972).

Behavioral approaches to assessment, therefore, stress that an individual's behavior is meaningful only when assessed in the appropriate environmental context, since behavioral frequency varies as a function of environmental changes (Patterson & Harris, 1968). Since the best indicator of future behavior is past performance in similar situations (Fulkerson & Barry, 1961), the goal of behavioral assessment is to predict human behavior by defining the parameters and situations governing the response, detailing behavior/environment interactions. This procedure gives the assessor a direct, non-inferential measure of the client's responses to relevant features of the environment, minimizing the need for interpretation.

The best way to provide a non-inferential basis for an assessment device is to use the sample approach to interpretation outlined by Goodenough (1949). Since the behaviors observed in an assessment situation are taken to be a representative sample of behaviors of interest, the most efficient way of obtaining such a sample is to maximize the similarity of the assessment response to the target behavior. Given the situation specificity of behavior (Mischel, 1968), the least inferential method of obtaining a sample of these target behaviors of interest would be a direct sampling of criterion responses in the naturalistic setting.

It has been pointed out many times that the taking of objective recordings of behavior in the natural and quasi-natural setting for both assessment and research purposes is one of the distinguishing characteristics of behavioral assessment (Eckman, 1973; Johnson & Bolstad, 1973; Kubany & Slogget, 1973; Lipinski & Nelson, 1974; Romancyzk, Kent, Diament, & O'Leary, 1973; Thomson, Holmberg, & Baer, 1975). Direct observation in situ would seem to be the ideal assessment device, since these observations provide a non-inferential, situation specific sample of behaviors. The observation period may be seen as a subset of the individual's interactions in similar situations, and the behaviors he emits during the observation period may be similarly considered typical of behaviors he would normally emit in that and similar stimulus situations (Johnson & Bolstad, 1973).

However, direct observations are far from being ideal assessment measures because of the many methodological problems associated with their use. These methodological problems limit the predictions about behavior one can make from the assessment session to non-observed situations. When direct observations are used as dependent measures in behavioral research, such methodological problems may impair the generalizability or external validity and the content or internal validity (Campbell & Stanley, 1966) of results. Lipinski and Nelson (1974) specify three major categories of methodological problems associated with the use of direct observations in the naturalistic setting as an assessment device: the reactive nature of "being observed", potential observer bias, and procedural problems in observation.

Reactivity of "Being Observed"

People tend to perform atypically on a task when they know their performance is being observed, as opposed to those times they are unaware of observation (Johnson & Bolstad, 1973). Observers function as stimuli which change the subject's environment (Patterson & Harris, 1968). Since the subject's environment has changed, one would expect a concurrent behavioral change (Bandura, 1969). These effects of the observer's presence on the subject's behavior may make it impossible to obtain a truly representative sample of the client's behavior; observed behavior in the naturalistic setting may not generalize to unobserved situations (Johnson & Bolstad, 1973). Such changes in behavior occurring as a function of the observation process itself are attributed to the reactive effects of "being observed" (Patterson & Harris, 1968).

Patterson and Harris (1968) were involved with studying the effects of observers on the behavior of subjects being observed in their own homes. They found that high interactors tended to decrease interaction rates after the baseline sessions, while low interactors increased rates slightly. Both groups tended to regress towards the mean of interaction frequencies over time. It is suggested that this change in behavior patterns may be attributable to a habituation to the observational process phenomenon. It is also mentioned that "observer effects" may be limited to certain classes of behaviors. Patterson and Harris suggest that differences in interaction levels could be related to the idea that being observed may function as a stimulus which elicits different behaviors than would be expected in the absence of that discriminative stimulus. Further, observing and recording someone's behavior may be perceived as aversive by the person being observed and may lead to an increase in escape and avoidance behaviors. This last interpretation was suggested by the fact that the family members being observed tended to spend a lot of the observation time in the bathroom or playing solitary games or reading.

In a later study, Patterson and Cobb (1971) analyzed the stability of each of the behavior interaction categories coded and found a stability in behavior rates over time. This suggests that the family did not habituate to the observers, at least during the time limits imposed by the 1968 study.

Earlier, Bechtel (1967) reached similar conclusions on the reactive effects of "being observed" that Patterson and Harris (1968) did. Subjects were requested to look at and rate pictures in order of preference in a room in an art museum. The results indicated that people in the group that knew they were being observed spent less time in the room and covered less floor space. Bechtel suggests that the subjects perceived the observations as aversive stimulus conditions. Leaving the room faster was one way of escaping this aversive situation.

White (1973) used activity level as defined by the distance subjects covered in a room as the dependent measure in a deception experiment designed to study the effects of the presence or absence of an observer in a quasi natural setting which allowed for concealed observation. White found that the knowledge of being observed reduced the activity level of families as opposed to conditions where the subjects were unaware that their behaviors were being monitored.

Polansky, Freeman, Horowitz, Irwin, Papanis, Rappaport, and Whaley (1949) overtly observed the disruptive behaviors of children in a summer camp for delinquents. They suggest that the many aggressive responses shown towards the observers were a function of resistance to being observed.

One additional study by Horton, Larson, and Maser (unpublished manuscript) may be interpreted as contributing to the evidence of the reactive effects of "being observed". A teacher, receiving special instruction in raising the frequency of her classroom approval behaviors, emitted higher rates of approval behaviors when she knew she was being observed.

These studies offer evidence that in the observer/observee kind of situation, the process of being observed can effect the behaviors of those being observed. The observer functions as a discriminative stimulus in whose presence certain behaviors become more likely, and others less likely. Several variables have been suggested as contributing to these reactive effects of "being observed".

One source that may contribute to the reactivity of "being observed" is the obtrusiveness of the observer. The more novel and conspicuous the observer's presence, the greater the distracting effect on the subject should be. Callahan and Alevizos (1973) studied the reactive effects of the observation process of the male and female patients on a chronic ward. Minimal behavioral differences were found to be related to the intensity and obtrusiveness of observations. Callahan and Alevizos point out that reactive effects are probably a function of the behavioral options available to those being observed.

Roberts and Renzaglia (1965) manipulated varying degrees of the obtrusiveness of assessment devices. Subjects made significantly more favorable comments about themselves when they knew their verbalizations were being recorded, than when the recording process was covert. Similarly, subjects made significantly more unfavorable comments about themselves in the covert as opposed to overt recording conditions. These results suggest that the obtrusiveness of detectable observation devices had a differential effect on verbal behavior.

The personal characteristics of the observer himself, age, sex, race, professional status, may also influence the behaviors emitted by the subject. For this reason, it has been suggested by O'Leary (O'Leary, Romancyzk, Kass, Deitz, & Santogrossi, 1971) that the observer should try to become as neutral a stimulus as possible in the subject's environment. Another factor that may influence the behavior of subjects is the rationale given for the observations. Also, some people are probably more prone to emitting reactive responses to the observation process. Individual differences and differing rationales may be seen to affect anxiety responses in the observation situation. Also, observers in an applied setting may transmit their expectancies to the subjects thereby altering subject behaviors such that they conform to the experimental hypothesis.

A study by Johnson and Lobitz (1974) illustrates that people can alter their behavior as a function of the demand characteristics of the situation. They found that parents can make their children appear more or less deviant when so instructed by the experimenter. The parents modified their own behaviors and interactions with their children in accordance with experimenter instructions.

Mash and Hedley (1975) studied the effect of an adult observer on a child's performance of a simple motor task. Mash and Hedley suggest that the observer functions as a discriminative stimulus; certain behaviors will be facilitated and others inhibited as a function of the observer's presence. The nature of this acquired discriminative property of the observer may in part be determined by history of social interaction.

Several ways of minimizing these reactive effects of "being observed" have been suggested, including using invisible monitoring devices, minimizing the obtrusiveness of the observers, allowing time for the subjects to habituate to the observer's presence, and finding the minimum number of observations to provide the necessary data and using that data sampling frequency (Callahan & Alevizos, 1973; Johnson & Bolstad, 1973). One way of minimizing observer expectancy effects is to employ naive or misinformed observers, or to use different observers in final phases of the study. However, this solution may introduce a design problem of confounding observers with treatments, discussed by O'Leary and Kent (1973).

In summary, the process of having one's behavior observed by others may alter ongoing behavior. These changes in behavior may be attributed to the reactive effects of "being observed" (Patterson & Harris, 1968). Several variables have been postulated to account for these reactive effects such as the obtrusiveness of the observer, individual subject differences, unique characteristics of the observer, rationale given for the observation, the expectancies of the observer which are communicated to the subject, or the demand characteristics of the situation (Johnson & Bolstad, 1973).

Observer Bias

O'Leary and Kent (1973) describe the human observer as a "faulty cumulative recorder". Errors in the recording of behaviors by observers are generally assumed to be distributed symmetrically in some random manner around what has actually transpired during the observation period. However, when such errors are distributed in an asymmetrical or unidirectional manner (generally consonant with the experimenter's hypothesis) these errors can no longer be considered random variations and are considered to be the result of observer bias (Johnson & Bolstad, 1973).

Systematic direct observations may lack validity insofar as these recordings may be a function of factors other than the subject's behaviors. The problem of observer bias poses a particularly serious threat for the validity of naturalistic observations. While other methodological problems associated with the use of observations such as the reactive effects of "being observed" and questions of observer agreement or accuracy (discussed in the next section on procedural problems) tend to remain constant or vary randomly between experimental conditions, observer effects or bias may interact and be confounded with treatment effects (O'Leary & Kent, 1973) since they tend to be asymmetrically aligned with experimental hypotheses (Johnson & Bolstad, 1973).

Some sources of biased recordings of behavior include knowledge of predicted results or expectancies and experimenter feedback (Lipinski & Nelson, 1974). The effect of the observer's expectancies on the subject's behaviors, as opposed to the effects of his expectancies on his recording behaviors, is another possible source of observer bias (Johnson & Bolstad, 1973) and has already been discussed in the previous section on the reactive nature of the observation process.

Observers' recordings and interpretations of behaviors may be modified by knowledge of expected results (Johnson & Bolstad, 1973). The results of several studies suggest that systematic behavioral recordings of relatively untrained observers are particularly susceptible to errors in recording and interpretation attributable to observer expectancies or instructional sets.

Scott, Burton, and Yarrow (1967) used two "groups" of observers to rate the pre-recorded verbal behaviours of a child interacting with classmates. Ratings of an observer familiar with predicted results were substantially more consonant with the experimental hypothesis than the ratings of the uninformed observers. Another study by Rapp (1966) used eight pairs of observers to observe children in a nursery school. Those observers in the "below par" group described the children in such a way that it sounded as if the child must in fact not be feeling well, while those observers in the "above par" group described the same children in accordance with the expectancy given them by the experimenter.

Here again, observer expectancies seemed to bias subjective reports of child behavior.

Other studies, employing highly trained observers using complex behavioral codes have tried to manipulate observer expectancies to see what effects different expectancies will have on systematic behavioral recordings.

Kass and O'Leary (1970) had three groups of undergraduates record from the same videotaped presentations of disruptive behaviors of children in the classroom setting, using a nine category behavioral code (O'Leary, Kaufman, Kass, & Drabman, 1970). The group that expected soft reprimands to increase disruptive behavior recorded a significantly smaller decrease in disruptive behavior from baseline to treatment than the observers who expected a decrease.

Skindrud (1972) unsuccessfully attempted to replicate the Kass and O'Leary (1970) results using a thirty category family interaction code (Patterson, Ray, Shaw, & Cobb, 1969). There was no significant difference in recordings between groups on the incidence of deviant child behaviors recorded arising from experimenter induced expectancies. These results may be interpreted as showing no evidence for observer bias arising as a function of expectancies concerning experimental outcome.

Using a similar design, Skindrud (1973) compared observations of observers who were informed about the "normal" or "deviant" status of the family being observed and whether the family was "in treatment" to reduce deviant behaviors to observations of the same family carried out by a "blind calibrating observer". This "blind" observer was uninformed about family status or treatment variables. The results showed no significant difference in the recordings of either type of observer, but informed observers recorded significantly higher rates of deviant behaviors than uninformed observers across all experimental conditions. Skindrud does not interpret this difference as observer bias; rather he discusses "the possibility that information about a study sensitizes observers to the variables involved..."

A final study by Kent, O'Leary, Diament, and Dietz (1974) was designed to assess the effects of expectancies as a source of observer bias while avoiding the methodological flaws in the Kass and O'Leary (1970) study in which experimental group was confounded with experimental teams. The observers were given differential and noncontingent feedback that observational data seemed to be consistent or not consistent with predicted results. The results indicated that expectancies did not bias objective behavioral recordings, but that the observers' global evaluations were in accordance with the expectancies they had been given by the experimenter.

Another source of potential observer bias stems from the effects of accuracy contingencies on the observer's data collection imposed by the experimenter. Two studies support the notion that observers' recording behaviors may be biased by evaluative feedback. O'Leary, Kent, and Kanowitz (1975) found that although knowledge of predicted results did not seem to bias observers' recordings, knowledge of results plus differential reinforcement of recordings of specific behavioral categories consistent with the expectancies observers had been given was sufficient to distort observational recordings. Another study by Romancyzk _et al._ (1973) also indicated that observers may adjust their applications of a behavioral code as a function of the feedback they receive from reliability checkers. Knowledge of which assessor inter-observer agreement levels were being computed with produced a shift in the observational criteria normally employed by the observer, to match the criteria used by the reliability checker.

Another problem associated with the contingencies imposed on observers to produce accurate recordings is the phenomenon of observer cheating. Observers are likely to perceive that the experimenter is interested in accurate recordings, reflected by high inter-observer agreement levels. Trying to keep the experimenter happy can lead to observer collusion or cheating to obtain high agreement levels (O'Leary & Kent, 1973).

Cheating may take the form of overt collusion wherein observers communicate with each other during the observation period in order to match up behavioral ratings. O'Leary and Kent (1973) mention one of their studies in which the experimenter supervised the observation period and computation of agreement coefficients during certain phases of the observer training program. During other phases of training, the experimenter was not present in the observation room. Average inter-observer agreement levels were significantly higher when observers were left unsupervised.

The differential coding of behavioral categories by observers to match the rating behaviors of an identified assessor employing a modified version of a behavioral code, discussed in the Romancyzk _et al_. (1973) study is another example of observer cheating to obtain higher reliabilities. Another example of observer cheating involves computational errors made when estimating inter-observer agreement levels. An interesting result of the Kent et al. (1974) study was that there was a significant difference in estimates of inter-observer agreement within observational groups, when these estimates were calculated by the experimenter as opposed to the observers. The observers consistently overestimated the accuracy of their observations.

Potential observer bias may be seen as a factor impairing the validity of observational data collected in the laboratory and naturalistic settings. Experimental data can be distorted if observer effects are confounded with treatment effects (O'Leary & Kent, 1973), for example, if different observers are used to record data from different experimental conditions. Two major sources of observer bias are expectancies and evaluative feedback from or accuracy contingencies imposed by the experimenter. These potential sources of bias may be seen as contributing to the phenomenon of observer drift. Individual members of groups of observers recording, computing agreement levels, and discussing differences in recording behaviors together will tend to modify their applications of the behavioral code to match the definitions used by other members of the group. These modifications tend to be random in nature and thus might not be seen as observer bias but rather consequences of observer bias (O'Leary & Kent, 1973).

Some procedures that have been suggested to minimize the confounding effects of observer bias include keeping observers uninformed about the experimental hypothesis (Johnson & Bolstad, 1973), having the experimenter compute levels of inter-observer agreement, and eliminating feedback to observers on the results of inter-observer agreement assessments (O'Leary & Kent, 1973). A final suggestion is using well defined behavioral codes which might inhibit interpretive bias (Johnson & Bolstad, 1973), which leads into the issue of the procedural problems involved in naturalistic observations.

Procedural Problems

Some procedural problems that can impair the validity of systematic observations in the naturalistic or laboratory setting include behavioral information lost as a consequence of the use of observational codes, the question of how long to continue collecting observational data, and the issues surrounding the calculation and implications of observer accuracy (Lipinski & Nelson, 1974). Since one primary goal of behavioral assessment is the collection of a representative sample of behaviors of interest (Goldfried & Kent, 1972), a methodology by which to collect a valid sample is a fundamental issue.

A popular method of collecting systematic observations of behavior is grouping similar behaviors into coded categories. Behaviors included in and excluded from each category are carefully defined. The occurrence of behaviors defined by the code are either noted on a precoded observation sheet or the observer records letters which represent code categories as the behavior occurs. No record is kept of behaviors the subject emits that are not specifically defined by the code. Two popular codes currently in use as assessment and research devices are the Patterson family interaction code (Patterson _et al_., 1969) and a series of O'Leary disruptive classroom behavior codes (O'Leary _et al_., 1970; O'Leary _et al_., 1971). The Patterson code consists of thirty five behavior categories used to record child behavior/familial response interactions. It allows for a rapid sequential recording of the child's behavior, the family member's response to him, the child's ensuing response, etc. The O'Leary codes delineate nine categories of behavior that are typical of the kind of disruptive behaviors a child is likely to emit in the classroom (viz. out of chair, modified out of chair, touching other's property, vocalization, playing, orienting, noise, aggression, and time off task). If one of these behaviors occurs during a given interval, the symbol for that behavior is circled on a precoded data sheet.

One advantage of using codes is that the observer can devote most of his attention to observing the subject's behavior. If instead, the observer had to write out everything the subject did, most of his attention would have to be focused on the recording of behaviors (Lipinski & Nelson, 1974). There are two major disadvantages associated with using codes. By a priori selection of certain categories for inclusion in the behavioral code, certain other categories of behavior are necessarily excluded, so data on the occurrence of these excluded categories are lost (Johnson & Bolstad, 1973). A second problem is that the only aspect of the behavior recorded by the coder is that the behavior occurred sometime during the interval. Usually no information is recorded about the duration of the behavior or when in the interval the behavior was emitted (Lipinski & Nelson, 1974).

The decision of how long to continue baseline data collection by way of systematic observation is usually a function of a subjective judgement by the experimenter. It is necessary to determine how many data points are required to get a true baseline estimate of the occurrence of a behavior or to be able to assume that a stable sample of target behavior frequency has been obtained. The experimenter must decide how low the variance of the behavior around its mean occurrence should be (Lipinski & Nelson, 1974). Given the reactive nature of "being observed", Callahan and Alevizos (1973) suggest that the experimenter's goal should be to find the minimum number of data points needed to get an adequate sample. Eckman (1973), discussing the cost involved in data collection in terms of time and funds, advocates a similar approach. Patterson and Harris (1968) point out that varying amounts of time are needed to obtain stable estimates of behavior, depending on the variables controlling the behavior. There is therefore no absolute criterion of how many data points are sufficient to give a stable representative sampling estimate of the target behavior.

For systematic observations to be valid as assessment or research measures, a minimum requirement is that the observers' recordings be an accurate representation of behaviors that have transpired. Low levels of inter-observer agreement increase the chance of making a Type II error or failing to reject the null hypothesis because true differences in procedures may not be detected (Johnson & Bolstad, 1973; Reid, Skindrud, Taplin, & Jones, 1973). For this reason, most studies employing observations as a data source will periodically use a second observer to check the reliability of the primary observers' recordings. As Baer, Wolf, and Risley (1968, p. 93) point out:

> If humans are observing and recording the behavior under study, then any change may represent a change only in their observing and recording responses, rather than the subject's behavior. Explicit measurement of the reliability of human observers becomes not merely good technique but a prime criterion of whether the study was appropriately behavioral.

Of course, although high reliability is a necessary condition for high validity, it is not a sufficient condition. High agreement does not imply high validity for the data collected or that the observation category is a valid measure of the target behavior (Lipinski & Nelson, 1974). Rather, high reliability is only a minimum criterion for high validity of naturalistic observations.

An observer's recordings of video or audio taped materials may be compared to some previously established criterion profile to obtain an index of observer accuracy. Ratings of behavior by one observer either recorded on tape or presented in vivo may be compared to the ratings of another observer, observing the same subjects, using the same recording technique to obtain an estimate of inter-observer agreement (Johnson & Bolstad, 1973). Here, the terms observer accuracy, inter-observer agreement, and reliability will be used interchangeably to assess the extent to which two observers record the same behavioral frequencies for a particular subject. Some variables that effect observer accuracy are the method by which reliability is calculated (Repp, Deitz, Boles, Deitz, & Repp, 1976), whether or not observers know reliability checks are in progress, the nature of the observers (Skindrud, 1973) and the observational setting (Patterson & Harris, 1968) and the recording procedure used (Mash & McElwee, 1974).

Two common methods of analyzing observational data to calculate an index of inter-observer agreement are per cent agreement and correlation analysis of the two observers' recordings. To calculate per cent agreement indices for continuous or high frequency behaviors, the observation period is commonly divided into units or intervals of arbitrary length. Agreement is calculated by dividing the number of intervals in which the two observers agree on the occurrence of the target behavior divided by the total number of intervals for which the behavior is observed:

$$\frac{\text{agreements}}{\text{agreements + disagreements}} \quad \text{x} \quad 100.$$

Variations in this method include using different interval lengths and counting intervals where no response is recorded by either observer as agreement intervals or not counting those intervals at all. A second method of obtaining a per cent agreement score is more appropriate when the dependent measure is a frequency count of the behavior over time. This procedure involves counting the number of instances of the target behavior rated by each observer, comparing the smaller to the larger number of recorded occurrences and multiplying this ratio by 100:

$$\frac{\text{smaller}}{\text{larger}} \quad \text{x} \quad 100$$

(Repp _et al_., 1976).

One problem of using a per cent agreement criterion is as interval length increases, it becomes increasingly difficult to ascertain if the two observers are actually coding the occurrence of the same behavior during the interval, or if they are actually attending to two discrete responses emitted by the subject during that interval period. This is an example of how high reliability does not necessarily imply high validity. Another problem involves the base rate of the behavior. The percent agreement obtained to the percent agreement that could have been obtained by chance, chance agreement being defined as the square of the base rate of the behavior (Johnson & Bolstad, 1973).

Johnson and Bolstad (1973) suggest that whenever possible, a correlational approach to agreement calculation should be used. This method is particularly useful when the base rate of chance agreement approaches 1.00, when there is a limited sample of monitored as opposed to not monitored for accuracy, or when the observations are based on extended time samples. One problem associated with the use of correlational reliability analysis is that high correlation coefficients may be obtained if one observer consistently over or under estimates behavioral frequencies.

One study by Repp _et al_. (1976) illustrates how different methods of computing inter-observer agreement can lead to significant differences in the reported "reliability" of observers. Two observers working simultaneously rated the behaviors of five children in terms of five behavioral categories. All five behaviors were rated at once. After these observational data were collected, the two transcripts were compared for agreement analyzing each behavioral category seperately and using several different ways of computing agreement which are found in the literature. The mean percentage of inter-observer agreement across all behaviors ranged from 64% to 94%, depending on the computational method used. In all cases, an exact agreement method where agreement was defined as both observers recording the same number of responses for an interval resulted in the lowest per cent agreement. A response intervals only method where only those intervals that both observers agreed that a response had occurred were counted as agreement intervals also consistently yielded a lower per cent agreements than counting those intervals where neither observer recorded the occurrence of a response as agreement intervals also. Interval length was also manipulated. The length of the interval had no significant effect on agreement levels, but as interval length increased, differences in agreement percentages obtained using the different methods of agreement computation increased. This suggests that percentage agreement scores are a misleading index of observer reliability, since these percentages may be more a function of the method the experimenter selects to compute agreement than true levels of observer accuracy.

While a major component of levels of inter-observer agreement may be computational artifacts, independent of true levels of observer accuracy or training, agreement itself can be conceptualized as a function of three major factors: observer characteristics, the observation setting, and most important, the recording procedure used (Cronbach, Gleser, Nanda, & Rajarratnam, 1972). Observer characteristics include sex, age, intelligence (Skindrud, 1973), expectancies (Johnson & Bolstad, 1973), and prior observational experiences (O'Leary & Kent, 1973; Reid, 1970). Characteristics of the observation setting include the number of subjects being observed, their personal characteristics, and the nature of the behaviors they emit in terms of frequency, rate, and temporal sequencing. For example, accuracy is known to vary as a function of which settings and categories of responses are sampled (Patterson & Harris, 1968). Components of the recording procedure include the nature of the observation procedure and the complexity of code categories (Mash & McElwee, 1974).

One characteristic of the observation setting which may differentially effect observer accuracy is the use of overt as opposed to covert methods of assessing the accuracy of the observers' recordings of behaviors, or reliability. Given the reactive effects of "being observed", it is only logical that when an observer knows his recordings are being checked for accuracy, the quality of such recordings will differ from when he is unaware of such reliability checks (Callahan & Alevizos, 1973). Johnson and Bolstad (1973) point out that when an observer knows the accuracy of his recordings is being assessed, he will tend to be particularly careful to code behaviors accurately on that occasion. For this reason, it is difficult to generalize about the overall accuracy of observations from a sample of observations monitored by an overt assessment procedure, since the stimulus situations are not the same. Recording behaviors is itself a behavior, and thus the observer's recording behaviors may be expected to be situation specific (Mischel, 1968).

It is likely that these unrepresentative estimates of reliability computed with the observers' knowledge are inflated estimates of the true overall level of accuracy of observations. Several studies have shown that when observers know that reliability assessments are in progress, they will record more accurately, attempt to match idiosyncratic definitions of the behavioral code employed by the designated reliability checker, and make computational errors which will tend to inflate inter-observer agreement levels (O'Leary & Kent, 1973).

A series of studies by Reid indicate that overt reliability assessments, at regular intervals, may not be an accurate representation of the accuracy of data generated on a day to day basis. Reid (1970) found that inter-observer agreement dropped an average of more than twenty five percentage points from the last day of training and overt assessment to the first day covert assessment procedures were in operation. Reid and DeMaster (unpublished manuscript) found that observer agreement dropped from .82 to .27 or an overall drop of forty five percentage points from the last day of training and overt assessment to the first day of covert assessment. Reid and DeMaster also found that on overt, spot check reliability days, the mean level of inter-observer agreement exceeded levels reached during training.

The results of these studies suggest that the observer's recording behaviors may vary as a function of knowledge of reliability checks. They also suggest the nongeneralizability of observer accuracy from overt assessment situations to those situations where the observer is unaware of assessment or when there is no assessment being performed. Taplin and Reid (1973) found that random accuracy assessments without observer awareness of these checks gave the best reliability estimate across all observational sessions. Observers informed that accuracy checks would be conducted at random, unspecified times showed the lowest decrement in accuracy between the last day of training and overt reliability assessment and the beginning of covert assessment.

Given the reactive nature of reliability assessment demonstrated by these studies, Romancyzk _et al_. (1973) conducted a study to determine whether observers will systematically modify their recordings of behavioral categories to match the idiosyncratic definitions of categories employed by identified assessors. The results suggest that observers will adjust rating criteria as a function of the feedback they receive and the idiosyncratic definitions of behavioral categories employed by other observers with whom they compute reliability, leading to less reliable and valid recordings of behavior. In addition, observers tended to record less deviant behaviors when assessment was covert, suggesting that observational data may be systematically biased towards underestimates of disruptive behaviors when reliability checks are not in progress, limiting the generalizability of the data.

In addition to the overt or covert nature of the reliability assessment procedure, the components of the recording procedure, the nature of the observation procedure and the complexity of the coding categories used, may also differentially effect observer accuracy levels. The nature of the observation procedure may consist of taking a time sampling, frequency count, or duration measure of the behavior of interest (Lipinski & Nelson, 1974). It would seem that the ideal observation procedure is one that permits one observer to be responsible for the behaviors of one subject exclusively. However, while research settings often will make use of such a one to one arrangement, the observer to subject ratio for projects conducted in the applied setting is usually less favorable (Thomson, Holmberg & Baer, 1975).

Using time sampling as an assessment device, Thomson _et al_. (1975) divided up observation time in different ways to find what method of intermittent time sampling could give the most accurate sampling estimate of behaviors when compared to an ongoing observational record. Three teachers were observed for 60 four minute periods, divided into ten second intervals. A continuous ongoing record time sample was made of two categories of behavior, reinforcement of peer interaction and priming of peer interaction. The observation period was divided into sixteen four minute segments. Three methods of time sampling were used.

The Contiguous method was designed to monitor the behavior of a subject for the longest possible unbroken time span. The observation period was divided into quarters. Only those behaviors occurring during the first quarter of the period were sampled from the ongoing record. The Alternating method divided the observation period in half. Behaviors were sampled from the ongoing record for alternating four minute segments for the first half of the observation period (32 minutes) only. Behaviors occurring during the second half of the period were not considered. The Sequential method divided the observation period into quarters. Behaviors were sampled from the ongoing record for one four minute segment in every sixteen minutes of observation.

All observation methods sampled one quarter of the entire observational period for each subject. The experimental methods were compared to the criterion method of ongoing recording wherein each subject's behavior was represented by the entire 60 four minute observation period. From this ongoing record or criterion protocol, those time segments of recorded behaviors that were sampled by each of the three experimental methods was separated from the ongoing record. The frequency of each behavior, recorded during these sampled segments, amounting to one fourth the observation period, were prorated to estimate the frequencies that theoretically should have been obtained if the observations had been continued throughout the period, and if the sampled segments did in fact constitute a representative sample of behaviors throughout the entire observation period.

For the two behaviors, the average error of estimate for each experimental method compared to the ongoing criterion method, ranged for the Contiguous method from 25% to 50%, and from 30% to 52%; for the Alternating method, from 18% to 48%, and from 11% to 55%; and for the Sequential method, from 1% to 38%, and from 4% to 11%. The Sequential method was associated with the smallest percentage of error overall. It is suggested that the Sequential method gave the best estimate of behavioral frequency, because it gave the most widely dispersed sample of the entire observation period.

In addition to the nature of the observation procedure, a second component of the recording procedure, category complexity, also has an effect on observer accuracy. Briefly, category complexity is a "measure of the number of discriminations required of an observer during a data collection session" (Reid, Skindrud, Taplin, & Jones, 1973, p. 2). Complexity depends on the number of coded interactions or behavioral categories used. An increase in the number of behaviors the observer must keep track of at one time leads to a concurrent increase in the difficulty associated with the number of different discriminations the observer must make to code the subject's behavior accurately (Mash & McElwee, 1974). Several studies reviewed by Reid _et al_. (1973) have shown that complexity is inversely related to observer agreement and accuracy.

Taplin and Reid (1973) compared the accuracy of observations to the relative complexity to be coded. Category complexity was defined as the number of different categories used divided by the total number of entries made. The correlation between complexity and accuracy was -.52 which suggests that accuracy tends to decrease as interaction complexity increases. Skindrud (1972) found significant negative correlations between observer agreement and the per cent of unrepeated interactions within each observation segment, which also may be taken as an index of complexity. Finally, Reid (1970) found a significant negative correlation of -.75 between complexity of the observers' protocols and per cent agreement.

Reid et al. (1973) suggest several implications that category complexity as it effects observer accuracy has for the generalizability or external validity of observational data. Complexity of observed behaviors may vary from session to session and subject to subject. When training observers for use as data gatherers, usually some predetermined criterion accuracy level must be reached before the observer may participate in the experiment. In many cases, the observer may reach this criterion accuracy level if by chance the sessions or subjects to be rated on accuracyassessment days are extremely simple ones. This can lead to an inflated index of observer skill, or the extent to which accuracy levels obtained during training will generalize to or represent the reliability of post training data. In a similar way, spot check methods of reliability assessment may overestimate the reliability of unchecked observations. Reid et al. also suggest that the complexity of interactions may be used to predict mathematically an estimate of reliability of unmonitored interactions of one particular level of complexity, given the reliability of monitored interactions of another particular level of complexity.

Mash and McElwee (1974) suggest that observer accuracy is situation dependent and agree that the stability of observer accuracy is doubtful. A study was conducted that manipulated the complexity of code categories, the patterning of behavior, and prior observational history of the observer to assess the effects of these variables on recording accuracy. Category complexity was defined both in terms of number and kind.

Observers were required to code a series of pre-recorded verbal statements in accordance with categories defined by two coding systems. Increased code complexity was obtained by dividing more inclusive behavioral categories into finer units. The broader code divided verbalization into four categories. The eight category system divided each of these four categories into two subcategories. Other variables manipulated were the predictable vs. unpredictable nature of interactions, and the observers' prior experience with coding interactions in terms of this predictability.

Mash and McElwee found that observers using the less complex four category system seemed to learn the code faster. There was a significant inverse relationship between complexity of the coding system and criterion agreement scores. They suggest that observer accuracy is a situation specific response, dependent on observer characteristics, conditions of observation, and recording procedure characteristics. Consistent with the results of the Reid (1970) and Romancyzk _et al_. (1973) investigations, an observer's past accuracy levels were shown to not always provide accurate estimates of future performance if situational variables are not consistent. Training conditions should then approximate the observation setting as closely as possible, or observer training should be conducted in a diverse sample of observation conditions.

Procedural problems involved in the direct observation of behavior in the applied or laboratory setting can affect the degree to which the observational record represents a truly unbiased sample of the behaviors of interest, or the internal validity of the data points. Some procedural problems associated with the use of systematic observations include the loss of information resulting from the use of codes, deciding how many data points are sufficient to obtain stable estimates of behavior, calculating indices of observer agreement, and the generalizability of observer accuracy data across situations, observational methods, and behaviors.

In the literature reviewed so far dealing with the methodological problems associated with the use of direct systematic observations of behavior as assessment and research measures, it is apparent that methods of observation adequate for laboratory research may not be as ideally suited for the applied setting. For example, academically situated researchers often have sufficient grant funds to pay observers or a substantial population of graduate or undergraduate students to serve as unpaid observers. This is rarely the case in applied settings (Eckman, 1973; Thomson *et al.*, 1975). Also, many experimental observation laboratories are equipped with two way mirrors and covert video and audiotaping facilities which may be used to minimize the obtrusiveness of observers. In the naturalistic setting, such equipment is usually not available, so that observers have to work in the same room as their subjects. In this way, these systematic observations, while often valid representations of behavior in the controlled laboratory setting, may not represent a true sample of behaviors of interest in the naturalistic settin, owing to the reactive effects of "being observed" (Patterson & Harris, 1968). Further, it has been suggested that the more obtrusive the observer, the greater these reactive effects may be (Callahan & Alevizos, 1973).

In both laboratory and naturalistic setting observations, another common practice that may be considered far from ideal is having one observer track between nine (O'Leary _et al._, 1970, 1971) and thirty five (Patterson et. al., 1969) behavioral categories at once, with the observer taking a continuous time sampling of each of these code categories simultaneously. This common practice may have limited utility in that the accuracy of an observer's recordings is seen to decrease as a function of the number of interactions observed. To get a more representative sample of behaviors of interest, one must control for the variation in accuracy of the observations resulting from the inverse relationship between category complexity and observer accuracy (Repp _et al._, 1976). Current methods generally do not control for this complexity dimension.

It has already been pointed out that a basic goal of behavioral assessment is to obtain a non-inferential situation specific sample of behaviors of interest (Goldfried & Kent, 1973). The more representative and objective the sample is, the less need there should be to make subjective judgements about the observed behaviors. Given the methodological problems associated with the use of typical methods of time samplings of behaviors for assessment and research purposes, it is difficult to obtain such a truly representative sample of target behaviors.

What is needed is an observational method, suited to the applied as well as the experimental laboratory setting, which minimizes observee reactivity and maximizes observer accuracy by decreasing category complexity, to provide a true sample of behaviors. Such a procedure would improve the validity of naturalistic observations, promoting the generalizability of the data. An alternate procedural method of time sampling is proposed to compensate for some of these problems, which if uncontrolled may jeopardize the reliability and validity or generalizability of the data.

Statement of Purpose

The purpose of the present study was to assess the reliability, representativeness, and utility of a particular procedural method of time sampling. Reliability was assessed in terms of how closely two observers recording the same material would agree on what had transpired during the recording session. Representativeness or "relative validity" was assessed by comparing the behavior frequencies generated by the use of each observational method. Utility was assessed subjectively, in terms of how useful an observation method would be for the clinician, given its level of reliability and validity. In order for an observation method to be truly useful for the clinician, it must be suited to the applied as well as the laboratory setting, and be economical to use in terms of training time and the number of observers required to get a truly representative sample of behaviors.

The proposed method of observations used one observer to monitor the behavior of one subject during the observation period. Four categories of behavior were monitored. To minimize the complexity of the coding system, only one behavior was recorded during a particular interval. To get a truly representative sample of that behavior during the observation period, the target behavior that was tracked during a particular interval was varied systematically between intervals so that behavior category one was observed during interval one, behavior two during interval two, etc. This pattern was repeated for each observation period.

This sequential method of intermittent time sampling of behaviors was proposed since it was expected to have several advantages over methods currently in use:

1) Since reliability is seen to decrease as category complexity increases, this method was designed to have each observer record only one behavioral category per interval.

2) The cost of observations in terms of both time and funds can be minimized by using only one observer to rate all of a subject's behaviors of interest, except during those infrequent sessions when reliability assessments are being conducted.

3) Since reactivity to "being observed" may be effected by the obtrusiveness (e. g. number) of the observers, reactivity should be minimized by this method which requires only one observer to sample a subject's behavior, again except during reliability assessments when two observers must necessarily be present at one time.

4) In addition, training time on the sequential method should be minimized, owing to the simple nature of the recording procedure (Mash & McElwee, 1974).

To assess the reliability and representativeness of the proposed time sampling procedure, 10 observers used five different recording procedures to rate four categories of disruptive child behaviors. The five different methods of observation varied along two dimensions: the complexity of the coding system (number of behavioral categories) and the nature of the sampling procedure (continuous vs. intermittent) used to record these behaviors (see Fig. 1). Method 1 involved taking a continuous time sampling of all four behavioral categories simultaneously. Method 2 (which had two variations to account for all four behaviors) involved taking a continuous time sampling of two behaviors per interval. Method 5 (which had four variations to account for all four behaviors) involved taking a continuous time sampling of one behavioral category per interval. For these three methods, each behavior category was observable during each of 24 observation intervals.

The technique of primary interest in this investigation, Method 3, involved taking an intermittent time sampling of all four behavior categories, one per interval in a sequential manner. Each behavior category was observable for six of 24 observation intervals. Method 4 (which had two variations to account for all four behaviors) involved taking an intermittent time sampling of two categories of behavior in a sequential manner. Each behavior category was observable for 12 of 24 intervals.

Inter-observer agreement levels were calculated between the observers and the experimenter for each of the five observation methods. To calculate reliability, an exact agreements formula (Johnson & Bolstad, 1973) was used, where reliability was expressed as the number of intervals the observer and reliability checker agreed a behavior had occurred, divided by the number of agreements plus disagreements on the occurrence of the target behavior. It was predicted that inter-observer agreement levels would be the highest between the reliability checker and the observers when reliability on Method 5 was computed, since Method 5 involved the least complex observational operation (Reid et al., 1973).

Behavior A	Behavior A	Behavior A	Behavior A
Behavior B	Behavior B	Behavior B	Behavior B
Behavior C	Behavior C	Behavior C	Behavior C
Behavior D	Behavior D	Behavior D	Behavior D

Method 1

Behavior A	Behavior A	Behavior A	Behavior A
Behavior B	Behavior B	Behavior B	Behavior B

Method 2

Behavior A	Behavior B	Behavior C	Behavior D

Method 3

Behavior A	Behavior B	Behavior A	Behavior B

Method 4

Behavior A	Behavior A	Behavior A	Behavior A

Method 5

Figure 1. The number of behavioral categories recorded and the nature of the sampling procedure (continuous vs. intermittent) used to record these behaviors by each of the five methods of observation. Four observation intervals are shown.

High levels of agreement were also predicted for Methods 3 and 4 since the observer was still required to track only one behavior per interval. In contrast, lower levels of agreement were expected for Methods 1 and 2, since the observers were required to use a more complex recording procedure.

It seemed likely that the sequential methods of observation would generate highly reliable recordings of behavior, since category complexity was minimized. However, high agreement alone does not imply high validity for the data collected; the observational record may not accurately reflect what has really transpired during the observation period (Johnson & Bolstad, 1973). Behavior frequencies may be grossly distorted.

A further purpose of the present investigation was to assess the representativeness of the observational record generated by the sequential methods of intermittent time sampling. To study the effects of the nature of the sampling procedure on the representativeness of the data, sampling estimates of the behavioral frequencies obtained using the sequential methods of time sampling were compared to data obtained using the other observational methods. It was assumed that Method 5 would yield the most accurate record of what transpired during the observation period. Thus, if behavioral frequencies generated using Method 5 and the sequential methods are comparable, using the sequential methods should provide a reliable estimate of behavioral frequency. It was predicted that relying on a recording procedure that sampled behaviors every second or fourth interval would not lead to a gross under or overestimate of behavioral frequencies, compared to methods that sampled the target behavior during each interval.

The utility of an observational method was determined according to three criteria: reliability, representativeness, and practicality. Once reliability and representativeness (frequency) issues had been considered for a particular method, of paramount importance was the usefulness of the method for the clinician. Several issues to consider in deciding on a method's utility included the length of training time required for observers to use the method reliably and maintenance of high levels of reliability after training, such that retraining is not necessary. Also important are the number of observers required to obtain a reliable sample of behaviors and the amount of observation time required to record a representative sample of behaviors.

It was predicted that of all the methods under investigation, the sequential methods of observation would present the most practical alternative. Given adequate reliability and representativeness, training time, cost factors, and reactivity related to use of these sequential methods should be minimal. All of these factors combined promise an efficient observation package for the clinician.

CHAPTER II

METHOD

Experimental Design

Ten observers took a time sampling of four categories of disruptive classroom behaviors emitted by two children. These behaviors had been prerecorded on videotapes. Five different methods of recording, which involved ten observational operations to account for all four behaviors, were used to record these behaviors. These recording methods will be explained in detail in a subsequent section.

Each observer used one of these ten observational operations to record the behaviors of one of two children during each tape segment. A videotape segment consisted of 12 minutes of taped child behaviors. In all the observers recorded behaviors for 40 segments. During 20 of these segments they recorded behaviors of one of the two children. During the other 20 segments, the other child's behaviors were recorded. The order in which each child was observed was determined randomly. Each observer used each observational operation two times per child, four times in all during the course of the experiment. Different observers used different operations during each segment. For example, during segment one, observer one might use operation one; observer two, operation two; etc. To control for sequencing effects, the ten observers used each of the ten operations in different sequences. A Latin square design was used to determine the order in which the observers used each observation operation. A typical series of observations could be represented by the following hypothetical matrix of observations (see Fig. 2) which allows for the observer to use all ten observational operations to observe the child by the end of the tenth session.

Subjects

Observers were ten undergraduate research assistants enrolled in an independent studies course (Psychology 333) at the University of North Carolina at Greensboro. In addition to training, each observer participated in forty observation sessions for a total of eight hours of observation time per observer.

The observers were told that the purpose of the project was to determine which method of recording would be easiest for them to learn and the most accurate for them to use. They were instructed to try hard to record the behaviors as carefully as possible at all times. They were told that the length of time they would be required to participate in the experiment would be determined by how accurately they recorded the videotapes: the more reliably they recorded the tapes, the sooner they would be finished.

No indication was given of expected results. The observers were repeatedly told during the experiment that the experimenter was not sure which method was best because of conflicting reports in the literature. At the conclusion of the experiment, the observers were informed of the expected results. Some commented that it had seemed obvious to them that the simpler methods would be easiest to use.

Observer	Observation Order									
1	1	2A	2B	3	4A	4B	5A	5B	5C	5D
2	2A	2B	3	4A	4B	5A	5B	5C	5D	1
3	2B	3	4A	4B	5A	5B	5C	5D	1	2A
4	3	4A	4B	5A	5B	5C	5D	1	2A	2B
5	4A	4B	5A	5B	5C	5D	1	2A	2B	3
6	4B	5A	5B	5C	5D	1	2A	2B	3	4A
7	5A	5B	5C	5D	1	2A	2B	3	4A	4B
8	5B	5C	5D	1	2A	2B	3	4A	4B	5A
9	5C	5D	1	2A	2B	3	4A	4B	5A	5B
10	5D	1	2A	2B	3	4A	4B	5A	5B	5C

Figure 2. A hypothetical matrix of observations representing an order in which the observers could have used each observational operation to complete an observation series on a child. The letters A, B, C, and D represent the different variations of the five observational methods.

Stimulus Materials

The disruptive classroom behaviors of two children, which had been prerecorded on videotape were used as stimulus materials. The two children were approximately eight years old and were enrolled in Point of Woods Laboratory School at the State University of New York at Stony Brook. Point of Woods is a special school for children with behavior problems that is run by the Psychology department at Stony Brook. The classroom behaviors of these two children had been recorded on videotape for use in a series of observational studies conducted at Point of Woods by Dr. R. N. Kent and Dr. K. D. O'Leary, principal of Point of Woods.

The videotapes used in this investigation were copies of selected twelve minute segments of the master tapes prepared by Drs. Kent and O'Leary. The method of preparation of these tapes is outlined in Kent et al. (1974). At the beginning of each twelve minute tape segment, there is a verbal countdown recorded ("three, two, one, go!") to facilitate stopwatch synchronization. The tapes were played on a Panasonic NV 3020 videotape deck and were viewed on a Concord Solid State videotape monitor. The observers used stop watches to keep track of the beginning and end of each observation interval. Observations were recorded on precoded data sheets (see Appendix A) that indicated specifically which behavior(s) the observer needed to record during each interval.

Dependent Variables

Four categories of behavior defined by O'Leary's disruptive behavior code (O'Leary et al., 1971) were recorded. These target behaviors were Out of Chair, Playing, Time Off Task, and Orienting. These four categories of behavior may be briefly defined as:

Out of Chair: child moves body from chair without teacher's permission.

Time Off Task: child fails to attend to assigned work for an entire observation interval.

Playing: child uses his hands to play with his or another's property.

Orienting: child moves face more than 90° from point of reference while seated.

(For more complete definitions of these categories selected from the O'Leary code, see Appendix B.)

For each of these four target behaviors, there were two dependent variables of interest, the reliability and frequency with which each behavior was recorded. Frequency data was collected to assess how reliably each observation method portrayed what had transpired during the observation period, compared to the behavior frequencies estimated by the use of Method 5, assumed to be the most reliable observational method.

Of primary concern was the reliability of the observational data generated by a particular method, since reliability of the data is a necessary (although not sufficient) condition for the validity of the data. Reliability was represented by the level of inter-observer agreement obtained between each subject and an independent reliability checker (the experimenter) using the same method of observation to record equivalent segments of videotape, in an interval by interval comparison of their recordings. Levels of inter-observer agreement were computed by comparing the number of intervals in which both the observer and the reliability checker agreed a particular behavior had occurred to the number of intervals the two agreed plus the number of intervals the two disagreed on whether a target behavior had occurred, or,

$$\text{Reliability} = \frac{\text{no. of agreements on the occurrence of the behavior}}{\text{no. of agreements + disagreements.}}$$

This method has been suggested (Johnson & Bolstad, 1973) as the most appropriate way to compare recording behaviors of two observers taking a time sampling. This exact agreement method does not inflate reliability levels as a function of computational artifacts as is often the case with other methods of reliability computation (Repp et al., 1976).

The second dependent variable of interest was the frequency with which each behavior was recorded by the observers using different methods. These frequencies or sampling estimates were considered to represent the representativeness: of the observational data, answering the question: To what extent was each observational method generating data that represented a reliable picture of what had transpired during a tape segment? To evaluate the representativeness of a particular observation method, as assessed by the frequencies generated by its use, a sampling estimate for the occurrence of each behavior was computed for each method of observation. These sampling estimates were obtained by prorating the frequencies with which each behavior was observed by comparing the total number of times an observer recorded the occurrence of a behavior to the total number of times it would have been possible to record the occurrence of the behavior. Or,

$$\text{Frequency} = \frac{\text{No. of intervals behavior is observed}}{\text{No. of intervals behavior could be observed.}}$$

Frequency estimates were computed separately for each of the four target behaviors as recorded by each of the five observational methods throughout all forty tape segments.

Given the reliability and representativeness of each method, a method's utility was assessed by combining these measures with several other characteristics of the method to estimate the practicality of its use by the clinician. Other factors considered in deciding a method's utility included the length of training time required for observers to use the method reliably, the number of observers required to obtain a reliable sample of behaviors, and the amount of observation time required to record a truly representative sample of behaviors.

Methods of Observation

Five methods of observation which varied along two dimensions: the complexity of the coding system (number of behavioral categories) and the nature of the sampling procedure (continuous vs. intermittent) were compared in terms of the reliability and representativeness of the observational records they generated and the overall utility of the method. Observation periods were twelve minutes long and were divided into twenty four, 30 second intervals. The first twenty seconds of each interval was used for observing and the last ten seconds for recording the target behaviors.

Three methods of observation involved taking a continuous time sampling of behaviors of interest, while the other two procedures involved an intermittent time sampling procedure (See Fig. 1). Methods 1 and 2 consisted of taking a continuous time sampling of behaviors for the 20 seconds of observation time per interval. Method 1 involved rating four behavioral categories simultaneously during each interval. Method 2 involved rating two of these categories during each interval. There were two variations of Method 2 to account for all four behavioral categories. These types of recording procedures, comprised of rating several behavioral categories simultaneously are commonly found in the literature (Johnson & Bolstad, 1973).

Methods 3 and 4 consisted of taking an intermittent time sampling of behaviors, one at a time, for the 20 seconds of observation time per interval in a sequential manner. The target behavior to be recorded during each interval was alternated systematically by interval. Method 3 involved rating all four categories of behavior in the sequential manner described above. Method 4 involved rating two of the behavioral categories in this alternating, sequential manner. There were two variations of Method 4 to account for all four behavioral categories. These two sequential methods were designed to provide a widely dispersed sample of behavior throughout the observation period, while minimizing category complexity.

Method 5 involved taking a continuous time sampling of a single behavior throughout the entire observation session. There were four variations of method 5 to account for all four behaviors. Recording of only one behavior at a time was expected to provide the most accurate estimate of occurrence for a behavior since it has been shown that observer accuracy is an inverse function of category complexity (Mash & McElwee, 1974).

Training of Observers

The observers were trained over a two week period to record the four categories of behavior sampled from the O'Leary disruptive behavior code (O'Leary *et al*., 1971) using the five methods of observation. Each observer had to reach an agreement level of at least .75 with the trainer for two consective four minute videotape segments to complete training on that method.

Each observer was given a copy of the behavioral code to study before the first meeting. This introductory meeting lasted approximately one hour during which time the experimenter redefined the code categories verbally for the ten observers and role played sample behaviors that would and would not be included under each category. The observers then watched but did not record from a twelve minute sample videotape segment. The experimenter pointed out those behaviors emitted by each child that would be coded under each of the four relevant categories, stopping the videotape periodically so that the observers could ask questions about why a particular response would or would not be coded as Out of Chair, Playing, Time Off Task, or Orienting.

The formal training sessions were conducted over a ten day period. The observers were broken up into five groups of two for more individualized instruction in the five observational methods. Different segments of videotape were used to train each method with each observer pair. Both members of each of the five pairs of observers learned the same method at the same time. The order in which each pair of observers learned each method was varied systematically using a Latin Square design to control for the effects of order of exposure on facilitating learning to use each method reliably (see Fig. 3).

Observer Pair Number	Training Order Method				
1	1	2	3	4	5
2	2	3	4	5	1
3	3	4	5	1	2
4	4	5	1	2	3
5	5	1	2	3	4

Figure 3. The order in which each observer pair learned each observational method.

Training sessions were standardized across methods and observer pairs and conducted as follows:

At the beginning of each training session, observers were asked if they had any questions about the code categories or the recording procedure to be used. The definitions of the behavioral categories were repeated, stressing the critical points of each response class.

To introduce the observers to a new recording procedure, one minute of videotape was initially presented for the observer to rate. Discrepancies in recording between the observers and the trainer were discussed and misconceptions about code categories explicated. This procedure was repeated once. Then, a two minute segment of videotape was presented for the observers to rate, and again the observers were given feedback on their recordings with discrepancies between their and the trainer's recordings discussed.

Finally, to test the extent to which each observer had learned how to use the method reliably, four minute test segments of videotapes were presented for the observers to rate. After each four minute segment had been recorded, agreement indices between the observer and the trainer were calculated. Again, the observers were given feedback on the reliability of their recordings. When an observer's recordings reached a criterion reliability level of .75 or greater for each behavior with the trainer's ratings for two consecutive four minute videotape segments, training on that method was considered complete. When one member of an observing pair completed training before the other pair member, he was excused from further training sessions on that method. Once observers completed training on all five observational methods, the experimental phase of observations began.

Experimental Sessions

Experimental sessions were conducted over a two week period. Observers were seated at a comfortable viewing distance in front of the videotape monitor. Between two and four observers were present during any one particular session. They were shielded from each other to minimize the possibility of observer cheating or collusion to obtain high agreement levels. In addition, the experimenter was present during all sessions to reduce further the possibility of observer cheating.

Each observation session lasted twelve minutes. During that twelve minute period, the observer rated the behaviors of one of the two children, using one of the ten variations of the five observational methods. An observer used different methods during different sessions. By the end of ten sessions the observer had used each of the ten observational operations once to record the behaviors of one child, completing a series of observations on that child. Two series of observation were completed for each child by each observer (see Fig. 2), giving a total of 40 observation segments in all.

Once the observers completed an observation sheet (see Appendix A for sample observation sheets) for a particular session, they no longer had access to the sheet or the data summarized from it, because as O'Leary and Kent (1973) have pointed out, observers who compute their own agreement indices may often make systematic computaional mistakes in the direction that would tend to enhance their accuracy levels.

Reliability Checks by the Experimenter

An independent reliability checker, the experimenter, assessed the reliability of the data recorded by the observers. Since it is necessary to calculate the reliability with which observers use a particular mthod before comparisons between different methods of observation can be conducted, 25% of the segments each observer used each method of observation were randomly selected to be checked for agreement.

The experimenter rated each of these tape segments using the same observation method a particular observer had used to record behaviors during that tape segment. Those tape segments in which both the observer and the experimenter had used the same recording method were compared for agreement. Observers were given session by session feedback on the agreement levels obtained by comparing their observational recordings to the experimenter's. Johnson and Bolstad (1973) have suggested that giving observers differential feedback on the accuracy of their recordings will tend to promote more consistent applications of the behavioral code.

CHAPTER III

RESULTS

Training Phase: Time

An observer was required to demonstrate an accuracy level of at least .75 on two consecutive segments of videotape before pre-experimental training on a particular method was completed. Data were collected on the number of minutes of videotape each observer had to record in order to reach this minimal proficiency level. The one way analysis of variance performed between observational methods on the different methods on the different amounts of training time required for each observer to become "trained" on each method, (Table 1) indicated that there was a highly significant difference in the amount of training time required for each of the observers to become proficient in using a particular observational method, $\underline{F}$ (4,45) = 83.0091, $\underline{p} < .001$.

Results of Newman-Keuls post hoc comparison (Table 2) indicated that Method 1 (continuous recording of four behavior categories at one time), with a mean training time of 48.4 minutes, took significantly longer to train than any of the other four methods. There were no significant differences in mean training times between Method 2 ($\underline{M}$ = 15.2 minutes), Method 3 ($\underline{M}$ = 14.8 minutes), Method 4 ($\underline{M}$ = 12.8 minutes) or Method 5 ($\underline{M}$ = 12.0 minutes).

Training Phase: Reliability

All of the observers' recording behaviors during training were compared to the ratings of the trainer (the independent reliability checker).

Inter-observer agreement coefficients computed for each behavior were calculated by comparing the recording profiles of the observer and the reliability checker rating equivalent tape segments, using the same method of recording. Even though observers had to demonstrate reliability levels of only .75 to be considered "trained" on a particular observation method, most of the observers recorded behaviors with greater than .75 inter-observer agreement levels by the end of training. Hence, there were differences between methods in the reliability with which the observers recorded behaviors.

Agreement scores were analyzed across the five methods of observation. Each data point represented a per cent agreement score calculated from an interval by interval comparison of recordings. In this and all subsequent analyses, an arcsin transformation (Winer, 1971) was performed on the data in its decimal form. Multivariate analysis of variance was performed on the transformed data points arranged in this one way repeated measures design, considering all four behavioral categories at once. This analysis was followed by four separate univariate analyses of variance on each of the four target behavior categories considered individually.

At the conclusion of training, there were significant multivariate differences in the proficiency with which the observers used each observation method, $\underline{F}$ (4,45) = 14.428, $\underline{p} < .001$ (Table 3). Results of Neuman-Keuls comparisons on the canonical means (Table 4) indicated considering all behaviors simultaneously the observers used Method 5 less reliably at the end of training than any of the other observation methods. There was no difference in the reliabilities with which the observers used Method 1, Method 2, Method 3, or Method 4 at the end of training.

When considering the four behaviors individually, there were no significant differences between methods in the reliability with which Playing, F (4,45) = 1.30578, $p < .2816$ (Table 5) and Orienting, F (4,45) = 2.02497, $p < .1061$ (Table 6) behaviors were recorded. The average reliability with which the observers recorded Playing was 82% and Orienting, 85% at the conclusion of training. There were significant univariate differences in the reliability with which the observers used the recording methods for Out of Chair, F (4,45) = 12.63347, $p < .001$ (Table 7) and Time Off Task, F (4,45) = 6.82978, $p < .004$ (Table 9) behaviors at the conclusion of training. The average reliability with which the observers recorded Out of Chair was 96% and Time Off Task, 84% at the conclusion of training.

Post hoc comparisons (Table 8) on the reliability with which the observers used each recording method to rate Out of Chair behaviors indicated that the observers used Method 5 less reliably than all the other methods. There was no difference in the reliability with which the observers used Method 1, Method 2, Method 3, or Method 4 to record Out of Chair behaviors. Post hoc comparisons (Table 10) on the reliability with which the observers used each recording method to rate Time Off Task behaviors, indicated that the observers recorded Time Off Task behaviors more reliably using Method 3 than using Method 1, Method 5, or Method 2. They used Method 4 more reliably than Method 5 or Method 2. There was no difference in the reliability with which they used Method 3 and Method 4, or Method 1 and Method 2 and Method 5 at the end of training to record Time Off Task behaviors.

Experimental Phase: Reliability

Each observer's reliability level on the five methods, as demonstrated at the conclusion of the training phase, was used as a covariate to adjust all future reliability or agreement levels demonstrated on the methods. This process controlled for individual differences in reliability on the observational methods at the end of training effecting subsequent proficiency levels.

Twenty five per cent of the observational records the observers generated using each of the five methods of recording were randomly sampled and checked for agreement with observational records compiled by the independent reliability checker using the same observation method on that particular videotape segment. Inter-observer agreement scores were computed between the observer and the reliability checker for each method for each behavior category. A multivariate analysis of variance was performed between the five observational methods considering the four behavior categories simultaneously. This was followed by four separate univariate analyses of variance on each of the four behavior categories considered individually.

In terms of the relative reliabilities with which the observers used the five observational methods to record behaviors, there was a significant multivariate difference in the levels of inter-observer agreement obtained between the observers and the independent reliability checker between methods, $\underline{F}$ (4,41) = 13.69473, $\underline{p} < .001$ (Table 11). The average reliability with which the observers used each observation method was for Method 1, 51%; Method 2, 73%; Method 3, 86%; Method 4, 81%; and Method 5, 87%.

Neuman-Keuls comparisons of the canonical means (Table 12) indicated for all behaviors there was no difference in the reliability with which the observers used Method 3, Method 4, and Method 5 to record behaviors. Use of these methods generated more reliable recordings than the use of the other two methods. Also, using Method 2 was better than using Method 1. In general, those methods involving recording one category per interval were better than recording two categories per interval which was better than recording four categories of behavior per interval.

For all four categories of behavior considered individually, there were significant univariate differences noted in the reliability with which they were recorded using different methods. The average reliability with which each behavioral category was recorded was for Out of Chair, 86%; Playing, 87%; Time Off Task, 65%; and Orienting, 79%.

For Out of Chair, a significant difference in rating reliability was noted as a function of which recording technique was used, $\underline{F}$ (5,44) = 2.39217 , $\underline{p} <.05$ (Table 13). The average reliability with which the observers rated Out of Chair behaviors was by Method 1, 70%; Method 2, 86%; Method 3, 95%; Method 4, 87%; and Method 5, 90%. Results of Neuman-Keuls post hoc comparisons (Table 14) indicated that Method 3 was superior to Method 1 in terms of the reliability with which the observers recorded Out of Chair behaviors. There were no significant differences in the reliability of the recordings generated by any of the other methods for Out of Chair behaviors.

Similarly, for Playing behaviors, there was a significant difference in rating reliability as a function of recording method, $\underline{F}$ (5,44) = 5.56344, $\underline{p} < .0007$ (Table 15). The average reliabilities with which the observers rated Playing behaviors was for Method 1, 41%; Method 2, 71%; Method 3, 87%; Method 4, 70%; and Method 5, 89%. Results of Neuman-Keuls post hoc analysis (Table 16) indicated that Method 1 yielded the least reliable behavioral recordings compared to the other four methods. There were no significant differences in the reliability of recordings of Playing behaviors generated by any of the other methods.

In the case of Time Off Task, a significant difference in recording reliabilitv was also noted as a function of the recording method used, F (5,44) = 4.47, $p < .0025$(Table 17). The average reliability with which the observers recorded Time Off Task behaviors was for Method 1, 33%; Method 2, 57%; Method 3, 75% Method 4, 84%; and Method 5, 77%. Neuman-Keuls post hoc comparisons (Table 18) indicated that there were no significant differences in the reliability with which the observers rated Time Off Task behaviors between Method 1 and Method 2 and between Method 3, Method 4, and Method 5. Reliability of the recordings generated by Method 1 were significantly lower than those generated by Method 3, Method 4, and Method 5.

Finally, for Orienting behaviors, there was a significant difference in rating reliability associated with the recording methods used, F (5,44) = 4.2074, $p < .004$ (Table 19). The average reliabilities with which the observers recorded Orienting behaviors were for Method 1, 59%; Method 2, 77%; Method 3, 86%; Method 4, 82%; and Method 5, 92%. Neuman-Keuls post hoc comparisons (Table 20) indicated that there were no significant differences in the reliability with which the observers rated Orienting behaviors between Method 1 and Method 2 and Method 3, Method 4, and Method 5. Reliability of the recordings obtained by using Method 1 were significantly worse than those obtained using Method 3, Method 4, or Method 5.

Experimental Phase: Frequency

The degree to which each method of observation tended to under or overestimate the frequency of target behaviors was assessed by comparing the behavioral frequencies of sampling estimates of the occurrence of behaviors obtained for each tape segment using each of the five observation methods. Each data point represented the sampling estimate obtained by each observer using each method of observation on a single behavioral category. Sampling estimates were analyzed across the five observational methods within tape segments.

Multivariate analysis of variance was performed on these frequencies, considering all four behavior categories at once. This analysis was followed by four separate univariate analyses of variance on each of the four target behaviors considered individually. These analyses were designed to indicate whether use of the different observational techniques would result in different sampling estimates.

There were no singificant multivariate differences in the sampling estimates generated within tape segments by using different observation methods, $\underline{F}$ (4,195) = 2.47206, $\underline{p} <$ .10 (Table 21). There were similarly no univariate differences between methods for any of the behavioral categories (See Tables 22 - 25).

CHAPTER IV

DISCUSSION

Reliability

Results of this investigation indicate that the reliability of observers taking a time sampling decreases as a function of the number of behavior categories the observers are required to monitor during each observation interval. Reliabilities across all behaviors when recording one behavior per interval as opposed to two to four behaviors per interval decreased from an average of 91% (mean of methods 3, 4, and 5 combined) to 62% (mean of methods 1 and 2 combined).

Continuous recording of four behavior categories at one time (Method 1) produced the least reliable observational records; inter-observer agreement levels averaged only 48% across behavior categories. Recordings obtained using both sequential methods of recording and continuous recording of one category of behavior per interval consistently demonstrated highly reliable behavioral records. For certain behaviors, Out of Chair and Playing, Method 2 (which involved continuous recording of two categories of behavior per interval) was no less reliable than the sequential methods or Method 5. This can probably be explained in terms of the obvious nature of Out of Chair or Playing responses. It was apparently easier for the observers to discriminate these behaviors than Time Off Task and Orienting behaviors.

These findings are in agreement with previous studies which have found that observer accuracy is an inverse function of the number

(Taplin & Reid, 1973) and subtlety (Mash & McElwee, 1974) of discriminations required of the observers during a data collection session. Negative correlations between observer agreement and the complexity of the recording procedure have been consistently demonstrated (Reid, 1970; Skindrud, 1972; Taplin & Reid, 1973). Apparently, the more complex the recording procedure, the less chance there is that the observational record will represent a truly unbiased, reliable sample of the behaviors of interest.

Since the reliability of the data is a necessary condition for the validity of the data, this notion of the complexity of the recording procedure's impairing the accuracy of observations has serious implications for recording practices commonly employed in behavioral research. Many researchers have observers track seven (e. g., the O'Leary code) or more (e. g., the 35 category Patterson code) behavioral events at one time. These observational data are then used as dependent measures in behavioral outcome research. The problem of evaluating therapeutic success using an unreliable assessment device for both research and clinical purposes is obvious: changes noted in subjects' behavior may be merely a function of changes in the observers' recording behaviors (Baer, Wolf, & Risley, 1968), or conversely, true changes in subjects' behavior may go undetected (Johnson & Bolstad, 1973).

To insure that the observers' recordings of behavior are in fact reliable so that a study may be considered "appropriately behavioral" (Baer _et al_., 1968), the researcher is compelled to devote many hours to training his observers to use a complex behavioral code reliably before data collection may begin. An additional finding of the present investigation was that more complex recording methods took longer to train than simpler methods. Method 1, the most complex method, averaged three and a half to four times longer to train (mean training time = 48.4 minutes) than the other four recording methods (overall mean training time = 13.7 minutes). The obvious disadvantages of extended training time will be discussed in a later section on the utility of the recording methods.

Frequency

Although the sequential methods of observation consistently produced highly reliable recordings of behavior, one possible drawback associated with intermittent time sampling might impair the representativeness of the data generated by the use of these methods. Namely, it is possible that as a function of sampling artifacts (e.g., Repp, Berkler, Roberts, Slack, & Repp, in press; Thomson *et al.*, 1975), extremely biased estimates of behavior frequency might be obtained. By sampling every second or fourth interval exclusively, gross over or under estimates of behavior might occur. For example, a child may be out of his chair only during those intervals in which Playing and Orienting behaviors were recorded. The observational record would indicate less frequent Out of Chair responses when the opposite was true. Conversely, the child might be consistently out of his chair only during those intervals that Out of Chair responses were coded. This distortion of frequencies would necessarily impair the content or internal validity and the generalizability or external validity of the data (Campbell & Stanley, 1966).

Fortunately, distortion of behavioral frequencies as a function of sampling procedures was not a problem in the present investigation. Since no absolute, objective profile of what transpired on the videotapes could be created (unless the tapes presented the behaviors of actors carrying out a predetermined script), a comparison of the sampling estimates generated by using each method was conducted to see if compared to the other methods, the sequential methods lead to gross over or underestimates of behavior. This analysis was designed to provide an estimate of the representativeness the recordings. Since no differences were found between the behavior frequencies generated by the sequential methods and any of the other recording methods, it is safe to assume that the sequential methods were as "valid" as the other observation methods: they gave an accurate picture of what had transpired during the observation session.

Utility

Results of the present investigation provided substantial evidence that the sequential methods of observation can be a useful research tool for the clinician. This utility is a function of a combination of several advantages the sequential methods present.

Since the utility of each observational method cannot be determined by any absolute criteria, no "utility index" (Dodd & Schultz, 1973) can be computed. Rather, utility must be measured in terms of how practical a particular method is to use. Practicality can be assessed by examining a combination of several factors that were investigated in this study: e.g., training time, reliability, and representativeness; and then weighing the advantages and disadvantages offered by each method along these dimensions.

In terms of the amount of time it took to train observers to rate behaviors reliably, training time on the sequential methods was minimal. Observers could easily learn how to use the sequential methods reliably in one session. This provides a big advantage for the clinician. A lengthy delay between selecting target behaviors for remediation and starting baseline data collection is avoided. Also less man hours are wasted in a lengthy training program.

Training time was also minimal on those methods where the observers recorded one or two behaviors per interval in a continuous fashion throughout the observation session. However, these particular methods do not permit one observer to gather data on as many different behaviors as the sequential methods do. Using these other methods, at the end of an observation session, the observer has sampled only one or two categories of child behavior. Using the sequential method, as many as four categories of behavior were observed during the same period of time.

In addition, sampling two behaviors per observation interval does not always produce reliable recordings of behavior. Reliability in this case seems to be affected by the obvious vs. subtle nature of the target behaviors. For example, given two very obvious (easily discriminable) behaviors to track at one time, (e. g. Out of Chair and Playing) the observer would probably be able to code these behaviors more reliably than given two subtle (difficult to discriminate) behaviors to record at one time. However even under ideal (simple) recording conditions producing highly reliable recordings, this method still provides only half the information the sequential method does.

Method 1, continuous recording of four categories of behavior simultaneously, does provide information about as many behaviors as the sequential method does. However, the drawbacks of Method 1 are obvious. First, training time on Method 1 averaged more than three times as long as training time on the sequential method, an average of 48.4 minutes to 14.8 minutes respectively. Second, the observers used Method 1 about 60% as reliably as the sequential method. During the experimental phase of observations, average reliabilities were 51% to 86% respectively.

In terms of training time, reliability of the data, and cost factors, the sequential methods of observation provide the greatest advantages for the clinician in all of these areas. By having each observer record only one behavior category per interval, high reliability is almost guaranteed, given adequate training, since reliability and simplicity are correlated positively. The cost of observations in terms of both time and funds is minimized by using one observer to rate all of a subject's behaviors of interest, without a loss of accuracy in the data associated with more complex recording procedures. In addition, minimal training time can further reduce the cost of the data collection process and avoid having to wait for excessive periods for observers to become trained before beginning a project.

Another advantage the sequential methods offer the researcher in the applied setting (e. g. working in a classroom) is reducing one methodological problem associated with the reactive nature of the observation process. Reactivity to "being observed" (Patterson & Harris, 1968) may be effected by the obtrusiveness of observers or the number of observers in the observational setting (Johnson & Bolstad, 1973). Using a sequential method necessitates only one observer's being in the setting to collect reliable data (except during those sessions when reliability checks are conducted) the reactive nature of the observation process can be minimized without a concurrent loss of accuracy of the data compared to other recording alternatives where several observers would be required to get a reliable sample (e. g. Method 5). This advantage is particularily valuable to the average clinician researcher who does not have classrooms specially equipped with one way mirrors or videotape recorders in which data could be collected totally without the subjects' knowledge.

The basic model of the sequential method of observation could be generalized to other aspects of the recording situation. For example, if the researcher is interested in one inappropriate behavior emitted by several children in one classroom,one observer could be used to observe the target children in a sequential manner. Another possibility of information on many categories of behavior is desired, an observer could still record the behaviors sequentially, perhaps tracking two categories per interval. Whether reliability levels will be effected by tracking two categories at one time will depend on the nature (obvious vs. subtle) of the behaviors themselves.

Future research should be addressed to determining what sort of behaviors may be observed simultaneously without a concurrent loss of reliablility of the data. Another important issue for further consideration is the effect absolute behavior frequencies may have on the validity of data gathered by intermittent sampling methods. Although there was no evidence of distorted frequency estimates generated by the sequential methods in this investigation, a recent study by Repp _et al_. (in press) indicates that a recording system that has the observers observe behaviors for 20 seconds and record them for 10 seconds of each interval (the type used here) led to an accurate representation of low and medium rates of responding, but grossly underestimated by approximately 60%

high rates of responding. If non-sampling of only 10 of each 30 seconds of observation time could result in such an inaccurate estimate of behavior frequency, it is apparent that sequential sampling of only one of four intervals could lead to a similar distortion of behavior frequencies under certain conditions. Further study is required to determine what absolute rates of behavior are suitable for intermittent, sequential time sampling, and what other behavior rates require a continuous time sampling procedure.

REFERENCES

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

Bandura, A. *Principles of behavior modification*. New York: Holt, Rhinehart, and Winston, 1969.

Bechtel, R. B. The study of man: Human movement and architecture. *Transaction*, 1967, *4*, 53-56.

Callahan, E. J., & Alevizos, P. N. Reactive effects of direct observation of patient behaviors. Paper prepared for presentation at meetings of American Psychological Association, in Montreal, Quebec, Canada, August, 1973.

Cambell, D. T., & Stanley, J. C. *Experimental and quasi - experimental designs for research*. Chicago: Rand McNally, 1966.

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

Dodd, A. & Schultz, S. Computational procedures for estimating magnitude of effect. *Psychological Bulletin*, 1973, *79*, 391-395.

Eckman, T. A. Reducing the cost of obtaining reliability data in applied settings. Paper prepared for presentation at meetings of American Psychological Association, in Montreal, Quebec, Canada, August, 1973.

Fulkerson, S. C., & Barry, J. R. Methodology and research on the prognostic use of psychological tests. *Psychological Bulletin*, 1961, *58*, 177-204.

Goldfried, M. R., & Kent, R. N. Traditional vs. behavioral personality assessment: A comparison of methodological and theoretical assumptions. *Psychological Bulletin*, 1972, *77*, 409-420.

Goldfried, M. R., & Sprafkin, J. *Behavioral personality assessment*. New Jersey: General Learning Press, 1974.

Goodenough, F. L. *Mental testing*. New York: Rhinehart, 1949.

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. Chanpaign, Illinois: Research Press, 1973.

Johnson, S. M., & Lobitz, G. Parental manipulation of child behavior in home observations. Journal of Applied Behavior Analysis, 1974, 7, 23-31.

Jones, R., & Cobb, J. Teachers vs. observers as classroom data collectors. Paper presented at a meeting of the Western Psychological Association, Anaheim, California, April, 1973.

Kass, R. E., & O'Leary, K. D. The effects of observer bias in field experimental settings. Paper presented at a symposium on "Behavior Analysis in Education," University of Kansas, Lawrence, April, 1970.

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.

Kubany, E. S., & Sloggett, B. B. Coding procedure for teachers. Journal of Applied Behavior Analysis, 1973, 6, 339-344.

Lipinski, D., & Nelson, R. O. Problems in the use of naturalistic observation as a means of behavioral assessment. Behavior Therapy, 1974, 5, 341-351.

Mash, E. J., & Hedley, J. Observer effect as a function of prior history of social interaction. Perceptual and Motor Skills, 1975, 40, 659-669.

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.

Mischel, W. Personality and Assessment. New York: Wiley, 1968.

O'Leary, K. D., & 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. Illinois: 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.

O'Leary, K. D., Romancyzk, R. G., Kass, R. E., Dietz, A., & Santogrossi, D. Procedures for classroom observation of teachers and children. Unpublished manuscript. State University of New York at Stony Brook, 1971.

Patterson, G. R., & Cobb, J. A. A dyadic analysis of "aggressive" behaviors. In J. P. Hill (Ed.), *Proceedings of the Fifth Annual Minnesota Symposium on Child Psychology* (Vol. 5). Minneapolis: University of Minnesota, 1971.

Patterson, G. R., & Harris, A. Some methodological considerations for observation procedures. Paper presented at the meeting of the American Psychological Association, San Francisco, California, September, 1968.

Patterson, G. R., Ray, R. S., Shaw, D. A., & Cobb, J. Manual for coding of family interactions, 1969. Available from ASIS/NAPS, c/o Microfiche Publications, 305 E. 46th Street, New York, New York 10017. (Document #01234)

Polansky, N., Freeman, W., Horowitz, M., Irwin, L., Papanis, N., Rappaport, D., & Whaley, F. Problems of interpersonal relations in research groups. *Human Relations*, 1949, *2*, 281-291.

Rapp, D. W. Detection of observer bias in the written record. Cited in R. Rosenthal, *Experimenter effect in behavioral research*. New York: Appleton Century Crofts, 1966.

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

Reid, J. B. The relationship between complexity of observer protocols and observer agreement for twenty five reliability assessment sessions. In preparation, 1973.

Reid, J. B., & 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.

Reid, J. B., Skindrud, Taplin, P. S., & Jones, R. R. The role of complexity in the collection and evaluation of observation data. Paper prepared for the meetings of the American Psychological Association, Montreal, Quebec, Canada, August, 1973.

Repp, A. C., Berkler, M. S., Roberts, D. M., Slack, D. J., & Repp, C. F. A Comparison of frequency, interval, and time sampling methods of recording. *Journal of Applied Behavior Analysis*, in press.

Repp, A., Deitz, D., Boles, S., Dietz, S., Repp, C. Differences among common methods for calculating inter-observer agreement in applied behavioral studies. *Journal of Applied Behavior Analysis*, 1976, *9*, 109-113.

Roberts, R. R., & Renzaglia, G. A. The influence of tape recording on counseling. *Journal of Counseling Psychology*, 1965, *12*, 10-16.

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.

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

Skindrud, K. An evaluation of observer bias in experimental field studies of social interaction. Unpublished doctoral dissertation, University of Oregon, 1972.

Skindrud, K. Field evaluation of observer bias under overt and covert monitoring. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology, concepts, and practice. Illinois: Research Press, 1973.

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

Thomson, C., Holmberg, M. & Baer, D. A brief report on a comparison of time sampling procedures. Journal of Applied Behavior Analysis, 1975, 7, 623-626.

White, G. D. Effects of observer presence on family interaction. Paper presented at the meeting of the Western Psychological Association, Anaheim, California, April, 1973.

Winer, B. J. Statistical principles in experimental design. New York: McGraw - Hill, 1971.

O'Leary, K. D., Kaufman, K. F., Kass, R., & Drabman, R. The effects of loud and soft reprimands on the behavior of disruptive students. Exceptional Children, 1970, 37, 145-155.

Appendix A

Observation Sheet Used for Method 1

O P T @	O P T @	O P T @	O P T @
O P T @	O P T @	O P T @	O P T @
O P T @	O P T @	O P T @	O P T @
O P T @	O P T @	O P T @	O P T @
O P T @	O P T @	O P T @	O P T @
O P T @	O P T @	O P T @	O P T @

Appendix A (Continued)

Observation Sheet Used for Variation 1 of Method 2

O P	O P	O P	O P
O P	O P	O P	O P
O P	O P	O P	O P
O P	O P	O P	O P
O P	O P	O P	O P
O P	O P	O P	O P

Appendix A (Continued)

Observation Sheet Used for Variation 2 of Method 2

T @	T @	T @	T @
T @	T @	T @	T @
T @	T @	T @	T @
T @	T @	T @	T @
T @	T @	T @	T @
T @	T @	T @	T @

Appendix A (Continued)

Observation Sheet Used for Method 3

O	P	T	@
O	P	T	@
O	P	T	@
O	P	T	@
O	P	T	@
O	P	T	@

Appendix A (Continued)

Observation Sheet Used for Variation 1 of Method 4

O	P	O	P
O	P	O	P
O	P	O	P
O	P	O	P
O	P	O	P
O	P	O	P

Appendix A (Continued)

Observation Sheet Used for Variation 2 of Method 4

T	@	T	@
T	@	T	@
T	@	T	@
T	@	T	@
T	@	T	@
T	@	T	@

Appendix A (Continued)

Observation Sheet Used for Variation 1 of Method 5

o	o	o	o
o	o	o	o
o	o	o	o
o	o	o	o
o	o	o	o
o	o	o	o

Appendix A (Continued)

Observation Sheet Used for Variation 2 of Method 5

P	P	P	P
P	P	P	P
P	P	P	P
P	P	P	P
P	P	P	P
P	P	P	P

Appendix A (Continued)

Observation Sheet Used for Variation 3 of Method 5

T	T	T	T
T	T	T	T
T	T	T	T
T	T	T	T
T	T	T	T
T	T	T	T

Appendix A (Continued)

Observation Sheet Used for Variation 4 of Method 5

@	@	@	@
@	@	@	@
@	@	@	@
@	@	@	@
@	@	@	@
@	@	@	@

Appendix B

Disruptive Behavior Code

Out of Chair --- symbol = 0

Purpose: Out of chair is intended to monitor the gross motor behavior of the child removing himself from his seat entirely. Such behavior (when not permitted) may interfere with the child's learning and is potentially distracting to others e. g., running around the room.

Description: Observable movement of the child from his chair when not permitted or requested by teacher. None of the child's weight is to be supported by the chair, but the child may be in physical contact with chair.

Critical Points: None of the child's weight is to be supported by the chair.

Includes: Child is leaning on desk and has either lost all contact with the chair or none of his weight is actually being supported by the chair.

Time limits on the following beginning with teacher's permission. Allow 15 seconds for a child to get from the teacher's desk to his own. Allow 15 seconds for a child to return to his own seat after completing a task (i. e., placing a word card on the wall). Pencil sharpening 1½ minutes. Getting a drink - 1½ minutes (fountain.in room). Getting a book - 1½ minutes

Appendix B (Continued)

(time limit starts from the second that the child gets out of seat). Going to the bathroom: (a) 2 minute limit, (b) 30 second limit beginning when child leaves bathroom.

Note: If the child returns to the chair after 1½ (or 2 minutes, where applicable), but during the 10 second inter-interval period, the "0" will be recorded in the 20 second interval just prior to the 10 second interval.

Going to get a reading book during a math lesson. When child is fully standing and the back of legs touch chair, or child is fully standing and is touching back of chair with hands. Going to teacher's desk when not permitted. Throwing away papers. Stretching (if child actually leaves seat).

Excludes: Retrieval of an accidentally dropped task-related object. Leaning forward to pick up an object even if all contact with the chair is momentarily lost, providing the child is not standing fully erect on feet. Include if child begins crawling around on floor after retrieving

Appendix B (Continued)

object, also, include if child is moving from desk in a crouched position, so as not to let the teacher see him, etc.

Appendix B (Continued)

Playing --- symbol = P

Purpose: Playing is intended to monitor often subtle manipulative behavior that is distracting to the child and possibly also distracting to others.

Description: Child uses his hands to play with his own or community property, so that such behavior is incompatible (or would be incompatible) with learning.

Critical Points: Child uses his hands to manipulate his own or community property.

Includes: Playing with toy car when assignment is spelling. Playing with comb or pocket book. Eating only when the hands are being used - chewing gum is not rated as P unless child touches or manipulates it with his hands. Poking holes in workbook. Cleaning nails with pencil in such a manner as to make the behavior incompatible with learning e. g., shoving pencil back and forth on desk; waving pencil through air as an airplane. Picking scabs, nails, or nose if the desired "object" is separated from the body and manipulated. Looking into desk and moving arms, but does not come out with a task-related object. Working with or reading non-task related material e. g., reading page 25 when told to read page 1, doing math when told to do spelling, etc.

Appendix B (Continued)

Excludes: Touching others' property. Playing with own clothes.

Note: Include if article is removed from body, e. g., shoes, tie, buttons, scarf, etc., and is manipulated.

Lifting desk or chair with feet (rate N if this creates audible noise). Random banging of pencil on desk (rate N if audible). Simple twiddling pencil if it is not seen as being incompatible with learning.

Note: Rate twiddling pencil, banging pencil, or putting pencil in mouth, hair, behind ear, etc., if child attends to such behavior and ceases attending to assigned task. Operational definition of attending: child either looks at manipulated object or begins to manipulate object in non-random patterns for more than 5 seconds.

Picking scabs, nails, or nose if the desired "object" is not separate from the body.

Appendix B (Continued)

Time - off - task --- symbol = T

Purpose: Time - off - task is intended to monitor non-attending behavior, that, if excessive, is detrimental to child's performance.

Description: Child does not do assigned work for entire 20 second interval.

Critical Points: Child makes no attending response for the entire 20 second interval. Child must only attend i. e., "look at", his work. Inferences that, "he isn't really thinking about it", are not acceptable.

Includes: Child does not write when so assigned. Child does not read when so assigned. Child is working on inappropriate material e. g., on math during spelling, etc. Daydreaming - as reflected in not working. Child does not ask teacher for additional work or help when finished with assigned task, and merely sits at desk or begins to play for entire interval. When in corner, child's head must be within a 45 degree angle from the corner formed by 2 walls (e.g., if his head is facing either of the 2 walls directly, for a 20 second period, he would be rated X).

Excludes: Child has his hand raised to ask questions. Child is told he may cease working if he so desires.

Appendix B (Continued)

Orienting Response --- symbol = @

Purpose: Orienting is intended to monitor the gross motor behavior of turning around from the designated point of reference. Such behavior is distracting to child since it usually precludes attending to assigned task, and is often distracting to others.

Description: Child turns more than 90 degrees from point of reference while seated.

Critical Points: The child must be in his seat; he may be in a modified position; and orienting includes both the horizontal and vertical axis.

Includes: Turning to the person behind. Looking to the rear of the room. Turning around in chair or turning chair around. Leaning back in chair more than 90 degrees.

Note: Point of reference is typically child's desk but may be the teacher if the children are instructed to attend to her. If child should turn desk at some angle, point of reference becomes where desk was originally, not to where the child has moved it. Also, the child's chin should be used as the indicator of how far he has turned. Therefore orienting is rated when child's chin has turned more than 90 degrees from point of reference.

Appendix B (Continued)

Excludes: Orienting during class discussion when the teacher directs (either implicitly or explicitly) the class to attend to a child's explication of an answer.

Orienting while picking up a task related object.

When child is in corner or otherwise out of his chair.

Appendix C

Statistical Tables

TABLE 1

Analysis of Variance on Training Time Required for Observers to Become Proficient on the Five Methods of Observation

Source	df	MS	F
Methods of Observation	4	2426.08	83.0091*
Error	45	29.23	

* $p < .001$

Appendix C (Continued)

TABLE 2

Neuman-Keuls Comparison of Mean Training Time

	METHOD						
	1	2	3	4	5	r	C. V. for .05
1		33.2*	33.6*	35.6*	36.4*	5	8.43
2			0.4	2.4	3.2	4	8.03
3				2.0	2.8	3	7.47
4					0.8	2	6.53
5							

* $\underline{p} < .05$

Appendix C (Continued)

TABLE 3

Multivariate Analysis of Covariance on Reliabilities at the End of Training Considering All Behaviors Simultaneously

Source	df	MS	F
Method of Observation	4	.3206	14.428*
Error	45	.022	

*$p < .001$

Appendix C (Continued)

TABLE 4

Neuman-Keuls Comparison of Mean Reliability at the End of Training Considering Behaviors Simultaneously

	METHOD						
	3	4	1	2	5	r	C. V. for .05
3		.01	.15	.16	.45*	5	.20
4			.14	.15	.44*	4	.19
1				.01	.30*	3	.17
2					.29*	2	.14
5							

* $p < .05$

Appendix C (Continued)

TABLE 5

Univariate Analysis of Covariance on the Reliability at the End of Training with Which Playing Behaviors Were Recorded

Source	df	MS	F
Method of Observation	4	.214	1.30578*
Error	45	.164	

*$\underline{p} < .28$

Appendix C (Continued)

TABLE 6

Univariate Analysis of Covariance on the Reliabilities at the End of Training with Which Orienting Behaviors Were Recorded

Source	df	MS	F
Method of Observation	4	.4121	2.0249*
Error	45	.2035	

*$p < .11$

Appendix C (Continued)

TABLE 7

Univariate Analysis of Covariance on the Reliability at the End of Training with Which Out of Chair Behaviors Were Recorded

Source	df	MS	F
Method of Observation	4	.97	12.63347*
Error	45	.08	

*$p < .0001$

Appendix C (Continued)

TABLE 8

Neuman-Keuls Comparison on the Mean Reliability at the End of Training with Which Out of Chair Behaviors Were Recorded

	METHOD					
	4	3	2	1	5	r	C. V. at .05
4		0.0	.15	.15	.75*	5	.20
3			.15	.15	.75*	4	.19
2					.60*	3	.17
1					.60*	2	.14
5							

*$p < .05$

Appendix C (Continued)

TABLE 9

Univariate Analysis of Covariance on the Reliability at the End of Training with Which Time Off Task Behaviors Were Recorded

Source	df	MS	F
Method of Observation	4	1.13	6.82978*
Error	45	.166	

*$p < .0004$

Appendix C (Continued)

TABLE 10

Neuman-Keuls Comparison on the Mean Reliability at the End of Training with Which Time Off Task Behaviors Were Recorded

	METHOD						
	3	4	1	5	2	r	C. V. at .05
3		.21	.59*	.73*	.75*	5	.20
4			.38	.52*	.54*	4	.19
1				.14	.16	3	.17
5					.02	2	.14
2							

*$p < .05$

Appendix C (Continued)

TABLE 11

Multivariate Analysis of Covariance on Reliabilities During the Experimental Phase of Observation Considering All Behaviors Simultaneously

Source	df	MS	F
Method of Observation	4	.334	13.69473*
Error	41	.024	

$*p < .001$

Appendix C (Continued)

TABLE 12

Neuman-Keuls Post Hoc Comparison of Mean Reliability During the Experimental Phase of Observation Considering All Behaviors Simultaneously

				METHOD			C. V. at
	3	5	4	2	1	r	.05
3		.04	.14	.24*	.51*	5	.20
5			.10	.20*	.47*	4	.19
4				.10	.37*	3	.17
2					.27*	2	.14
1							

* $\underline{p} < .05$

Appendix C (Continued)

TABLE 13

Univariate Analysis of Covariance on the Reliability with Which Out of Chair Behaviors Were Recorded During the Experimental Phase of Observations

Source	df	MS	F
Method of Observation	5	.979	2.39217*
Error	44	.409	

*$p < .05$

Appendix C (Continued)

TABLE 14

Neuman-Keuls Comparison on the Mean Reliability with Which Out of Chair Behaviors Were Recorded During the Experimental Phase of Observations

	METHOD						
	3	5	4	2	1	r	C. V. at .05
3		.19	.31	.43	.93*	5	.8
5			.02	.14	.64	4	.76
4				.12	.62	3	.69
2					.50	2	.57
1							

*p < .05

Appendix C (Continued)

TABLE 15

Univariate Analysis of Covariance on the Reliability with Which Playing Behaviors Were Recorded During the Experimental Phase of Observations

Source	df	MS	F
Method of Observation	5	2.3155	5.56344*
Error	44	.4162	

*p < .0007

Appendix C (Continued)

TABLE 16

Neuman-Keuls Comparison on Mean Reliability with Which Playing Behaviors Were Recorded During the Experimental Phase of Observations

	METHOD						
	3	5	4	2	1	r	C. V. at .05
3		.09	.55	.58	1.34*	5	.824
5			.46	.49	1.25*	4	.773
4				.03	.79*	3	.702
2					.76*	2	.583
1							

*$\underline{p} < .05$

Appendix C (Continued)

TABLE 17

Univariate Analysis of Covariance on the Reliability with Which Time Off Task Behaviors Were Recorded During the Experimental Phase of Observations

Source	df	MS	F
Method of Observation	5	3.6073	4.46819*
Error	44	.8073	

*$\underline{p} < .0025$

Appendix C (Continued)

TABLE 18

Neuman-Keuls Comparison on Mean Reliability with Which Time Off Task Behaviors Were Recorded During the Experimental Phase of Observations

				METHOD			
	5	3	4	2	1	r	C. V. at .05
5		.23	.34	.92	1.54*	5	1.147
3			.11	.69	1.31*	4	1.076
4				.58	1.20*	3	.977
2					.62	2	.812
1							

*$p < .05$

Appendix C (Continued)

TABLE 19

Univariate Analysis of Covariance on the Reliability with Which Orienting Behaviors Were Recorded During the Experimental Phase of Observations

Source	df	MS	F
Method of Observation	5	1.2303	4.20741*
Error	44	.2924	

*$p < .0036$

Appendix C (Continued)

TABLE 20

Neuman-Keuls Comparison on the Mean Reliability with Which Orienting Behaviors Were Recorded During the Experimental Phase of Observations

			METHOD				
	5	3	4	2	1	r	C. V. at .05
5		.01	.3	.46	.94*	5	.69
3			.29	.45	.93*	4	.64
4				.16	.64*	3	.58
2					.48	2	.49
1							

*$p < .05$

Appendix C (Continued)

TABLE 21

Multivariate Analysis of Variance on the Frequency with Which Each Behavior Was Recorded Considering All Behaviors Simultaneously

Source	df	MS	F
Method of Observation	4	.0127	2.47206*
Error	195	.0051	

*p < .10

Appendix C (Continued)

TABLE 22

Univariate Analysis of Variance on the Frequency with Which Out of Chair Behaviors Were Recorded

Source	df	MS	F
Method of Observation	4	.21	.78271*
Error	195	.27	

*$\underline{p} < .5398$

Appendix C (Continued)

TABLE 23

Univariate Analysis of Variance on the Frequency with Which Playing Behaviors Were Recorded

Source	df	MS	F
Method of Observation	4	.195	.31730*
Error	195	.616	

*$p < .8665$

Appendix C (Continued)

TABLE 24

Univariate Analysis of Variance on the Frequency with Which Time Off Task Behaviors Were Recorded

Source	df	MS	F
Method of Observation	4	.04	.18715*
Error	195	.22	

*$\underline{p} < .9428$

Appendix C (Continued)

TABLE 25

Univariate Analysis of Variance on the Frequency with Which Orienting Behaviors Were Recorded

Source	df	MS	F
Method of Observation	4	.256	1.01188*
Error	195	.253	

*$\underline{p} < .4033$