INTEROBSERVER VARIATION IN RECORDING BEHAVIOR: RANDOM OR SYSTEMATIC ERROR?

A Dissertation

Presented to

the Faculty of the Department of Psychology

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Harry Hull

August, 1970

INTEROBSERVER VARIATION IN RECORDING BEHAVIOR: RANDOM OR SYSTEMATIC ERROR?

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Psychology

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Harry Hull

August, 1970

INTEROBSERVER VARIATION IN RECORDING BEHAVIOR: RANDOM OR SYSTEMATIC ERROR?

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Psychology

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Harry Hull

August, 1970

ABSTRACT

The number of investigations using direct behavioral data has increased in frequency during the last ten years. The interest in behavioral data has focused attention on observation methodology, and emphasized such variables as objectifying data language, training of observers, and using more refined methods of calculating interobserver reliability. The emphasis on these variables assumes observer homogeneity and does not acknowledge systematic differences among observers. Systematic observer bias, when acknowledged, is given only cursory attention or subsumed under recognized forms of observer bias such as knowledge of the experimental hypothesis. This study questioned the assumption of observer homogeneity and hypothesized that observers do show systematic differences in recording behavior events--differences which are related to how an observer indicates he would respond to (evaluate) such behavior.

Twenty-six Ss were categorized on the basis of their preferred mode of response (positively consequate, negatively consequate, or extinguish) to seven classes of behavior. Subsequently the Ss received familiarization training on

coding procedures and then coded the seven behavior classes for four kindergarten age males in four seperate coding sessions. The training and coding sessions utilized video tapes with an audio tone signalling alternate 10-sec observation periods. Each observer's coding records were compared with a coding standard (5 professionals' independent agreements on behavioral occurrence-nonoccurrence) to obtain a deviancy score which indexed the extent and direction of disagreement with the coding standard. Interobserver reliability was also calculated for gross reliability (a frequency ratio between two observers for an observation session) and agreement (paired observers' agreements divided by agreements plus disagreements for 10-sec observation periods within an observation session). An accuracy score (an agreement index with an observer's coding record compared with the coding standard for agreements-disagreements) was also obtained to determine the relative validity of the observers' coding records and the reliability indices.

When deviancy scores were compared for behavior classes the observers would positively consequate, negatively consequate, and extinguish, there were bias trends. However, wide variation in behavior rate for the seperate codes tended to mask systematic bias by the observers. The coding standard

indicated that the three ways of consequating behavior were comparable only for low behavior rates. When behavior rate was held constant for low rates, there were significant bias effects as well as strong evidence for response style (systematic bias across data sources). In the absence of ongoing behavior (according to the coding standard) behaviors positively consequated by the observers were significantly overrecorded, behaviors extinguished by the observers were intermediate in overrecording bias, and behaviors negatively consequated showed minimal overrecording bias by the observers. When behavior rate increased to one ongoing behavior per observation period, overrecording bias was reduced for behaviors the observers positively consequated or extinguished although the positively consequated behaviors were still overrecorded by the observers. The decrement in observer bias as behavior rate increased was consistent with other data which indicated that behavior rate and bias are inversely related for behaviors which are positively consequated and extinguished while behaviors which are negatively consequated show minimal bias across behavior rates. These results suggest that evaluative differences among observers are an important variable in behavior coding variability and thus the assumption of observer homogeneity can be rejected.

Of the methods of calculating interobserver reliability, the agreement index proved to be a more valid index than gross reliability. Gross reliability provided an inflated index on all behavior classes for most of the observation sessions. The validity of the gross reliability index is therefore seriously questioned and it probably should be discontinued as an index of interobserver reliability. The agreement index was most valid for codes which were negatively consequated, was marginal for extinction codes, and proved to be a conservative estimate of data validity for positively consequated codes. These differences in relative validity were directly related to the magnitude and direction of recording bias for the observers. The agreement index is, however, a useful estimate of interobserver reliability and data validity since any error is in a conservative direction.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

In recent years there has been an increased emphasis in child psychology on obtaining direct behavioral observations in both laboratory and field settings. Observation technology has been applied to studies dealing with parent-child interactions (Bernal, Duryee, Pruett, & Burns, 1968; Hawkins, Peterson, Schweid, & Bijou, 1966; Yarrow, 1963), parent training (O'Leary, O'Leary, & Becker, 1967; Patterson & Brodsky, 1966; Straughen, 1965; Wahler, Winkel, Peterson, & Morrison, 1965), classroom behavior (Charlesworth & Hartup, 1966; Buehler, Patterson, & Furniss, 1966; Patterson, Littman, & Bricker, 1967) and a number of other clinical and research investigations.

The interest in behavioral data and observational techniques has focused attention upon existing observational methodology (Wright, 1960) and on the development of a more sophisticated methodology (Bijou, Peterson, & Ault, 1968; Patterson & Harris, 1968; Wright, 1967). However, while observational procedures have become more popular, they contain a number of methodological problems. These revolve

around obtaining improved data validity through an increase in interobserver reliability--reliability which has been difficult to achieve in many studies using behavior classes. Some approaches to the problems of interobserver reliability, and subsequent inferences about data validity, consist of objectifying the data language, training of observers, and refining methods of calculating interobserver reliability (Bijou, et al., 1968); controlling for observer bias (Rosenthal, 1966); analyzing the effect of the observer upon observed behavior (Patterson & Harris, 1968); and formulating questions concerning the generality of observational data (Cronbach, Gleser, Nanda, & Rajaratnam, 1967). The methodological problem of immediate concern in this study was that of systematic influences affecting interobserver reliability --a methodological issue which has been largely ignored in the observational literature. It was the thesis of this study that variance between observers (leading to lowered interobserver reliability) contains a strong systematic component which is related to the observers' personal evaluation of the behavior under observation.

CHAPTER II

SURVEY OF THE LITERATURE ON OBSERVATIONAL PROCEDURES

The basic requirement for descriptive observations in scientific research or naturalistic observation is that such descriptions be devoid of surplus meaning as much as possible, and that the data language of observations be such that minimally qualified persons would agree on the descriptive content (Butler, Rice, & Wagstaff, 1962; Loevinger, 1965). In short, the reliability of any behaviorally descriptive system is largely a function of a lack of ambiguity in the data language interacting with a low margin of observer error (Maher, 1970).

The components of this requirement, objective description of behavior and observer agreement, are the primary criteria for the raw data of a behavioral science. They serve to give the investigator an objective record of what has occurred and to minimize, as much as possible, the subjective, personal biases of the observer. Objective descriptions of behavior and observer agreement thus provide the basis for an unambiguous description of behavior, the

determination of functional relationships between behavior and environmental events through experimentation, and the formulation of a surrounding theoretical structure.

In an attempt to decrease data ambiguity, investigators have tended to reduce the descriptive terminology to physical terms wherever possible, or minimally to operationalize data language in clearly defined behavioral referents. Indeed, Maher (1970) suggests that for observations to be labeled as such, it is necessary that they be recorded in the physical realm; all else is inferential. From such observations, inferences are derived to be verified by empirical testing. Thus direct behavioral-physical observations are considered to be the primary methodological root in the development of psychology as a science. As the data language deviates from this primary criterion, the reliability and subsequent validity of inferences drawn from the raw data decreases.

Interobserver agreement becomes necessary when behaviors are not capable of direct mechanical measurement and/or when judgements as to the occurrence or nonoccurrence of a response are required. High agreement among observers (high reliability coefficients) is assumed to be a reasonable confirmation that the response has or has not occurred. However, as Ayllon and Azrin (1968) have pointed out, observer agreement

can occur on the basis of shared subjective interpretations ("social bias") and indicate little or nothing about the physical basis of the observed behavior. In order to avoid such subjective influences, observational data are made as objective and specific as possible so that inferential factors and the observer's personal evaluations do not substantively influence the data.

Variation among observers in recording observed behavior may be attributed to three general sources of error: random error, systematic bias, and response style (Grosz & Grossman, 1968). The first, random error, is suggested if interobserver variation is not significantly different. Here, interobserver variation refers to differences between observers on the same data across observations. Variables such as poor behavioral specificity, inadequate observer training, and inappropriate reliability indices would contribute to random error variance.

Systematic bias is indicated if there are significant interobserver differences across observations. Here, observers are apparently using different criteria to rate the behavior, or record its occurrence or nonoccurrence. Detection of systematic bias does not specify the functional cause since such a bias could be the result of differential training,

different social biases of the observers, or any number of operative conditions.

The third source of error, response style, is suggested if the same systematic interobserver bias occurs across data sources. Thus if Observer A records a greater frequency of behavior X in a sample of situations than does Observer B, one could infer that a response style is operative for observers A and B for the given behavior or behavior class. Detection of a response style, like detection of systematic bias, does not pinpoint the functional cause. Any number of plausible conditions could account for such observer discrepencies.

Among behaviorally oriented psychologists, the most common variable associated with high interobserver agreement is the specification of the behavior to be observed (objective, physically directed data language). Such investigators are inclined to posit that interobserver reliability is a direct function of data specificity (cf. Ayllon & Azrin, 1968; Bijou, et al., 1968). As indicated earlier however, ambiguity in the data language interacts with observer agreement. While making the observations in objective data language may serve to reduce observer error to some extent, the relationship does not appear to be a direct linear one. This is made evi-

dent by numerous studies in the literature which report high interjudge or interobserver reliabilities on often substantively inferential data systems (Goldberg, 1968). In addition, Bijou, et al. (1968) report that when a behavior is adequately specified, interobserver agreement can still vary. The cases of high agreement among judges on inferential material and lowered agreement on some objective behavior codes would sharpely reduce any contention that interobserver reliability is directly related to the specificity of the data language. The determination of how much variance such a variable accounts for remains an open question.

Additional sources of interobserver variance, as reported by Bijou, et al. (1968) are observer training and the method of calculating reliability. It is assumed that if a behavioral code is servicable, observer training will firm up most of the gap in interobserver variance left after specifying behavior. Observer training generally consists of two steps. The first is familiarization with the observational code and coding materials (stop watch, coding sheet and symbols etc.). Coding apparatus or materials vary from paper and pencil recording (Madsen, Becker, & Thomas, 1968; Patterson, 1967) to sophisticated electro-mechanical recorders (Lovaas, Freitag, Gold, & Kassorla, 1965; Wahler, 1967).

The second step consists of training in behavioral control (attending to the behavior). The latter typically involves a formal or informal feedback program in which disagreements between observers' recordings are pointed out. Occasionally, discussion is involved to ascertain the basis for agreements and disagreements in order to clear up coding ambiguities or false assumptions of the observers. Observer training is continued until agreement among observers is 80% or better for all behaviors or behavior classes to be observed. Where small changes in behavior are anticipated, the interobserver reliability criterion is usually 90% or better.

It is not uncommon, however, that training procedures are not delineated in published material. Wahler, et al. (1965), as an example, specify the behavior coding system used and the time samples for calculation of interobserver agreement, but summarize observer training with the statement that "Observer agreement of ninety per cent or better was considered to be adequate; once this agreement was obtained on all behavior classes the baseline sessions were begun [p. 116] ." Such an offhanded description of training procedure would not be critical if it could be assumed that observers were drawn from a homogeneous observer population and standard training criteria were applied. The assumption

of a homogeneous observer population may be seriously questioned, and the only statement of training criteria is the level of acceptable agreement.

With respect to training criteria, it is legitimate to question whether all behavior classes required the same amount of training to reach 90% agreement. If not, one could argue that data language specification of the behavior classes may not have been equivalent. Given nonequivalence, the additional training required for some classes of behavior could well be functionally related to the introduction of idiopathic behavioral referents for these categories by the investigators and/or observers. Such unique judgements would impair not only interobserver reliability at later points in observational samples, but also replication efforts by other investigators using similiar behavior classes. Even if parallel replication were obtained, there would be an open question of whether the implicit behavioral referents were the same. It would appear then, that just on the basis of behavioral specification criteria, stating the level of agreement between observers as a training criterion is not completely acceptable. Stating the preconditions of reaching the interobserver criterion would seem necessary if methodological error is not to be replicated. The basic criteria of a science of behavior

--objective language with a minimum of surplus meaning, and interobserver agreement--is therefore given only cursory attention by such incomplete descriptions.

The other source of interobserver variance as stated by Bijou, et al. (1968) is that of the method of calculating reliability. Typically, interobserver reliability is calculated by forming a ratio between two observers' frequency records of an event for a given time period. If the summed frequency over this time period is equal, the reliability index is 100%. When sums are unequal, the smaller is divided by the larger to get a percentage of agreement (or reliability index). However, the gross nature of this reliability index may provide little or no indication of data reliability. For example, if the time sample is five minutes, a 100% reliability index could be obtained from observers recording, by category, an equal number of different behaviors. Thus perfect interobserver reliability could be obtained with little or no agreement on the behaviors being coded. Procedurally, Bijou, et al. (1968) suggest that if agreements are obtained over progressively smaller time samples, then confidence increases that observers are recording the same event. Thus if the observation period is broken down into smaller time samples such as five or ten seconds, then data

reliability increases. Such time samples have been used for both paper and pencil records (O'Leary, et al., 1967; Patterson & Reid, 1969) and for electro-mechanical recordings (Wahler, 1967; Wahler, et al., 1968). However, it would seem that the necessity of using small time samples with electro-mechanical equipment would be less since such data can be correlated by category and time.

Bijou, et al. (1968) also suggest that the technique of noncontinuous observation (e.g., record for 10 out of 20 seconds) may increase observer reliability as such nonobservation periods give the observer an opportunity to record and thus leave the observation period maximally free to attend to object behaviors.

Behavior coding specificity, observer training, and the choice of observation time samples are thus considered to be major factors contributing to interobserver variability, with behavioral specification being given the most weight. Treatment of such variables carries with it several basic assumptions. All three variables assume that interobserver variance is randomly assigned and that pruning on these variables will reduce such random variation. Systematic response biasing by the observer is not explicated, and when briefly acknowledged is generally considered a function of poorly de-

fined behavioral referents. The more basic assumption is that when interobserver response variability is considered a function of the three variables just outlined, observers are implicitly treated as part of a homogeneous population. These assumptions would at face value appear invalid, and as such would impose constraints on current observation technology.

Related to the assumption of observer homogeneity is the consideration by Cronbach, et al. (1967) that for data to be generalizable, the contributions of behaviors, observers, and environmental settings must be systematically sampled. For example, when only two observers, a first grade classroom, and the class of gross motor behaviors (running, jumping etc.) are included in a research design, there are constraints on the generality of the obtained data. When high agreement is obtained by two observers in such a context with a given class of behavior, it cannot be assumed that other observers would reach agreement within the same context or behavior class, or that the two observers would agree in other contexts or with other behavioral classes. When erroneous sampling assumptions of inter and intraobserver homogeneity across environmental and behavioral dimensions are made, the investigator runs the minimum risk of

introducing random error components and in all likelihood of introducing correlated errors of systematic bias or response style. As Campbell and Fiske (1959) have pointed out, random errors may decelerate observed relationships, but will not distort them as would a systematic error component.

An extension of the sampling consideration proposed by Cronbach et al. (1967) is contained in two additional methodological problems in observation technology--problems not directly related to interobserver variance. These concern the questions of the amount and kind of behavior sampling necessary for stable estimates of the frequency of behavioral classes, and whether the presence of an observer affects the behavior under observation (Patterson & Harris, 1968). Some investigators have already drawn conclusions about the extent of sampling required for stable estimates. For example, Ogden Lindsley (personal communication, 1968) requires at least five to seven data points before experimental variables are introduced. Patterson and Harris (1968) emphasize that while it is possible to establish stable estimates for some behavioral categories, questions concerning the minimum amount of sampling necessary for stable estimates, and how such estimates vary across environmental settings have not been clarified.

While investigation into the effect of the observer on observed behavior is still in its preliminary stages, Patterson and Harris (1968) tentatively conclude that though the presence of an observer can initially inflate or deflate the rate of observed behavior, there will be a "regression to the mean" effect if the behavior is observed over a period of time. In other words, subjects habituate to neutral investigatory stimuli if the overall time sample is adequate.

Sampling considerations underlie many of the major objections to current observational methodology, and are a significant component of interobserver variance--the dimension of interest in this study. It was suggested earlier, for example, that observers can share a common social bias which may indicate little or nothing about the physical basis of the observed behavior. For any given set of observers, environmental settings, or classes of behavior, a common social bias could yield a relatively high interobserver agreement index, and with observer training any random variation due to coding definitions, experience with the coding materials etc. would be reduced. The result would be an impressive, yet systematically biased agreement index which may not faithfully reflect the observed behavior or allow the investigator to assume he has valid data. Conversely, the rela-

tive absence of a common social bias between observers (or presence of opposing biases) may necessitate extensive training and/or data language reduction if interobserver agreement criteria are to be achieved.

Personal biases affecting the description and interpretation of observed data have a long and sustained representation in the history of psychology with individual variability in responding to relatively standard stimuli being the subject of both experimental and clinical interest. Response bias on questionnaires and rating scales is now taken as a given, and response style has been investigated under a variety of labels such as acquiescence, social desirability etc. (Goldberg, 1968; Rorer, 1965). As Rorer (1965) indicates, the literature on response bias and response set has grown so large that even the reviews have been reviewed.

It has only been within recent years however, that clinicians have looked to themselves or their procedures as contributors to data distortion. As has been pointed out by Rosenthal (1966), it may well be taken as fact that experimenter-observers obtain data more, not less, in accordance with their expectancies and personal biases. Such cases of personal bias are now controlled for in many experimental designs through such techniques as uninformed experimenters,

counterbalancing etc. As yet, however, the systematic influence of the investigator and/or observer has had minimal impact on observational technology.

One of the few behavioral studies to examine the effects of experimenter bias is that of Scott, Burton, and Yarrow (1967) who essentially confirm the conclusions drawn by Rosenthal. From observations, which took the form of the "stream of behavior" record, it was found that an informed observer produced a behavioral record which was more in the direction of the experimental hypothesis than the records of the uninformed observers. The investigators state that such supposed biasing did not alter the results of their study, but only affected the degree of change. The investigators even offer as an alternate explanation to experimental bias the possibility the informed observer may have been more "sensitive" to the relevant behavior and therefore produced more "precise" records. It would appear that had the investigators not had the uninformed observers as a comparison group, their results might have been interpreted less cautiously. The alternative explanation of the informed observer being a more "sensitive" recorder is a plausible hypothesis, but would have to be confirmed by means of some objective record, such as video taping or film media.

Experimenter bias, as it is commonly conceived, is however, only a small part of the interobserver variance arising from systematic biasing sources. Left unaccounted for is the entire range of attitudes or personal biases the experimenter and/or observer may have toward the specific contexts of the environment, the behavior class, or their interactions. It would appear to be almost an axiom to state that every observer has developed in the course of his individualized learning experiences, a set of implicit or explicit evaluations for many behaviors in most settings. As Ellis (1962) succinctly puts it:

> An individual evaluates (attitudinizes, becomes biased) when he perceives something as being 'good' or 'bad', 'pleasant' or 'unpleasant', 'beneficial' or 'harmful' and when, as a result of his perceptions he responds positively or negatively to this thing. Evaluating is a fundamental characteristic of human organisms...[p. 44].

It would be unreasonable to expect that an individual would discard such biases when functioning in the role of an observer, even though the observational record is not an evaluative one. Positive and negative 'halo' effects on observational records as a function of individualized attitudes could therefore be a major factor in interobserver variance after data specification, agreement indices etc. have been given full attention. Such biasing would be a plausible

reason for the relatively extensive training required of observers using adequately defined procedures (e.g., Madsen, Becker, & Thomas, 1968) as well as the necessity of periodically checking on and retraining for interobserver agreement (Patterson & Harris, 1968). If it could be assumed that the published observational literature has used representative samples from the observer population, then the effect of observer bias would seem to be minimal and effectively "neutralized" by training procedures. However, such assumptions may be misleading if investigators, for convenience or in terms of their own biases, retain observers who show relatively high initial agreement and discard those who do not agree or who do not respond favorably to training procedure. As several prominent investigators have indicated, this may not be an uncommon practice in observational research, particularly where only two observers are used (Paul Gump and Todd Risley, personal communication, 1968).

The critical point for observational methodology is that in ignoring systematic influences or failing to control for them, erroneous sampling assumptions are made and a margin of methodological error is introduced which could systematically distort investigatory results. While observational methodology may come close to meeting the basic criteria of a natural

science, if there is a systematic error source in the data it should be explicated. It is the thesis of this study that such systematic biasing is operative for observers. It is suggested that interobserver variance, if systematic, should be noted when the observers are differentiated on some meaningful intra individual variable, i.e. responses to an attitude scale (or estimates of their personal reaction to the behaviors to be observed. The assumption is that observers do place directional value (good, bad) on behavior within different contexts, and that such behavioral value loadings affect their recording of behavioral frequency (and subsequent reliability). Such biasing is seperate from, but interacts with variables contributing random error variance.

CHAPTER III

STATEMENT OF THE PROBLEM

Obtaining interobserver agreement for behavior classes has been a consistent problem in observational investigations whether experimental or descriptive. The difficulty in obtaining adequate interobserver agreement is generally posited to be a function of the objectivity of data language, observer training procedures, and methods of calculting reliability indices. The emphasis on these variables carries with it the assumption that interobserver variance is randomly distributed as well as the tacit assumption of observer homogeneity. The only acknowledged forms of systematic observer bias are knowledge of the experimental hypothesis, and an ill-defined "social bias" which is considered to be significantly reduced with data specification. This investigator questions the assumption of observer homogeneity and hypothesizes that interobserver variability is a function of systematic observer biasing in addition to random error factors --systematic bias which is directly related to how the observer personally evaluates the behaviors under observation. More specifically, it is hypothesized that differences be-

tween observers in the recorded frequency of behavior classes across observational samples reflects a systematic bias which is related to how the observer indicates he would consequate (evaluate) such behavior classes.

CHAPTER IV

METHOD

This study was divided into four phases. The first phase consisted of obtaining video tape records of childrens' nonverbal behaviors, deriving a relatively objective inter-judge coding record of the tapes, and developing a behavioral consequence and behavior coding record based on the taped behaviors. The second phase involved the observers rating taped illustrations of the behavior classes (codes) in terms of how they would respond to (consequate) such behavior, and the establishment of test-retest reliability for the behavior consequence instrument. The third phase was a training period in which observers were familiarized with behavior coding materials and procedure. The fourth phase consisted of the observers coding the taped behaviors.

Observers

The observers were 26 volunteer undergraduate students enrolled in junior-senior level psychology (N=15) and education (N=11) courses at a large southwestern state teachers college. Twenty-seven other observers were involved in the second phase, but were dropped from the study. One group of

12 observers were excluded for failure to attend the training period, and an observer group of 15 was excluded because of repeated video equipment problems.

Video Subjects

The video subjects were five kindergarten age males from a private day school. The video subjects were selected on the basis of the investigator's subjective estimate that they would contribute high rates of the behaviors to be coded.

Apparatus

The childrens' behaviors were recorded on an Ampex VR 5100 video tape recorder and shown on a Setchell Carlson 2100 SD monitor. Behavioral data in the observer coding phase was collected on data sheets (Appendix D) bearing a structural similarity to the coding sheets of Patterson (1967). Behavioral evaluations were collected on a behavior consequence record (Appendix A) which lists a series of behavior classes (and specific behavior examples), and possible consequences for each behavior class.

Procedure

Video tape record. For 20-min sessions over a two week period, the interactions of the five subjects were video taped. All recordings were taken in a 7.3m X 9.8m room partioned in half by 1.8m portable storage shelves. Several

storage shelves opened into the recording area and a number of academic and play materials (e.g., building cubes, bean bags, books) were available in the room and storage shelves for the video subjects.

Since the video camera and investigator were in the same room with the subjects, the subjects were told that the investigator was making a tape for himself, and that they could see it when he was completely through. The only other comments made were that they could play freely for 20 min and that they would lose 5 min of play time for throwing objects toward the video camera. Questions by the subjects were ignored except for restatement that the author was making a tape for himself or that they were to do with the time as they wished. No reactions were made to the subjects' activities.

The initial sessions were not video taped though the camera and investigator were present and to all appearances was recording. These sessions were expected to habituate the subjects to the presence of the recording equipment and to confirm that the investigator was a neutral adult who would not interfere with their activities. Habituation lasted for two sessions and was stopped when subjects' approaches to the investigator were one or less per session.

For each taping session following habituation, the taping focus was on the behavior of one video subject, with a different subject being taped in successive sessions. Selection of the subject to be taped in a particular session was determined by blindly drawing the names from a container and assigning each name to successive sessions. During the training and coding phases of the study, observers coded only the behaviors of the target subject for that session.

Three seperate video tape records were obtained from the taping sessions. The first was a continuous record of the video subjects' interactions which was used during the behavior coding phase and contained multiple instances of the behaviors to be coded. The final coding tape consisted of a total tape coverage of 72 min and a total observational footage of 36 min (where observational footage is defined as the alternate 10-sec time samples to be recorded by the observers). Seven behavior classes similiar to behavior codes currently used by other investigators (Madsen, et al., 1968; Patterson & Harris, 1968) were selected from the coding tape as the most frequent behavior classes. The seven coding categories, definitions, and number of recordable behaviors per code (based on the interjudge coding standard, p. 29) are listed in Table 1. The Behavior Consequence Record (Appendix

TABLE 1

CODING CATEGORIES, DEFINITIONS,
AND FREQUENCY OF OCCURRENCE

Code	Definition	Rate per Session			
		1	2	3	4
Isolate Play (I)	This code is to be used whenever a child is engaging in play or preparation for play by himself. The child must be seperated from the other children by two or more feet, and must engage in no interaction with them.	17	7	4	0
Destructiveness (D)	This code is to be used whenever a child damages or attempts to damage any object. Also included in this code would be behavior which is potentially destructive, but where no immediate damage is likely to occur. For example, a child pushing over a chair or throwing objects at the wall would be coded Destructiveness. Behavior which is potentially destructive but is part of a recognized game such as bean bag throwing would be coded Play or Isolate Play.	6	3	4	3
Play (P)	This code is to be used whenever a child is playing with another child in a nonhostile, nonaggressive manner. For example, the children tying each other to chairs or building a house with building cubes would be coded Play. Play is never a solitary activity; it always signifies two or more children interacting. If the child is playing alone it would be coded Isolate Play.	8	47	40	46

TABLE 1 (cont.)

Code	Definition	Rate per Session 1	2	3	4
Negative Physical Contact (-C)	This code is to be used whenever a child attacks another child. Such negative contact may be made with a part of the body or with an object. For example, a child hitting with his hand or a toy, pushing, throwing an object at a child would be coded Negative Contact. This code is used regardless of the intensity of the attack or apparent playfulness.	15	1	9	2
Positive Physical Contact (+C)	This code is to be used whenever a child makes contact with another child which is not negative. The contact may be either affectionate (arm around shoulder) or seemingly accidental or neutral. This code also includes instances where a child makes contact with an object (touching with a toy). Any of these contacts would be coded as Positive Contact.	5	11	11	2
Gross Motor (M)	This code is used whenever the child is engaged in locomotion such as running, walking, skipping etc. or is making exaggerated movements such as swinging his arms in windmill fashion.	49	16	40	31
Self-stimulation (SS)	While most behaviors could be considered self-stimulating, this code is used for those events in which a child stimulates himself in such ways as swinging his feet, rocking, rubbing his eyes, clapping hands etc. These behaviors always involve self contact.	11	3	6	4

TABLE 1 (cont.)

Code	Definition	
Self-stimulation (cont.)	Self-stimulation is not directed toward objects or other children. Also included in this category are those instances where the child uses an object to stimulate himself such as scratching with a pencil.	

A) and the Behavior Coding Record (Appendix B) were developed from the seven behavior classes selected from the coding tape. The Behavior Consequence Record lists the behavior classes, a series of possible consequences for the behavior classes, abbreviated definitions, and where feasible, several examples of the behavior class. The Behavior Coding Record lists the behavior classes and coding symbols for the behavior classes.

The second video tape was a continuous record used as a training stimulus during the observer training phase. The training tape consisted of 20 min total tape coverage and 10 min of observational footage, and contained at least three behavioral examples of each behavior code. The training and coding tapes were edited as little as possible so that the final tapes consisted of a series of naturally occurring behavioral interactions. The third tape contained three illustrations of each behavior code with the illustrations dubbed from the nonobservation periods of the training tape. The illustration tape provided the video stimuli for the observers' ratings on the Behavior Consequence Record.

In order to obtain a relatively objective estimate on the frequency of occurrence and nonoccurrence of the behavior classes, four judges analyzed the coding tape by 10-sec seg-

ments. The four judges (one child psychologist, two primary school directors, one doctoral level special education student) were given the definitions of the coding categories, shown the illustration tape and trained in the coding procedure. The training was the same as that of the observers. The judges were then asked to specify independently the behavior codes appearing in each observation time segment on the coding tape. When disagreement occurred among the judges for a coding category, the judges were told that a disagreement existed and the time segment was reviewed at reduced speed by manually turning the tape reel. When there was continued disagreement, the coding symbols agreed on were retained, and the time segment marked for a disagreement on the particular code. Thus an interjudge record was developed which listed successive observation samples, and the coding agreement or disagreement per sample. The interjudge agreement per time sample provided a relatively objective record against which the observers coded responses were judged as to accuracy. It was assumed that the analysis of small time samples, the review of time segments at reduced speed, the requirement of perfect agreement, and the absence of discussion or feedback on their categorization contributed to an objective coding standard. It is also assumed that the

diverse professional backgrounds and interests of the judges provided some substantiation to agreements on the occurrence-nonoccurrence of behavior classes and gave the coding standard a form of cross ecological validity. In addition, the judging procedure helped insure that the coding categories met the criterion of data language specification, as did the use of coding categories which are similiar to those already proven useful in current investigations.

For the remaining phases of the study, data were obtained in two small group settings where an example of one of the groups would be the volunteers from the psychology class. The groups met after their evening classes and were spaced apart by one hour.

Behavioral evaluation. The purpose of the Behavior Consequence Record was to provide an estimation of how an observer would respond to the occurrence of the behavior classes selected for observation. As was suggested in Chapter II, estimates of how an individual would personally react to behaviors, in terms of the consequences he would place on the behaviors, should constitute at least a self report statement of the observers attitude toward the behaviors. To determine the stability of the observers behavioral consequation, the behavior consequence instrument was analyzed for

test-retest reliability on all observers (test-retest interval of one week). Observer consequation (attitude) was then categorized by positive consequence (attention, approval, positive contact), negative consequence (isolation, disapproval, negative contact), extinction (ignore), and indeterminate (don't know, not certain).

During the test-retest behavioral evaluation phase, observers were given the following instructions on the Behavior Consequence Record:

> You have just been handed what is labeled a Behavior Consequence Record on which are listed a series of general behavior classes along with a listing of ways in which one could react to such behaviors. I will be asking you to indicate what your personal reaction would be to these behaviors after you have observed a video tape which will show examples of these behaviors.
>
> First, I want you to put your first, middle, and last initial at the top of the sheet as well as the name of this group. I am not asking you for your full name even though the information you give will be kept anonymous and will be looked at only by myself. However, I do need your initials since I will be collecting other information later and will need to keep the materials together.
>
> Very good. Now, I want you to look at the first class of behaviors which is Isolate Play. This behavior class consists of such behaviors as playing by oneself, and not interacting with other children as indicated on the consequence record. Such behaviors or behavior pattern could be responded to by you in a number of different ways. Some of the more likely reactions you might make are listed after the behavior class. They are

> (description). In the event one of these ways of responding to behavior is not the way you would react, or if you are not sure how you would respond, there is a category for not being certain as well as a blank space for you to fill in how you think you might respond.
>
> Now, lets assume that you are responsible for the actions of five kindergarten age boys who have been given access to a playroom. You are responsible for their behavior, but do not have to account to anyone for how you control their behavior. Try to visualize yourself in that situation as I show you on the video tape some examples of the Isolate Play class of behaviors. Then put a check to the left of what you believe your reaction would be to these behaviors. Do not put down how you think others would react, or base your response on what others would think of you. What is important is how you personally feel toward these behaviors. Are there any questions? Okay, here are the video examples of Isolate Play. Now indicate how you personally would respond to such behavior.

For all presentations, both a verbal description and video examples of the behavior class were given prior to the observers' ratings. Observers were retested on the Behavior Consequence Record after a one week interval in order to determine the stability of the observers' responses to the behavioral classes. The observers were informed that the retest was to establish the reliability of the instrument. Otherwise, the same instructions and procedure were followed as for the initial test.

The distribution of consequence categories among the behavior codes and the test-retest reliability for the codes

presented in Table 2. It will be noted that the distribution of consequence categories was relatively unidimensional except for Positive Physical Contact where observers split between the categories of positive consequence and extinction. Ideally, such splits would have been preferred on all codes for all possible consequence categories since the split groups would have provided within code reference points and would have allowed comparisons to be made within codes for the differential effects of consequence biases on the coding records.

It will also be noted from Table 2 that the category of indeterminate was used by only three observers. Because of its infrequent usage, the indeterminate category was dropped, and the observers categorized by positive consequence, negative consequence, and extinction. The observers were grouped by these three consequence categories for all analyses involving the observers' consequences for the behavior classes.

Training period. One week following the test-retest period, observers were provided with the Behavior Coding Record and given a brief description of the behavioral codes and their behavioral referents (cf. Table 1). The observers were then given five 4-min practice sessions on the coding procedure using the training tape. No feedback was provided

TABLE 2

DISTRIBUTION OF CONSEQUENCE CATEGORIES
BY BEHAVIOR CODE, AND CODE TEST-RETEST RELIABILITY

Code	Consequence Categories				Un-reliable	Test-Retest Reliab.
	+ Con.	- Con.	Ext.	Indet.		
Isolate Play	16	0	3	1	6	77%
Play	21	0	1	0	4	85%
Positive Contact	11	1	12	0	2	92%
Motor	12	1	3	0	10	68%
Self-stimulation	1	0	20	0	5	81%
Destructiveness	1	22	0	0	3	88%
Negative Contact	0	21	1	0	4	85%

on the appropriateness of the observers' practice codings. Thus the training session was seen as an opportunity to become familiar with the coding sheet and observation procedure without specifically training the observers in their coding responses.

The observation procedure during training was the same as for the coding phase of the study. The technique of noncontinuous observation was used with a 10-sec observation period followed by a 10-sec nonobservation period. During the nonobservation period, a continuous 20-kc tone (dubbed onto the training and coding tapes) was presented, and the start of the observation period was signalled by shutting off the tone. Likewise, the end of the observation period was signalled by the onset of the tone. Coding specifications were that a coding category could be used only once per 10-sec sample, though more than one coding category could be used in the same 10-sec sample. Familiarization with the coding definitions and coding procedure as well as the use of noncontinuous observation on small observation periods satifies observational criteria outlined by Bijou, et al. (1968). These criteria, along with adequate coding specification serve to reduce random error variance and increase the probability of detecting systematic bias by the observers.

For the training session, each observer group received the following instructions:

> In a few minutes, we will observe a video tape of five children in free play. First, however, I want to explain the coding sheet on which you will record the behaviors of the children and the procedure in recording these behaviors.
>
> You will notice that the coding sheet lists behavior codes along with the symbols for these codes. You may recall the definitions of these codes from previous sessions, but in the event you have forgotten, let me review them again. (code definitions) When recording a behavior which falls under one of the codes, you need only put down the symbol for that code. For example, if on the first observation period you observe a child running, you would put an M in the first block on the first line. If other behaviors occur during the same observation period which are included in the other codes, those code symbols would be placed in the same block. For eacy observation period, a coding symbol can be used only once, though more than one code may be recorded in the same block. Thus all of the coding symbols may be found in one observation block on the coding sheet, but none should be listed more than once in that block.
>
> Each observation period will last for 10 sec and each of the blocks on the coding sheet corresponds to one 10-sec observation period. Your observations are to be recorded from left to right on the coding sheet as the arrow indicates. The dotted line does not signify anything. It is only a more "aesthetic" way of seperating rows of observation periods. Your task then will be to look for the kinds of behavior which fall under the codes listed at the top of the coding sheet and to put the coding symbol down in the appropriate block when they occur. Following each observation period there will be another 10-sec period during which you will continue to see the taped behavior though it is not to be coded. To remind you that you are

not to record behaviors during this period, a tone will sound continuously.

You will notice on the board that there is a brief statement of the procedure rules. To briefly review, the rules are:

1) Each block on the coding sheet is for coding one 10-sec observation period.
2) Each code may be used only once per observation period though more than one code can be used.
3) Do not code behavior that you observe when the tone is on. Use that time for coding behavior from the previous period.

Okay, do you have any questions about the procedure rules?--We will now observe four minutes of video tape after which you may ask questions about problems you have in recording. The first part of the tape will have the tone on so it will be a nonobservation period. At that time I will point out the child you will be observing for the entire four minute period.

After the 4 min observation period, questions by the observers were answered by repeating the procedure rules from the instructions. Questions concerning the coding of specific behaviors ("Should this be coded?" or "How should this be coded?") were answered by repeating the coding definitions and leaving the final determination to the observer. Each of the remaining 4-min segments of the 20-min training tape was preceded by a statement of the procedure rules and followed by a question period.

Data recording phase. During the data recording phase, the third video tape was presented to the observers in se-

quential observation periods. The observation periods consisted of four 8-min samples for a total tape coverage of 72 min and a total observational footage of 36 min. The same observation procedures were followed as have been described in the training section. Instructions to the observers prior to each coding session consisted of a restatement of the coding categories and their definitions, the three general procedure rules, and a question period as outlined in the training section. The only technical change was that the observation tape was a continuous 18 min rather than interrupted at 4-min intervals and the observers were so informed.

CHAPTER V

RESULTS

The results are discussed in two sections. The first section deals with interobserver reliability and considers the relative validity of the two major reliability indices used in observation methodology. The second section considers the effect that an observer's attitude toward behavior classes has on his recording of those behavior classes (observer bias).

Interobserver reliability

In analyzing interobserver reliability, two major reliability indices (gross reliability, agreement index) were compared for each of the seven behavior classes. Since an observer's attitude toward behavior, and subsequent observer bias, was proposed as affecting interobserver reliability, each behavior class was classified as positive consequence, negative consequence, or extinction depending on how the majority of observers indicated they would respond to it (cf. Table 2). Those observers who held similiar attitudes toward the behavior class were randomly paired to obtain the reliability indices for that code. For example, on the be-

havior code Self-stimulation, 20 observers indicated that they would extinguish (ignore) occurrences of Self-stimulation. This provided 10 random pairs of observers for assessing the two types of reliability indices and their relative validity. In addition, keeping attitude uniform within codes allowed for an informal comparison of the reliability indices for those codes which observers consequated differently.

The first reliability index, called gross reliability, has been the more typical approach to calculating interobserver reliability (Bijou, et al., 1968). Gross reliability refers to the ratio between two observers for a behavior code over a total observation period, and is calculated by dividing the smaller sum of observer's recordings of the behavior class by the larger. The second reliability measure is called an agreement index, which Bijou, et al. (1968) consider a more accurate and refined method of calculating interobserver reliability. The agreement index refers to the ratio of two observers agreements and disagreements on the occurrence and/or nonoccurrence of a behavior class for small time samples within an observation period. The agreement index is calculated by scoring a small observational time sample (10 sec in this study) as agree or disagree, and dividing the total number of agreements by the number of agreements plus disagree-

ments.

A third index, also considered in this section, is called an accuracy score. The accuracy score is essentially an agreement index except that each observer's coding record is scored for agreements and disagreements with the judges' coding standard rather than another observer. The judges' coding standard, described in Chapter IV, refers to independent agreement among four professionals as to the occurrence-nonoccurrence of the seven behavior classes for each 10-sec observation sample on the coding tape. Granting the assumption that the judges' coding standard is an objective one, the accuracy score is essentially a validity measure. Mean accuracy scores for the group of observers on each code can thus be compared with the means for gross reliability and the agreement index to determine the relative validity of the two reliability indices.

Table 3 lists the means for the three kinds of indices on each child (observation period) and behavior code, and the total means for gross reliability and the agreement index for behavior occurrence. The first point to be noted is that the overall gross reliability index was not representative of gross reliability for the individual observation periods. In comparing the eight behavior categories (six be-

TABLE 3

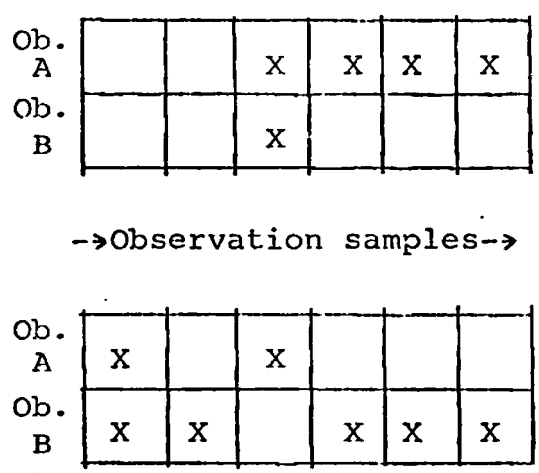
MEAN PERCENTAGE VALUES FOR GROSS RELIABILITY (GR), AGREEMENT (AG), AND ACCURACY (AC), FOR BEHAVIOR OCCURRENCE (o), AND OCCURRENCE-NONOCCURRENCE (on) ON ALL CODES

Code	Child 1					Child 2					Child 3					Child 4					Total		Comparisons			
	AC on	AG on	AC o	AG o	GR	AC on	AG on	AC o	AG o	GR	AC on	AG on	AC o	AG o	GR	AC on	AG on	AC o	AG o	GR	AG o	GR	AC_{on} AG_{on}	AC_o AG_o	GR_o AG_o	AC_o GR_o
I	82	76	57	49	71	88	81	40	33	58	77	71	42	35	70	72	59	--	28	55	38	82	xxx	xxx	xxx	xxx
P	70	66	50	45	59	82	74	76	73	82	79	75	70	61	83	80	68	62	53	71	60	86	xx	xx	xxx	xxx
$+C_+$	86	83	35	26	59	70	64	21	16	30	84	76	26	23	35	94	87	12	03	22	16	38	xxx	xx	xx	xx
M	80	68	69	66	73	84	70	54	47	66	80	69	65	63	68	82	73	68	61	79	61	78	xxx	xx	xx	xx
D	77	75	41	42	52	99	99	32	25	80	89	91	25	22	52	94	97	19	15	58	35	62	ns	ns	xx	xx
-C	81	80	38	35	64	96	94	21	20	36	88	89	34	31	63	98	98	24	21	50	31	62	ns	ns	xxx	xxx
SS	76	75	29	22	37	79	73	23	16	32	83	81	18	17	22	84	80	35	25	60	20	44	x	xx	x	x
$+C_e$	88	86	39	35	55	79	78	25	23	22	81	74	24	15	55	92	87	19	12	46	21	61	x	xx	x	x

xxx less than .02
xx less than .05
x less than .10

havior codes plus split consequences for Positive Contact) on the four children, overall gross reliability was inflated for 26 of the 32 means. For example, on Isolate Play (I), the overall mean was 82% while the individual means for the coding sessions were 71%, 58%, 70%, and 55%. There was not only inflation, but inflation which varied from 11 to 27 percentage points. Indeed, on 17 of the 26 inflated means, gross reliability was inflated by 10 or more percentage points. For the agreement index, inflation occurred on only 17 of the 32 means and was clearly a more representative index.

The large discrepency between total and subtotal reliability was largely due to the absolute nature of the gross reliability ratio as is illustrated in Figure 1. Observer A and B have unequal frequency counts during each observation period but reverse frequency counts on the two observation periods. This yields a cumulative frequency which is equal for the two observers and a reliability index which does not approximate the observers' reliability scores for either of the observation sessions. If the cumulative denominator had been a summation of the denominators for the two observation periods, the cumulative reliability would not have been as inflated and a more representative reliability index would have been achieved. The agreement index does just this by



Coding Session 1

GR	AG
$\frac{1}{4} = .25$	$\frac{1}{4} = .25$

Coding Session 2

GR	AG
$\frac{2}{5} = .40$	$\frac{1}{6} = .17$

Cumulative

GR	AG
$\frac{6}{6} = 1.00$	$\frac{2}{10} = .20$

FIGURE 1

HYPOTHETICAL DATA ILLUSTRATING
INFLATION OF GROSS RELIABILITY

making the denominator a constant which is relative to both observers. It seems apparent that the gross reliability measure contains serious distortions without even considering the issue of data reliability or validity. It might generally be assumed that increasing the number of time samples used to calculate gross reliability will increase the probability of an inflated reliability index.

When the agreement index for behavioral occurrence and the gross reliability index were compared directly, gross reliability was significantly greater than the agreement index for all codes except the extinction codes (*t* test for correlated means). For the extinction codes, the mean differences on Self-stimulation and Positive Contact were 17.75 and 23.25 respectively, which would appear to be a substantial difference. The variance of the mean differences was sizable however, primarily due to the small mean differences for Child 3 on Self-stimulation and Child 2 on Positive Contact. This variance detracted substantially from what was a more general significant difference. Gross reliability was thus significantly greater than the agreement index for behavior occurrence in almost every comparison. Since the method of calculating gross reliability has been shown to provide an inflated index, this was expected.

The discrepency between gross reliability and the agreement index took on additional meaning when the accuracy score for behavior occurrence was compared with the two indices. Reference to Table 3 indicates that the gross reliability inwas significantly greater than the accuracy score on all codes except the extinction codes where small mean differences for Child 3 on Self-stimulation and Child 2 on Positive Contact again detracted from the more general case of significant mean differences. Since the accuracy score represents a relative validity measure, this immediately suggests that gross reliability, by not accurately reflecting behavioral data, is not a valid index. Thus from the standpoint of inflated scores over observation samples or the relative validity of the index, the gross reliability measure is woefully inadequate.

In comparing the accuracy score with the agreement index for occurrence, significant differences were obtained for all codes except those that were primarily negatively consequated. Thus the agreement index appeared valid for those behavior classes which were negatively consequated, but questionable for the other attitude classifications. Since the differences in relative validity were apparently a function of differential observer bias, interpretation of the differ-

ences in relative validity were delayed until the discussion of observer bias.

To this point, discussion has been primarily focused on reliability indices of behavior occurrence. Frequently however, investigators are interested in an observer's detection of behavioral nonoccurrence as well as occurrence as well as occurrence. The agreement index for occurrence and nonoccurrence provides this estimation of data reliability. Comparison of this agreement index with accuracy scores for occurrence-nonoccurrence is presented in Table 3.

The most apparent feature of the comparison between accuracy scores and the agreement index was that there were differences between the codes which were primarily positively consequated, negatively consequated and extinguished. For both negative consequence codes (Destructiveness, Negative Contact) the agreement indices and accuracy scores were similar. For the extinction codes (Self-stimulation, Positive Contact), accuracy and reliability differ, but not at a significant level. For positively consequated codes (Isolate Play, Play, Gross Motor, Positive Contact), the differences were uniformly significant, with accuracy being greater than the agreement index in all cases. Thus for the negative consequence codes and marginally for the extinction codes, the

agreement index was a relatively valid one while the validity of the agreement index for positively consequated codes was questionable. As with the agreement index for occurrence, attitudinal differences significantly affected the relative validity of agreement indices, which suggests that there were differential biases operating in the coding records which affected agreement among observers. The nature of the bias was not indicated by the scores of Table 3 since the accuracy score was in almost all cases greater than the reliability indices. In order to account for the relatively larger decrement in reliability for the positive consequence and extinguish categories, it was necessary to directly analyze the observers' records for coding bias or irregularities.

Observer bias

The major intent of this study was to determine if an observer's attitude toward different behavioral classes (in terms of his stated consequences for those behavior classes) would bias his observational coding record of the behavior classes. Bias, for the purposes of this study, means that an observer either records more behavior than is actually occurring (positive bias), or records less behavior than is actually occurring (negative bias). The judges' coding standard provided a convenient reference point for determining

the extent of such observer bias as well as the sign of the bias (positive or negative). Thus a meaningful dependent variable in comparing the effects of the observer's attitude on his coding of the behavior is the observer's deviation from the judges' coding standard. Since the direction and extent of an observer's bias were considered important dimensions, individual observation samples could not be used since an observer's deviation from the judges' standard would yield only a 0 or $\pm$ 1. Instead, blocks of six 10-sec observation samples were arbitrarily chosen to compute deviancy scores since an observation block would yield deviation scores from 0 to $\pm$ 6, and provide more of a bias continuum. Since the positive and negative signs would have been difficult to handle statistically, a constant of seven was added to make the deviancy scores all positive in sign. Thus the deviancy scores could range from 1 to 13, with a 7 indicating agreement with the judges' coding standard, a 13 maximum positive bias (overrecording), and a 1 maximum negative bias (underrecording). For example, if during a block of six 10-sec observation samples the judges were in agreement that a behavior class had occurred four times and an observer recorded one occurrence, the observer's deviancy score, or bias, would be negative and yield a score of 4. The obtained deviancy score

thus provides a relative index of the direction and magnitude of observer bias.

In addition, the 26 observers all responded to at least one of the seven behavior classes with either positive consequence, negative consequence or extinction. However, the number of behavior classes the observers consequated similarly varied within the observer group. For example, one observer indicated he would provide positive consequences for Positive Contact, Play, and Isolate Play, negative consequences for Destructiveness and Negative Contact, and extinguish Gross Motor and Self-stimulation. For another observer, positive consequence was indicated for Positive Contact, Play, Gross Motor, and Play, negative consequences for Destructiveness and Negative Contact, and extinction for Self-stimulation. Here the behavior class of Gross Motor was responded to differently by the two observers, but both observers responded to one or more behavior classes with all three consequence categories. Since there were shifts across observers in the way they consequated specific behavior classes and the consequences for the behavior classes were relatively unidimensional, formal analysis of individual behavior classes was not possible. However, deviancy scores were calculated on codes the observer consequated the same, and a mean de-

viancy score was obtained for these combined codes which served as a representative index of the observer's bias for each consequence category. With mean deviancy scores as the dependent variable, comparisons could be made for the group of observers on the three attitudes (ways of consequating behavior) across observation blocks. The group of observers are thus replicated on attitude and observations. Since this study was concerned with systematic bias (significant differences across observation samples) as well as response style (systematic bias across data sources), this provided a three factor replicated design where one dimension of interest was attitude, the second was observations within a coding session (within child), and the third was observations across coding sessions (across children). A schematic representation of this design is presented in Figure 2.

The group means for the three factor replicated design are graphically presented in Figure 3 a,b. It can readily be seen that while there were some bias differences between attitudes, the most significant effect appeared to be an interaction of attitude and child. The deviancy scores appeared to be randomly distributed across observation samples and a systematic bias within child, or response style across children was not apparent for any of the attitude categories. For

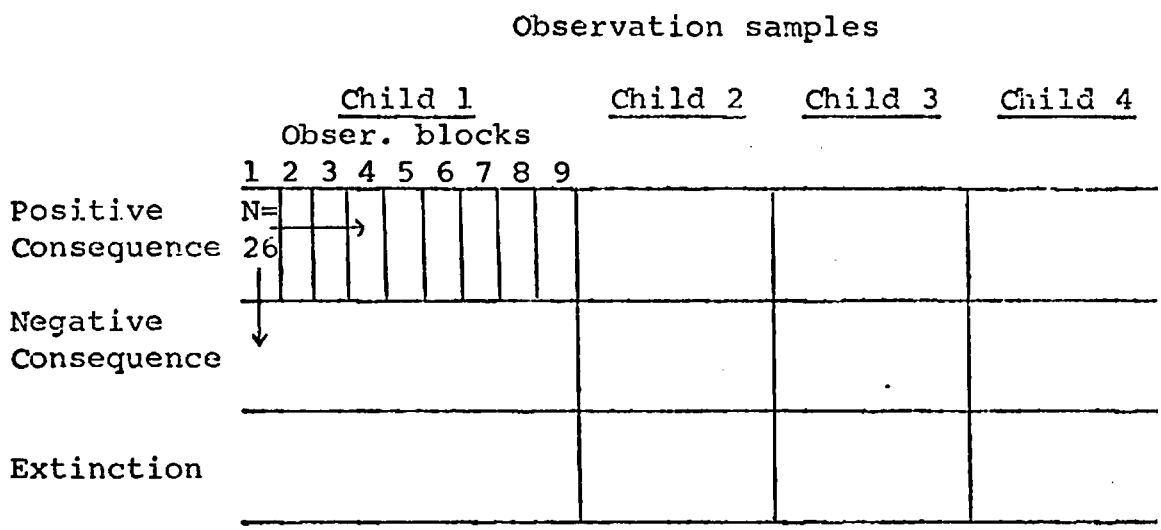
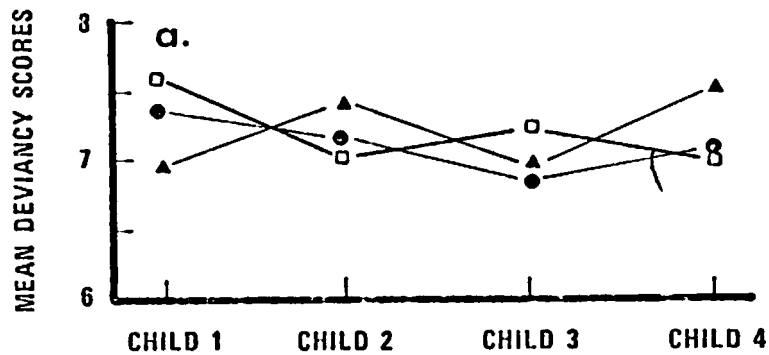


FIGURE 2

ILLUSTRATION OF A THREE FACTOR
REPLICATED DESIGN WITH ATTITUDE COMPARED
FOR OBSERVATION BLOCKS WITHIN AND BETWEEN CHILDREN



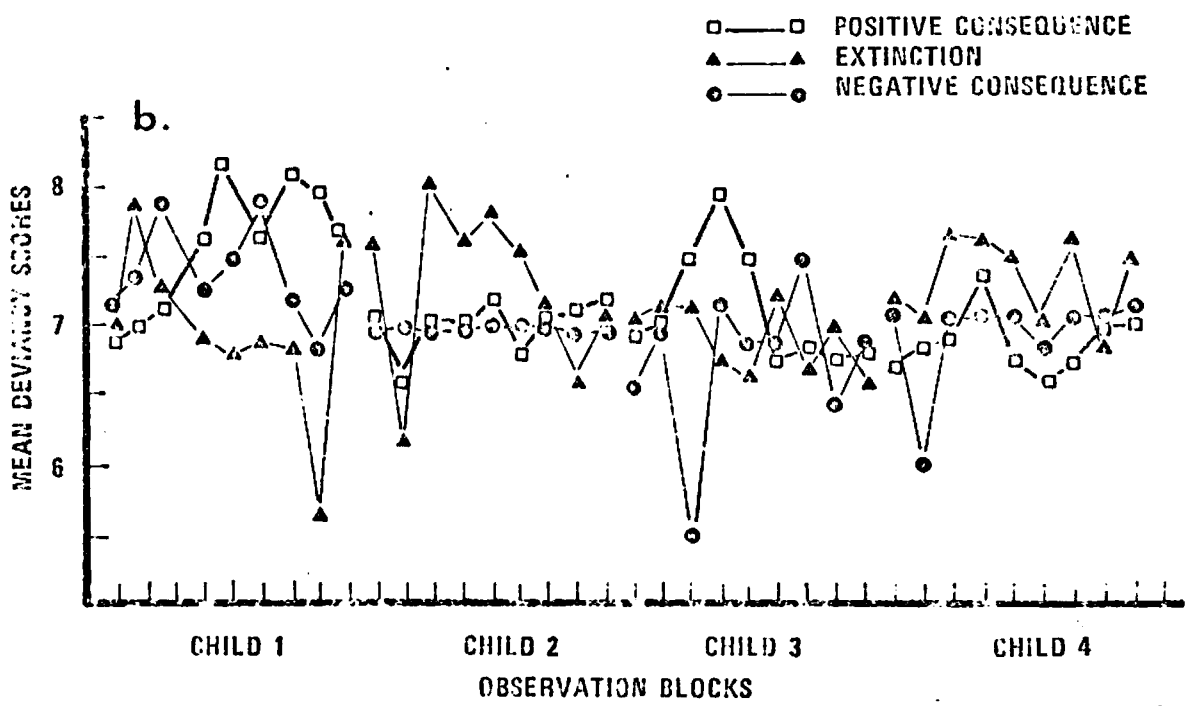


FIGURE 3

COMPARSON OF DEVIANCY SCORE MEANS
FOR ALL OBSERVERS ACROSS OBSERVATION BLOCKS
FOR EACH CHILD, OR CODING SESSION (b), AND OVERALL
DEVIANCY SCORE MEANS FOR EACH CHILD, OR CODING SESSION (a).

within child comparisons, there were some biasing trends between the positive consequence and extinction categories for children 1, 2, and 4. However, there was enough interaction within and between these attitudes to suggest that any systematic error attributable to attitude was heavily weighted by a variable(s) other than those considered. Similiar effects were noted for the other direct comparison--that between the 11 observers who positively consequated Positive Contact, and the 12 observers who indicated extinction for this behavior class (cf. Table 2). This comparison is graphically presented in Figure 4. Since systematic error was not uniformly apparent in the graphed data for either Figures 3 or 4, the data were not treated statistically.

In order to isolate possible contaminating factors, it was decided to inspect the individual codes contained within the attitude categories. Reference to Table 1 indicates that the behavior codes subsumed under the attitude categories varied widely in frequency of occurrence across children. For example, Isolate Play occurred 17 times for Child 1 and 0 times for Child 4, while Play occurred 8 and 46 times respectively. It seems apparent that the relative magnitude of the behavior codes and their cumulative frequency for given observation samples would affect the magnitude of the

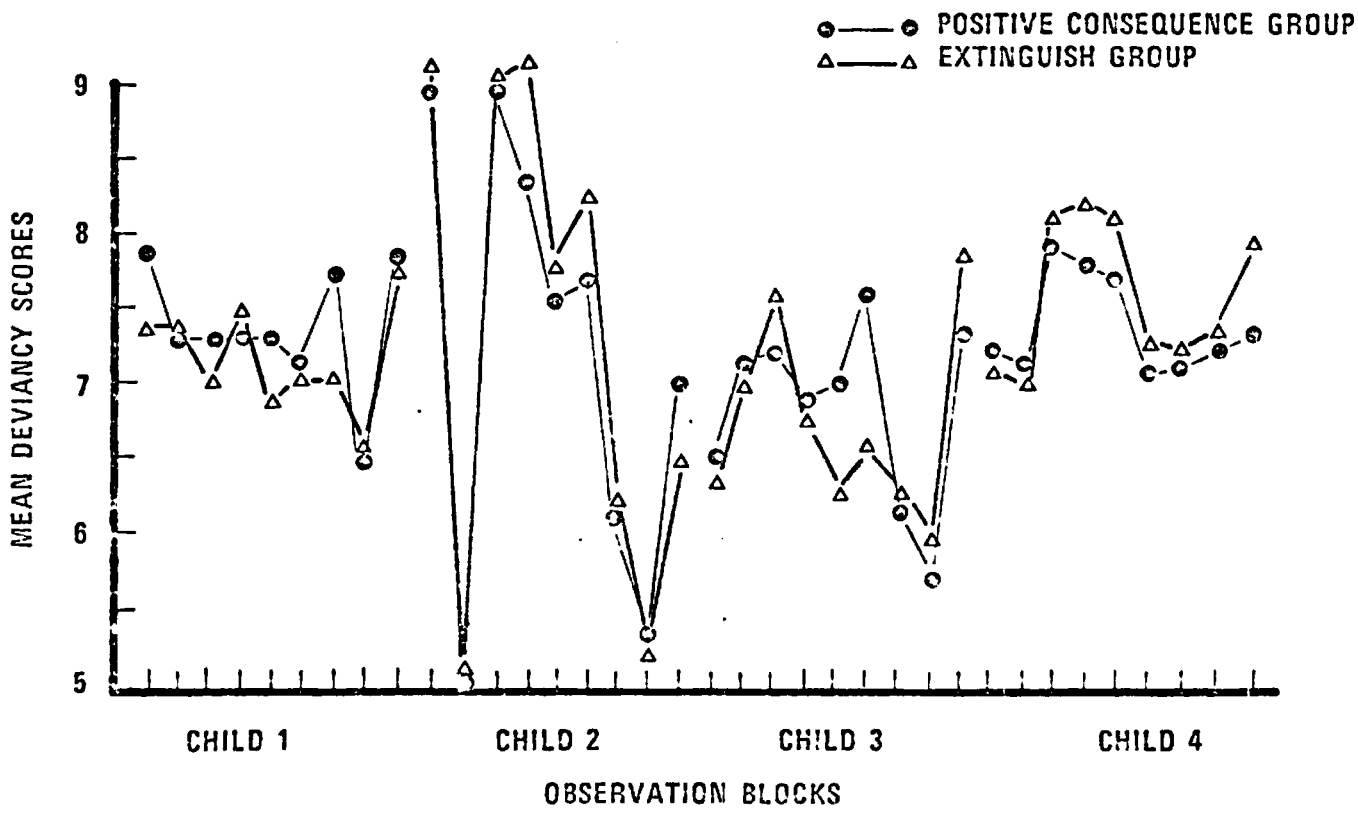


FIGURE 4

COMPARISON OF DEVIANCY SCORE MEANS FOR THE BEHAVIOR CLASS POSITIVE CONTACT

the mean deviancy scores within and across children. This would account for bias variation across observation samples, but would not account for changes in deviancy score sign (shifts from positive to negative bias).

Changes in sign for deviancy scores could be partially affected by the nature of the coding procedure. According to the coding procedure, a block of six observation samples can contain a maximum of six behavioral occurrences for a given code and a minimum of zero. When the judges' coding standard indicates that no behavior occurred for a specific code during an observation block, the only biasing possibilities are positive. When the maximum number of behaviors occur in an observation block, the only biasing possibilities are negative. Obviously, if the trend of combined codes within attitude is toward the maximum rate and regression effects occur, a deviancy score with a negative sign will result. Thus bias variation in magnitude and sign can occur across observation samples partially as a function of the relative magnitude of combined codes and/or regression effects at the extreme ends of the behavior rate continuum.

In order to generally assess the effect of behavior rate on deviancy scores across attitudes, a Lubin test (Lubin, 1961; Lyerly, 1952) was run on each of the individual codes.

The Lubin test essentially compares correlated ranks with hypothetical ranks to determine if a trend can be specified by the theoretical rank order. The Lubin test is equivalent to an average Spearman rank-order correlation and provides a correlation coefficient (K) and a z score for each set of ranked comparisons. Since the Lubin is one-tailed, a negative z indicates a significant relationship between the hypothetical and observed ranks. For the data of this study, the behavior rates specified by the judges' standard allowed each child to be ranked from highest to lowest in frequency of occurrence for a behavior class. The judges' ranks constituted the hypothetical ranks, which were then compared with the observers' ranks (obtained from summed frequency scores within child) to obtain difference scores for computation of K and z. Since individual codes were compared, the observers' ranks were drawn from those observers who provided the majority consequence for an individual code (cf. Table 2). The results for the individual codes, by consequence, are presented in Table 4.

The results of the Lubin tests on the individual codes indicate some interesting relationships. First, with the exception of Positive Contact, the only codes showing rankings which significantly correspond to the judges' rankings were

TABLE 4

SUMMARY OF K AND z SCORES
FOR THE INDIVIDUAL CODES

Behavior Class	Consequence	K	z
Isolate Play	Positive	.58	4.92
Play	Positive	.61	4.51
Positive Contact	Positive	.88	-2.92
Motor	Positive	.61	3.75
Self-stimulation	Extinction	.62	3.78
Positive Contact	Extinction	.88	-2.97
Destructiveness	Negative	.87	-3.50
Negative Contact	Negative	.88	-3.94

those codes which were primarily negatively consequated. The other codes confirmed the null hypothesis of the Lubin test, which is that for each subject, all the permutations of the ranks are equally probable. Thus the negatively consequated codes and the Positive Contact code were fairly accurately presented by the observers while the other codes deviated in some random or systematic fashion from the judges' rankings. To determine how the null hypothesis codes deviated from the judges' rankings, it was necessary to informally examine the trend for ranks. Inspection of the ranked data for the codes falling under the null hypothesis proved quite interesting. For all the null hypothesis codes, the relationship between the judges' ranking and the observers' ranking was in a negative direction. Thus the K values presented in Table 4 can be essentially interpreted as negative ones although the degree of relationship was deflated somewhat because of the directional nature of the Lubin test. This means that on the basis of the gross rankings, those codes which were positively consequated or extinguished (excepting Positive Contact) tended to show bias which was inversely related to behavior rate, while negatively consequated codes tended to reflect ongoing behavior rate more accurately, or show less coding bias. That is, observers who positively consequated or extinguished

behavior tended to show positive bias (overrecording) when the judges' standard indicated a child's behavior rate was low, and negative bias (underrecording) when the coding standard indicated that the child's behavior rate was high. Observers who negatively consequated behavior tended to report behavior rates which were parallel to the judges' coding standard across the behavior rate dimension. These conclusions have to be modified somewhat however, since inspection of Table 1 indicates that only two codes were high rate--both positively consequated. Thus one could generally conclude that for low rate behavior, most of the codes which are positively consequated or extinguished will show positive bias, while the negatively consequated codes will show minimal bias.

If rate could be held constant, then biasing differences between the positive consequence, negative consequence and extinction categories should be more apparent than when behavior rate is variable. Inspection of the judges' coding standard indicated that the three observer attitude categories were comparable for those observation blocks were no behaviors occurred, and where one behavior occurred. The distribution of rate was not sufficient however, for comparison of higher rates. Since both 0 and 1 rate blocks represent

low rate activity, the results of the Lubin tests would suggest that the positive consequence and extinguish categories will generally show a stronger positive bias than the negative consequence category, which should show a close approximation to the judges' coding standard. This analysis is again a three factor replicated design (cf. Figure 2) with consequence categories (attitude), rate (0 and 1), and child being the dimensions of interest, and deviancy scores for the group of observers the dependent variable. Within child comparisons were not possible in this design since the frequency of 0 and 1 rate blocks within child was not equivalent for the three consequence categories. The results of this analysis are graphically presented in Figure 5, the ANOVA summary in Table 5, and comparisons among means (Duncans Multiple Range Test) in Table 6.

The results of the ANOVA strongly suggest that attitude does affect how one records behavior. Attitude along with rate and child-attitude interaction were highly significant by the Box Conservative Test (Winer, 1962). In comparing the mean differences between attitudes, it will be noted from Figure 5 and Table 6 that, with only one exception, there were significant differences between positive and negative consequence categories across children and rate. In all com-

TABLE 5

ANALYSIS OF VARIANCE
FOR RATE X CHILD X ATTITUDE

Source	df	SS	MS	F	
Between Ss	25	.62			
Within Ss	598	522.92			
Child (C)	3	8.32	2.77	2.20	
Error C	75	94.60	1.26		
Attitude (A)	2	103.37	51.65	30.74	xx
Error A	50	84.17	1.68		
Rate (R)	1	39.50	39.50	79.00	xx
Error R	25	12.58	.50		
C X A	6	26.29	4.38	10.95	xx
Error C X A	150	60.67	.40		
C X R	3	2.82	.94	3.92	
Error C X R	75	17.77	.24		
A X R	2	4.08	2.04	6.00	x
Error A X R	50	16.97	.34		
C X A X R	6	7.97	1.33	4.59	x
Error C X A X R	150	43.81	.29		

xx less than .01
x less than .05

TABLE 6

COMPARISON OF MEANS
FOR SIGNIFICANT Fs OF TABLE 5

Attitudes (+,-,e) by child (1,2,3,4) for 0 rate

	4-	2-	3-	1e	1-	2+	4e	3e	2e	3+	4+	1+
$\bar{X}$	7.08	7.08	7.23	7.77	7.77	7.77	7.77	7.84	7.85	8.35	8.58	8.62

Attitudes (+,-,e) by child (1,2,3,4) for 1 rate

	3-	4-	1e	3e	2e	4e	1-	2-	2+	4+	1+	3+
$\bar{X}$	6.58	6.77	6.84	6.88	7.12	7.19	7.27	7.27	7.38	7.70	8.04	8.27

Attitudes (+,-,e) across rate (0,1)

	1-	1e	0-	0e	1+	0+
$\bar{X}$	6.97	7.01	7.29	7.81	7.85	8.33

Note.--Means not connected by a line are significantly different at less than .01.

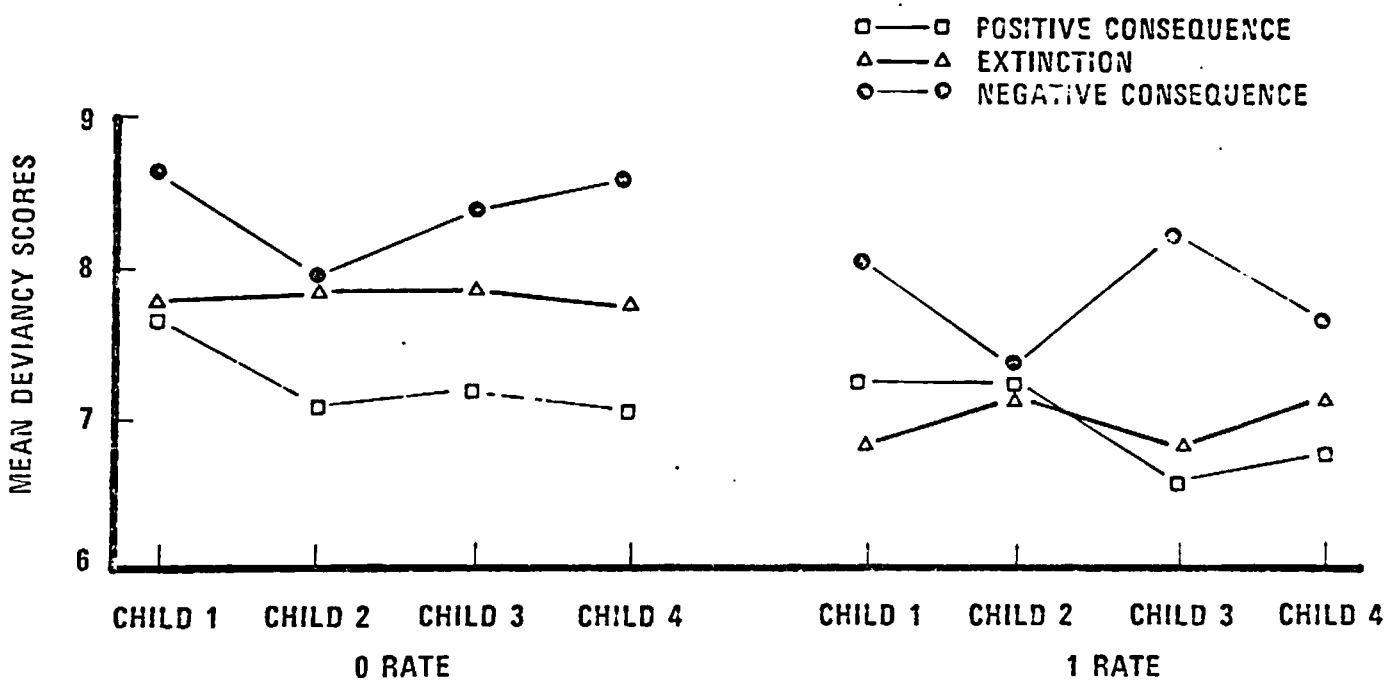


FIGURE 5

MEAN DEVIANCY SCORES FOR

EACH CHILD HOLDING BEHAVIOR RATE CONSTANT

parisons for these two consequence categories, the observers showed positive bias (overrecorded) for the behavior classes they positively consequated, but showed only slight biasing trends for those behavior classes they negatively consequated. For the low rate behaviors in this comparison, the data clearly confirmed the conclusions of the Lubin tests--that for the behavior classes which were positively consequated, observers showed significantly stronger positive bias than they did for behaviors they negatively consequated. Indeed, the graphed data in Figure 5 strongly suggest that the differences between observer bias for behaviors which were positively and negatively consequated corresponded to a response style since systematic differences existed across children (across data sources). For the behaviors the observers would extinguish, there were significant differences from the other attitude categories, but the effect was less dramatic than for the differences between positive and negative consequence.

For rate there were some interesting relationships. For both positive consequence and extinction categories, there were significant decrements in observer bias from 0 to 1 rate when means by child were averaged for the two rates. There was no significant decrease for the behavior classes the observers consequated negatively. These results suggest that

observer bias for the positive consequence and extinction categories shifted toward negative bias as behavior rate began to increase. This was consistent with the Lubin tests which generally indicated a negative relationship between behavior rate and observer bias for these two attitude categories.

The interactions, with the exception of child-rate, were all significant with the strongest being the child-attitude interaction. There were two major sources of variation which contributed to the significant interaction terms. The first arose from the bias decrements for the positive consequence and extinction categories when behavior rate changed from 0 to 1 while there was only a slight decrease for the negative consequence category. In addition, the slope was steeper for the change in the extinction category. This accounted for the significant Attitude X Rate interaction and contributed to the triple interaction. The Attitude X Rate term was an important one since it suggests that there were different functional relationships between observer recording bias and observer attitude; differences which were influenced by behavior rate.

The second source of variation, as evidenced by Figure 5, arose from the nonparallel variation in observer bias a-

cross children for the three consequence categories. This variability in observer bias contributed to both the Child X Attitude and the triple interactions. The meaning of the Child X Attitude interaction is that there are unaccounted for sources of variation in observer bias. Holding behavior rate constant managed to reduce variation in deviancy scores to the point that observer bias differences between attitudes was more apparent, but other variables were influencing the deviancy scores. This variance could be contributed by random error factors such as training, coding specification, etc. However, random error was anticipated in this study and considered markedly reduced by the use of relatively objective codes, training on coding procedure, noncontinuous recording of behavior, and standard stimuli.

One source of variance contributing to the interactions was the variability of observers within attitude categories. Variability was most marked for the positive consequence category where the range of deviancy scores for any observation block varied from two to seven. The classification system was refined enough to detect what were apparently very strong attitude effects, but in its relative simplicity the categorizations only marginally reduced subject variance.

The use of combined codes in this analysis offers another

plausible hypothesis for the interaction effects. While behavior rate was held constant for individual codes by averaging deviancy scores for each observer on positive consequence, negative consequence, or extinction codes, the observation blocks from which the 0 and 1 rate means were drawn were not. For example, on Child 1, 0 rate behavior was noted in the judges' standard in observation blocks 1,2,3,7, and 9 for Self-stimulation (primarily extinction), and blocks 2, 3, and 9 for Positive Contact (extinction for half the observers). The common observational blocks for behavioral nonoccurrence were blocks 2, 3, and 9 while blocks 1 and 7 contained 0 rate for Self-stimulation and ongoing behavior for Positive Contact. Thus behavior rate was only a relative rather than an absolute constant in comparing attitudes. It would seem plausible to hypothesize that when behavior rate is zero for all codes under an attitude such as extinction, the probability of bias would be greater than if there was ongoing behavior in one or more codes for an observation block. For the example of Self-stimulation and Positive Contact, the mean deviancy score for Self-stimulation when both codes were 0 rate was 8.33. When only Self-stimulation was 0 rate the mean deviancy score dropped to 6.80. If this effect were to hold for the three attitudes across children, it can readily be seen that there

would be some variability in the coding records which would be a product of code covariance within attitude. The logical next step would be to look for covariance between attitudes.

The major importance of the interactions is that they limit the generality of any claims that can be made for response style (systematic bias across children). This is particularly true for the child-attitude interaction. While systematic bias was indicated, a pure response style was not readily apparent except for the differences between positive and negative consequence. Response style was indicated for other attitude differences, but the interactions were sufficient enough to limit the conclusions to specific children and/or rates. The effects were there however, and were hopefully limited only by the error introduced by the relatively simple method of classifying observers and/or attitude covariance.

The results of the Lubin tests and ANOVA made possible an interpretation of the differences between the accuracy scores and agreement indices in Table 3. Since many of the behavior codes in this study were essentially low rate behaviors, the above analyses would suggest that the bias for the positively consequated codes would tend to be positive, and neutral to positive for extinction codes. For negative

consequence codes the bias would tend to be neutral. With this continuum in mind, the following logic applies. When two observers tend to overrecord behavior, there will be a decrease in the accuracy score, but the agreement index will decrease only to the extent that the overrecording of the observers is in disagreement or uncorrelated. However, decrements in the accuracy score and agreement index are disproportionate since uncorrelated overrecording will, for every observation above actual, feed two disagreements into the denominator of the agreement index and only one into the denominator of the individual observer. Thus for uncorrelated observations, the reliability index drops at approximately twice the rate of the individual accuracy score or combined mean of the accuracy scores. The meaning of this for the data in Table 3 is that the discrepency between accuracy and agreement increases as the bias and corresponding possibility of uncorrelated observations increases. The discrepency between accuracy and agreement is greatest for positive consequence and least for negative consequence. Thus positive consequators should show more positive bias than the extinction category, which should be greater than negative consequence. This is essentially what the results of the Lubin tests and ANOVA have indicated.

CHAPTER VI

DISCUSSION

The data in the present study support the hypothesis than an observer's attitude toward behavior he is observing does affect his coding response. The effects are, however, not independent and do interact with other behavioral variables. The most critical variable in analyzing observer bias appears to be behavior rate, at least for those observers who indicate they would positively consequate or extinguish certain behavior classes. For these two attitude classifications, the data suggest that observer bias is positive (overrecording) when behavior is not occurring, or occurring at a very low rate. The data would also imply that when behavior is occurring more frequently, the positive consequence and extinguish observers tend to introduce a negative bias, or record a lower rate of behavior than is actually occurring. Observers who negatively consequate behavior are not as affected by behavior rate variation nor do they bias the coding records to the extent that the positive consequence or extinguish observers do.

Of the two reliability indices, the agreement index was

most likely to reflect the accuracy (agreement with the judges' coding standard) with which observers detected behavior occurrence-nonoccurrence. The gross reliability index proved to be a highly inflated measure, and an inadequate index of either interobserver agreement or accuracy. A such, the gross reliability measure probably should be discontinued in observation methodology. These results confirm the conclusions drawn by Bijou, et al. (1968) that, of the two indices, the agreement index is to be preferred. However, it was apparent that the relative validity of the agreement index varied as a function of observer bias. For observers who positively consequate behavior, the agreement index was significantly lower than their accuracy scores. For the extinction classification, the agreement index was lower than accuracy, but the differences were of marginal significance. Only the negative consequence classification had agreement scores which paralleled accuracy. Thus observer bias (positive) tends to deflate the agreement index. This is not a critical problem however, since error introduced by observer bias is in a conservative direction. Investigators should be aware of these influences nonetheless.

The data thus strongly suggest that an observer's evaluation of the behavior he is observing has a marked effect

on his recording of that behavior, as well as influencing interobserver reliability indices and subsequent implications for data validity. It seems apparent from the results of this exploratory study that systematic bias (systematic differences between observers across observation samples) and response style (systematic bias across data sources) associated with attitudes toward behavior should be considered relevant variables in observation methodology. Considering that the results were probably dampened by the relatively simple observer classification system, and sources of error such as code covariance within attitude, interobserver variance due to systematic bias and response style should be considered a critical source of error. Though the results tend to confirm the assumptions on which this study was based, there is still the necessity for research which systematically samples behavior rate so that the interaction of rate and bias suggested by, and partially confirmed in this study can be fully explicated. It will also be necessary to investigate more extensively the effects of code covariance within attitude, and perhaps covariance between attitudes, so that these possible sources of error can be controlled for to reduce coding variation in future investigations.

It is also recognized that classifying the observers by

how they indicate they would consequate behavior may reflect and be contaminated by factors such as social desirability (Edwards, 1957) rather than the individual's personal attitude toward the behavior or behavior class. The design of this study did not differentiate between social and individual (public and private) bias. While it was assumed for the purpose of this study that such biases do not differentially affect how one would categorically consequate behavior or code its frequency of occurrence, such discriminative possibilities were not excluded. To partially guard against such confounding variables, and to ensure personal evaluations, the behavior consequence rating instructions emphasized that the ratings were confidential, and stressed that the observers rate the behaviors in terms of their personal evaluation. As indicated in Chapter V however, there was some variance in deviancy scores within attitude categories, particularly for positive consequence. This suggests that the observer classification system used in this study did not completely differentiate behavioral attitudes, even though it was successful in seperating observers to the extent that observer bias could be detected. It seems plausible that some observers might have indicated how they personally would respond toward the behavior classes, while others indicated what

would be considered socially appropriate. The latter observers could either be suppressing a more personal response to the behavior class, or simply did not have a personal reaction. Since the variance in observer bias was greatest for the positive consequence category and social expectations are more likely to bias one toward favorable responses, the distinction between public and private attitudes might be an important dimension in developing a more discriminative observer classification system.

The constraints on generalizing the results of this study do not, however, lie in the relative imprecision of either technique or procedure. The question of generalizability, for the statistics of this study, is primarily focused on the failure to systematically sample what would appear to be a critical variable in an observation technology--behavior rate. The contribution of behavior rate to observer bias was not anticipated since neither variable has been dealt with in the observation literature. Any other constraints imposed by the sampling of observers, behavioral situations, or behaviors (Cronbach, et al, 1967) would appear subordinate to the sampling of behavior rate since the behavior rate-bias interaction, if confirmed, would introduce error in observation procedures across the other sampling dimensions.

The effects of systematic bias and response style pose methodological questions for design technology and data interpretation which investigations using observation methods will need to take into account. As an example, consider a clinical case where one goal is to increase the rate of some valued behavior for a child. If we grant the full effects of the behavior rate-bias interaction, it should be apparent that if the child's behavior rate is low, the baseline rate would be inflated by observers who would positively consequate the behavior they are coding. The effect of the inflated baseline would be to mask the actual behavior increments occurring as a function of introducing a therapeutic variable. As the behavior rate increases, the tendency of the observers would be to introduce a negative bias, thus dampening the effects of the therapeutic variable even further. If the observers were those who would negatively consequate the behavior, the coding records would be more accurate, but the therapeutic variable would appear more effective for one set of observers than for another. Where attitudes toward behavior are relatively uniform, as they were for most of the behavior classes in this study, investigations are more likely to show stronger effects for deviant behavior and less change for more socially approved behavior--assuming

one is manipulating relevant independent variables. Thus any replication design where observers act as data recorders would be subject to systematic error which could easily lead to misinterpretation of the results.

The partial findings of this study also raise some general questions about the frequently observed discrepency between major social agents (parents, teachers) on the reported rate of behaviors under clinical investigation. If the results of this study have any generality, then differences in the social agents behavioral reports could well be related to different attitudes toward the behavior(s) under investigation. That various social agents respond differentially to the same classes of behavior has been well documented (Bandura, 1969; Bandura & Walters, 1967). Given different responses by social agents to the same behavior class, one can question whether discrepent reports on behavioral incidence are a function of the "actual" behavior rate in different behavioral situations or a cofunction of the reporting bias held by different observers in those settings. Such questions have numerous practical and methodological implications for therapeutic assessment and treatment planning. This is particularly the case where therapeutic efforts do not employ direct observations with children, but instead rely on re-

ports from parents or other social agents to develop behavioral programs for the child (Lakin, 1967; Lindsley, 1966). Observer bias has been shown in this study to be a strong source of error under fairly precise observation conditions. Where direct observation is not employed and observation technology minimized, observer bias as a systematic source of error would seem to be even more critical.

Bijou, et al. (1968) suggest that a third observer can be used on occasion to determine if observers are introducing systematic bias into their coding records. However, the use of a third observer may not necessarily uncover bias trends. If, as was the case for the sample of observers in this study, observers responses to a behavior class are fairly uniform, the probability of the third observer having the same behavioral attitude and coding bias would be quite high.

If the results of this study are replicated and systematic bias for samples of behavior rate, observers, situations, and behaviors is confirmed, controls for systematic error will become a technological problem of immediate concern. One possible way of minimizing observer bias would be to introduce a correction factor into an observer's coding data. Once observer bias is detected, with procedures similiar to the ones used in this study, mean deviancy scores could be

derived and this score used as a correction term. Ultimately however, detection and control of systematic bias will depend on defining precisely the antecedents of the bias, and controlling the antecedents in such a way that high reliability and minimal bias in recording behavior is insured. In terms of the results of this study, this could mean using only observers who would negatively consequate the behavior, or using such observers to check the reliability of those whose bias has been reduced with training procedures.

The reference to training raises the question at this point as to what effect training procedures would have in reducing systematic error arising from behavioral attitudes. This study provided only familiarization training for coding procedure, which, it is interesting to note, was sufficient to produce relatively high mean agreement and accuracy scores, at least for behavior occurrence-nonoccurrence indices. The next step would be to provide feedback on the accuracy of the observer's recording. Assuming that such training can counter the effects of systematic bias, there remains the possibility that the depolarization of attitude will not be stable. Patterson and Harris (1968) indicate that even after a few hours of observation, interobserver reliability begins to show decrements. A plausible explanation for such findings is that

lowered interobserver reliability after training might be a function of "creeping bias"--observers returning to the systematic recording bias they held prior to training. Thus another area of investigation would be to examine the effects of training on coding bias, and to a sess the stability of training results.

BIBLIOGRAPHY

Ayllon, T., & Azrin, N. The token economy: A motivational system for therapy and rehabilitation. New York: Appleton-Century-Crofts, 1968.

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

Bandura, A., & Walters, R. H. Social learning and personality development. New York: Holt, Rinehart and Winston, 1967.

Bernal, M. E., Duryee, J. S., Pruett, H. L., & Burns, B. J. Behavior modification and the brat syndrome. Journal of Counseling and Clinical Psychology, 1968, 32, 447-455.

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

Buehler, R. E., Patterson, G. R., & Furniss, J. M. The reinforcement of behavior in institutional settings. Behavior Research and Therapy, 1966, 4, 157-167.

Butler, J. M., Rice, L. N., & Wagstaff, A. K. On the naturalistic definition of variables: An analogue of clinical analysis. In H. H. Strupp, & Luborsky (Eds.), Research in psychotherapy. Vol. II. American Psychological Association, 1962.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Charlesworth, R., & Hartup, W. W. An observational study of positive social reinforcement in the nursery school peer group. Educational Testing Service, Research Bulletin, 1966, 52-66.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Multifacet studies of generalizability. Technical Report OE 6-10-268, September, 1967, U. S. Office of Education. Cited in Patterson, G. R., & Harris, A., 1968.

Edwards, A. L. The social desirability variable in personality assessment and research. New York: Dryden, 1957.

Ellis, A. Reason and emotion in psychotherapy. New York: Lyle Stuart, 1962.

Goldberg, L. R. Simple models or simple processes? Some research on clinical judgements. American Psychologist, 1968, 23, 482-496.

Grosz, H. J., & Grossman, K. G. Clinicians response style: A source of variation and bias in clinical judgements. Journal of Abnormal Psychology, 1968, 73, 207-214.

Hawkins, R. P., Peterson, R. F., Schweid, E., & Bijou, S. W. Behavior therapy in the home: Amelioration of problem parent-child relations with the parent in a therapeutic role. Journal of Experimental Child Psychology, 1966, 4, 99-107.

Lindsley, O. R. An experiment with parents handling behavior at home. Johnstone Bulletin, 1966, 9, 27-36.

Loevinger, J. Measurement in clinical research. In B. B. Wolman (Ed.), Handbook of clinical psychology. New York: McGraw-Hill, 1965.

Lovaas, O. I., Freitag, G., Gold, V. J., & Kassorla, I. C. Recording apparatus and procedure for observation of behaviors of children in free play settings. Journal of Experimental Child Psychology, 1965, 2, 108-120.

Lubin, A. A rank order test for trend in a set of correlated means. Bulletin du C.E.R.P., 1961, 10:4, 433-444.

Lyerly, S. B. The average Spearman rank correlation coefficient. Psychometrika, 1952, 17, 421-428.

Madsen, C. H., Becker, W. C., & Thomas, D. R. Rules, praise, and ignoring: Elements of elementary classroom control. Journal of Applied Behavior Analysis, 1968, 1, 139-150.

Maher, B. Introduction to research in psychopathology. New York: McGraw-Hill, 1970.

O'Leary, K. D., O'Leary, S., & Becker, W. C. Modification of a deviant sibling interaction pattern in the home. Behavior Research and Therapy, 1967, 5, 113-120.

Patterson, G. R. Manual for the behavior rating sheet. Unpublished manuscript, Oregon Research Institute, University of Oregon, 1967.

Patterson, G. R., & Brodsky, G. A behavior modification programme for a child with multiple problem behaviors. Journal of Child Psychology and Psychiatry, 1966, 7, 277-295.

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

Patterson, G. R., Littman, R. A., & Bricker, W. Assertive behavior in children: A step toward a theory of aggression. Monographs for the Society of Research for Child Development, 1967, 32, 1-38.

Patterson, G. R., & Reid, J. B. Reciprocity and coercion: Two facets of social systems. In C. Neuringer, & J. Michaels (Eds.), Behavior modification for clinical psychology. New York: McGraw-Hill, 1969.

Phillips, E. L. Behavior change among children through use of adults as change agents. Unpublished manuscript, George Washington University, 1967.

Rorer, L. G. The great response-style myth. Psychological Bulletin, 1965, 63, 129-156.

Rosenthal, R. Experimenter effects in behavioral research. New York: Appleton-Century-Crofts, 1966.

Scott, P. M., Burton, R. V., & Yarrow, M. R. Social reinforcement under natural conditions. Child Development, 1967, 38, 54-63.

Straughan, J. H. Treatment with child and mother in the playroom. Behavior Research and Therapy, 1964, 2, 37-41.

Wahler, R. G. Child-child interactions in free field settings: Some experimental analyses. Journal of Experimental Child Psychology, 1967, 5, 278-293.

Wahler, R. G., Winkel, G. H., Peterson, R. F., & Morrison, D. Mothers as behavior therapists for their own children. Behavior Research and Therapy, 1965, 3, 113-124.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

Wright, H. D. Observational child study. In P. Mussen (Ed.), Handbook of research methods in child development. New York: Wiley, 1960.

Wright, H. D. Recording and analyzing child behavior. New York: Harper & Row, 1967.

Yarrow, M. R. Problems of methods in parent-child research. Child Development, 1963, 34, 215-226.

APPENDIX A

BEHAVIOR CONSEQUENCE RECORD

BEHAVIOR CONSEQUENCE RECORD

Initials __ __ __ Group_____

Isolate Play (Where the child is not interacting with other children--is playing by himself)

___Attention (look toward, listen to)
___Approval (smile, "That's good")
___Positive contact (rumple hair, arm on shoulders)
___Ignore
___Isolation (seperate from the other children)
___Disapproval (frown, "Don't do that")
___Negative contact (spank, restrain)
___Not certain or don't know
___Other_____

Destructiveness (Damage or potential damage to an object)
(Examples:
(A boy kicking a cushion)
(A boy knocking over a chair)
(A boy breaking a toy)

___Attention (look toward, listen to)
___Approval (smile, "That's good")
___Positive contact (rumple hair, arm on shoulders)
___Ignore
___Isolation (seperate from the other children)
___Disapproval (frown, "Don't do that")
___Negative contact (spank, restrain)
___Not certain or don't know
___Other_____

Play (Involving two or more children in non-hostile activity)

___Attention (look toward, listen to)
___Approval (smile, "That's good")
___Positive contact (rumple hair, arm on shoulders)
___Ignore
___Isolation (seperate from the other children
___Disapproval (frown, "Don't do that")
___Negative contact (spank, restrain)
___Not certain or don't know
___Other_____

Negative Physical Contact (one child attacking another child)
(Examples:
(A boy hitting another)
(A boy hitting another with an object)
(A boy pushing another)

___Attention (look toward, listen to)
___Approval (smile, "That's good")
___Positive contact (rumple hair, arm on shoulders)
___Ignore
___Isolation (seperate from the other children)
___Disapproval (frown, "Don't do that")
___Negative contact (spank, restrain)
___Not certain or don't know
___Other_____

Positive Physical Contact (affectionate or seemingly accidental or neutral contact)
(Examples:
(A boy touching another)
(A boy holding or hugging another)
(A boy patting another)

___Attention (look toward, listen to)
___Approval (smile, "That's good")
___Positive contact (rumple hair, arm on shoulders)
___Ignore
___Isolation (seperate from the other children)
___Disapproval (frown, "Don't do that")
___Negative contact (spank, restrain)
___Not certain or don't know
___Other_____

Gross Motor (locomotion or exaggerated body movements)
(Examples:
(A boy running)
(A boy jumping)
(A boy rolling)

___Attention (look toward, listen to)
___Approval (smile, "That's good")
___Positive contact (rumple hair, arm on shoulders)
___Ignore
___Isolation (seperate from the other children)
___Disapproval (frown, "Don't do that")
___Negative contact (spank, restrain)
___Not certain or don't know
___Other_____

Self-stimulation (where a child is making self contact or is not directing his movement to other objects or children)
(Examples:
(A boy swinging his feet)
(A boy rubbing his eyes or forehead)
(A boy scratching his head or body)

___Attention (look toward, listen to)
___Approval (smile, "That's good")
___Positive contact (rumple hair, arm on shoulders)
___Ignore
___Isolation
___Disapproval (frown, "Don't do that")
___Negative contact (spank, restrain)
___Not certain or don't know
___Other_____

APPENDIX B

BEHAVIOR CODING RECORD

BEHAVIOR CODING RECORD

Initials __ __ __ Group_____

Behavior Codes

(I) Isolate Play
(D) Destructiveness
(P) Play
(M) Gross Motor
(+C) Positive Contact
(-C) Negative Contact
(SS) Self Stimulation