DOCUMENT RESUME

ED 072 105 24 TM 002 346

AUTHOR Skindrud, Karlton D.
TITLE An Evaluation of Observer Bias in Experimental-Field Studies of Social Interaction. Final Report.
INSTITUTION Oregon Research Inst., Eugene.
SPONS AGENCY National Center for Educational Research and Development (DHEW/OE), Washington, D.C. Regional Research Program.
BUREAU NO BR-1-J-054
PUB DATE 31 Jul 72
CONTRACT OEC-X-72-0001(057)
NOTE 105p.

EDRS PRICE MF-$0.65 HC-$6.58
DESCRIPTORS Analysis of Variance; *Bias; Conceptual Schemes; Control Groups; Data Collection; *Evaluation Criteria; Experimental Groups; Females; Interaction Process Analysis; Measurement Instruments; *Observation; *Psychometrics; Questionnaires; Research Methodology; *Social Relations; Technical Reports; Training Techniques

ABSTRACT

Twenty-eight mature women were recruited from the community and trained in a 21 category observation code of family interaction. Observers were assigned randomly to three experimental groups and given different expectancy rationales about the outcomes of the studies for which they would be collecting data. All groups were told they would be observing a family under a father-present and father-absent condition. One group was led to expect an increase, another a decrease, and a third no change in the rate of deviant behavior for the boys in the family as conditions changed from father-present to father-absent. None of the groups were told they would be observing identical videotape recordings of family interaction permitting comparison of observation data across groups. Results indicated that the expectations of experimental outcomes differed significantly for the three groups. However, observers were totally unbiased in their reports of deviant behavior in group comparisons. Failure to obtain evidence for observer bias in spite of the demonstrated manipulation of observer expectations was attributed to the precautions taken to assure high levels of observer accuracy throughout the study. (Author)

ED 072105

Final Report

Project No. 1-J-054
Grant No. OEC-X-72-0001(057)

Karlton D. Skindrud
Oregon Research Institute
P.O. Box 3196
Eugene, Oregon 97403

AN EVALUATION OF OBSERVER BIAS IN EXPERIMENTAL-FIELD STUDIES OF SOCIAL INTERACTION

July 31, 1972

U. S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Office of Education

ED 072105

An Evaluation of Observer Bias in Experimental-Field Studies of Social Interaction

Karlton D. Skindrud

Oregon Research Institute

Twenty-eight mature, women were recruited from the community and trained in a twenty-one category observation code of family interaction. Observers were assigned randomly to three experimental groups and given different expectancy rationales about the outcomes of the studies for which they would be collecting data. All groups were told they would be observing a family under a father-present and father-absent condition. One group was led to expect an increase, another a decrease, and a third no change in the rate of deviant behavior for the boys in the family as conditions changed from father-present to father-absent. None of the groups were told they would be observing identical videotape recordings of family interaction permitting comparison of observation data across groups.

Results indicated that the expectations of experimental outcomes differed significantly for the three groups. However, observers were totally unbiased in their reports of deviant behavior in group comparisons. Failure to obtain evidence for observer bias in spite of the demonstrated manipulation of observer expectations was attributed to the precautions taken to assure high levels of observer accuracy throughout the study.

FINAL REPORT

Project No. 1-J-054
Grant No. OEC-X-72-0001(057)

An Evaluation of Observer Bias in Experimental-Field Studies of Social Interaction

Karlton D. Skindrud

Oregon Research Institute
Eugene, Oregon

July 31, 1972

U. S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education

ACKNOWLEDGEMENTS

To the memory of my father, who died during the planning of this dissertation His high expectations provided the initial stimuli for continuing my education. He was able to balance hard work with a redeeming sense of humor. His oldest son is still trying to achieve the latter.

To the many who have seen me through the completion of this dissertation:

Dr. Gerald R. Patterson, Chairman of my dissertation committee. His support, delegation of responsibility, and scientific integrity are responsible for this dissertation. He has the gift of making behavioral research an adventure, a challenge, the conquest of a social problem, and fun--all at the same time.

Drs. Barbara Bateman, Stephen Johnson, Richard Rankin, and Hill Walker, who served on my dissertation committee. Each is noted for outstanding achievement and a rich schedule of social reinforcement to students--behavior I have appreciated.

Drs. Richard Jones and John Reid, who gave generously of their time to provide valuable criticism and assistance with data ar..ysis.

Dr. Roberta Ray, who assisted with the writing of a successful grant proposal.

Betty Brummett and Jonni Johnson, who spent many hours coding and recoding the videotape recordings used in this study. They served as co-trainers, co-experimenters, and collaborators throughout.

Ardis, Erik, and Sonia Skindrud, who gave continual encouragement, unflagging support, and served as subjects for the observer training tape.

Past and present colleagues at "Middle Earth" who provided the reinforcing environment and invaluable assistance required for the completion of this dissertation.

And the 28 observers who submitted to the rigors of the training program and proved their maturity throughout the study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Two kinds of expectancy effects in behavioral research are potentially damaging to the results obtained. One type affects the actual response of the subject of the experiment and the other the data collection process (Rosenthal, 1969, p. 182). The latter is a particular problem where human observers rather than automated methods of data collection are employed. This chapter will review the methodological problems presented by these expectancy effects together with relevant studies. The design for the present investigation[1] of observer bias can be found at the end of the chapter.

While meaningful research could hardly be conducted without hypotheses regarding outcomes, the experimenter's expectancies have the potential for subtly confounding the results. Intentional or unintentional communication of the experimenter's expectancies differentially affect subject or observer responses as a function of the subject's treatment condition. Furthermore, while the research design, procedures, and interpretation of the data are public matters, the effect of the experimenter's expectancies upon the subject's or observer's behavior is not open to public scrutiny and may occur without the experimenter's awareness. Not even independent replication of experimental results guarantees control against such expectancy effects (Rosenthal, 1969, pp. 195-196).

Expectancy effects should be of particular concern to investigators conducting evaluations of treatment outcomes of child behavior therapy. Pawlicki (1970) cites the lack of control groups and lack of controls for observer bias as the two most frequent methodological deficiencies in his review of research on child behavior therapy. It is appropriate that most of the studies reviewed below have been drawn from contemporary research in behavior modification.

Expectancy Effects upon the Subject's Behavior

Clever Hans, the horse belonging to Mr. von Osten, a German mathematics teacher, illustrates the subtle communication of expectancies to the subject of an experiment. Clever Hans could add, subtract, multiply, and divide by tapping with his hoof the answers to problems presented by his master and others. His master was unaware of cuing the horse in any way, although careful evaluation by Pfungst (1911) revealed that when the horse could not see his questionaer he ceased to be clever. When he arrived at the correct number of taps, the horse was cued by a nodding of the questioner's head.

Such expectancies may be directly communicated to the subject by the experimenter as in the case of Clever Hans. However, cultural expectancies may affect the subject's behavior independently of the experimenter's expectancies. Hathaway (1948) has argued persuasively for the cultural pressures on patients to appear "sick" upon entering and "well" upon leaving therapy. Such effects created by the apparent expectancies of the situation are referred to as "demand characteristics" by Orne (1969, pp. 147-148). Since cultural expectancies frequently converge with those of the experimenter, especially where the study is of therapy outcomes, no clear distinction will be made between demand characteristics and experimenter expectancy effects upon the subject's behavior in this chapter.

Research on the demand characteristics involved in naturalistic observation has been conducted by several behavior modifiers. Johnson and Lobitz (1972) provide convincing evidence that it is possible for parents of "normal" children to "fake" good and bad child management during home observations. Twelve sets of parents of preschool children were asked to do everything in their power to make their children appear "good" on three days of a six-day home observation period and "bad" on the remaining three days. Parents alternated from "good" to "bad" days in a counterbalanced design. Rates of deviant behavior, parental commands, and "negative responses" by the parents consistently and significantly differed from "g[illegible]" to "bad" days across families.

If parents of "normal" children can potentially "fake" the data according to the demands of an experimental situation, is it possible that the treatment effects reported for the families of deviant children undergoing behavior therapy are due merely to "fakeability" according to the demands of the situation? A placebo study by Walter and Gilmore (1972) suggests that the effects of behavioral intervention in deviant families cannot be accounted for by the demand characteristics of the treatment situation or observer expectancies. The investigators had 12 families with socially aggressive, predelinquent boys come to a prestigious research institute for treatment of their boys' behavior problems. Half of the families received group behavioral intervention focused on the treatment of specific behavior problems (Patterson, Cobb, & Ray, 1972) and half a plausible, leaderless group placebo treatment. Expectations for change remained high for subjects in both groups. Observers collecting data in the homes were kept uninformed regarding group membership. The experimental families showed a significant change while the placebo families remained unchanged. It is hypothesized by the present author that parents of deviant children have less control over the behavior of their children than the parents of "normal" children, making it difficult for the former to "fake" the data as "normal" families could in the Johnson and Lobitz (1972) study.

No studies have yet been conducted which examine the effects of the observer's expectancies upon the behavior of the subjects in naturalistic observation (Johnson & Bolstad, 1972). Rosenthal's review of experimenter effects in studies of human learning and ability, psycho-physical judgement, reaction time, inkblot tests, structured laboratory interviews, and person perception suggests this possibility. Critiques point out errors in Rosenthal's analysis and interpretation of the data (Barber & Silver, 1968; Snow, 1968; Thorndike, 1968), but the possibility for observer expectancy effects on the subject's behavior remains. However, this author sees the effects of experimenter expectancies upon the _observer's data recording behavior_ as a more serious methodological problem.

Expectancy Effects upon the Observer's Behavior

A different expectancy effect is illustrated in the physical sciences by the case of the infamous N-rays (Rostand, 1960). In 1903, a distinguished physicist, M. Rene' Blondlot, Professor of Science at the University of Nancy, reported a discovery during his research on X-rays.

Blondlot came across new rays quite distinct from X-rays. They were stronger in that they could penetrate metals and a great many other substances normally opaque to all known spectral radiation. In particular, when they struck a small spark or flame or any luminous object, they increased the brightness of these sources of light. He chose to call them "N-rays" to honor the site of their discovery. For two years physicists replicated Blondlot's findings to the point of producing photographs of the effects of the N-rays upon electric sparks and by means of prisms, lenses, and other measures independently assessing the wave lengths of the N-rays with good agreement. The reflective and refractive properties of the N-rays were shown to be unique, supporting the significance of the discovery. Such unintended distortions of the data by a group of respected scientists continued to grow through the two-year period until skeptics with opposing biases accumulated evidence to the contrary. Rostand (1960) attributed the collective delusion to pre-conceived ideas and auto-suggestion coupled with the possibility of an overzealous laboratory assistant bent on flattery or deception.

Observer bias may also be a significant problem in the behavioral sciences today. Current research in behavior modification relies almost exclusively upon naturalistic observation as the method of data collection and the criterion of treatment effectiveness. Various reviews (Johnson & Bolstand, 1972; O'Leary & Kent, 1972) document the fallibility of the human observer as a data collector.

The problem of observer bias has received less attention from Rosenthal and his colleagues than experimenter bias and demand characteristics. However, Rosenthal (1966, p. 14) presents the most complete catalogue of possible sources of observer bias in the literature with documentation from the various sciences. Observer bias may occur in the form of recording errors (Kennedy & Uphoff, 1939; Rosenthal, Friedman, Johnson, Fode, Schill, White, & Vikan, 1964) where an average of 1% of the recordings were in error and 71% of the errors were biased in the direction of the experimental hypothesis, computational errors (Laszlo & Rosenthal, 1967; Rosenthal et al., 1964; Rosenthal & Hall, 1968) which, when rechecked, showed errors by 65% of the 34 experimenters, of which 73% of the errors were biased, interpretive errors (Smith & Hyman, 1950) where recordings of interviews matched for content were interpreted differently as a function of the political labels placed on the respondents being interviewed, and intentional errors (Azrin, Holz, Ulrich, & Goldiamond, 1961; Rosenthal & Lawson, 1964) where undergraduates in laboratory psychology classes distorted data to confirm well-known theories of learning and personality.

Rosenthal (1966) considers interpretive errors as the least difficult to control as the data upon which interpretations are based are generally open to public scrutiny and reinterpretation. Scientific integrity and failure to replicate tend to prevent intentional errors. Computational errors may be controlled by careful rechecking of the data. Least public and most difficult to control are the recording errors made by observers. The major focus of the present study is the effect of the experimenter's expectancies, directly communicated, upon the recording errors made by observers. As behavior modifiers use naturalistic observation as their sole criterion of treatment outcomes and rarely control for the effects of observer bias (Kass & O'Leary, 1970; Pawlicki, 1969), it behooves them to carefully study the circumstances under which observer bias occurs. The few studies where observer bias has been systematically evaluated will be reviewed below. Special attention will be given to the conditions associated with the occurrence of bias.

Azrin et al. (1966) had untrained, undergraduate observers track "expressions of opinion" by adults with whom they were conversing. When observers were given an operant interpretation of the phenomenon under study, observations were the mirror image of later reports when observers were exposed to a psychodynamic reinterpretation. It was unlikely that a slight modification in the experimental procedures (shifting from extinction to disagreement) could produce the highly significant differences reported. Simultaneous observations of the same phenomenon from audiotape recordings by a group of the student observers produced very poor inter- and intra-observer agreement. Use of a confederate during a replication of the study revealed fabrication of the data to confirm the theoretical notions advanced by the class instructor. A further attempt to replicate the study with graduate student observers failed, confirming that the results originally reported by the undergraduate observers were due to intentional errors.

Rapp (1965), cited in Rosenthal (1966, p. 21), had eight pairs of observers describe the behavior of a given nursery school child for one minute. A member of each pair had been falsely told that the child under observation was feeling "under par" and the other that the child was "above par." Seven of the eight pairs of observers wrote descriptions that differed significantly in the direction of the expectations given them. Clearly, the definitions of such global behaviors as "above and below par" are vague. The description of the study suggests that any measures of inter-observer agreement taken prior to the differential biasing of the observer pairs would have been low.

Scott, Burton, and Yarrow (1967) compared the observations of an informed observer (Scott) with uninformed observers as they observed the same nursery school child's behavior. Inter-observer agreement on 12 discrete categories of peer interaction was relatively low (.54), but when categories were combined to form frequencies of "positive" and "negative" peer interactions, agreement rose to .89. Both sets of observations confirmed the experimental hypothesis but the informed observer's results provided significantly stronger support. The amount of training, experience, and background of the uninformed observers used in this study were unspecified. It is difficult to determine whether the differences were due to the degree of information given the two sets of observers or to selection differences. Furthermore, the small number of informed observers (N = 1) makes generalization to other informed observers risky.

A field study employing uninformed "calibrating" observers to assess the accuracy and objectivity of a staff of informed observers was reported by Skindrud (1972). The two calibrating observers were given the same training as the informed observers but were uninformed as to the treatment or "deviant" vs. "normal" status of the families observed in their homes. There was a significant tendency for the uninformed calibrating observers to underestimate the deviant behavior relative to the informed observers across all treatment conditions. However, no observer bias was found with the relatively small number of paired observations available for study. The relatively high reliabilities reported by a stringent measure of observer agreement (82%) may have precluded the occurrence of bias even with the differing amounts of information available to the two sets of observers. However, a test for observer bias under a second condition where the informed observers were unaware of monitoring for accuracy and objectivity and where observer agreement is known to drop significantly (Reid, 1970; Romancyzk, Kent, Diament, & O'Leary, 1971) also produced no measurable bias. The small N, a possible selection confound, and an incomplete design make these results tentative. A large-scale replication of such a field study in which expectancy effects upon both observers' and subjects' behavior may occur should be carried out by behavior modifiers when feasible.

An attempt was made to more systematically evaluate observer bias in a simulation of naturalistic observation. Kass and O'Leary (1970) trained 27 undergraduate observers in a nine-category code of disruptive classroom behaviors. Groups informed, uninformed, and misinformed as to the effects of loud and soft teacher reprimands on disruptive classroom behavior coded videotape recordings of classroom interaction. The rates of disruptive behavior reported differed significantly ($F = 7.67$; df = 2, 24; $p < .005$) in the direction of the expectations given the observers. However, O'Leary and Kent (1972), after a re-analysis of the Kass and O'Leary (1970) data, report that the results were confounded by a tendency for the groups to drift apart on code definitions. Johnson and Bolstad (1972) point out that observer drift and observer bias may be the same phenomenon in this case. If observer drift is prevented by anchoring observers to standard code definitions, observer bias should be less likely to occur. Had Kass and O'Leary (1970) checked observer agreement across groups or observer accuracy against some outside criterion during data collection, observer drift, and, consequently, observer bias, may have been minimized.

Summary

There is sufficient evidence that uncontrolled expectancy effects may pose a major threat to the internal validity of experimental-field studies under certain conditions. Demand characteristics may confound the observations for relatively "normal" families (Johnson & Lobitz, 1972). However, demand characteristics had no measurable effect on the families of grossly deviant children (Walter & Gilmore, 1972). Since most child behavior therapy is designed for deviant cases, this may not be a major problem in the evaluation of therapy outcomes.

Observer bias may prove a more serious threat. The evidence existing prior to the present study suggests that observer bias is likely to occur where undergraduate observers are faced with a difficult observation task (Azrin *et al.*, 1966), where global or ambiguously defined code categories such as "expressions of opinion" (Azrin *et al.*, 1966), "above and below par" (Rapp, 1965), and "positive and negative peer interactions" (Scott *et al.*, 1967) are used, or where observer drift from standard code definitions is not controlled (Kass & O'Leary, 1970). Unfortunately, only some of the conditions where observer bias may exist have been adequately investigated. Does observer bias occur under more carefully controlled observation conditions, e.g., where well-trained, mature observers, discrete code categories, and precautions to prevent observer drift are employed? One preliminary study (Skindrud, 1972) which used behaviorally defined code categories and mature women observers monitored for observer agreement during data collection suggested that observer bias is minimal under these conditions.

Differential sensitivity to the dependent variables of a study have been reported in two of the studies reviewed. Kass and O'Leary (1970) noted that their uninformed "control group reported lower levels of disruptive behavior" and "had a lower level of motivation than the other two groups" (pp. 13-14). Skindrud (1972) found that his uninformed observers reported significantly lower frequencies of the dependent variables over all treatment conditions. Only the informed observers knew which 13 of the 29 code categories were the dependent variables of interest. Such observer differences in sensitivity to the dependent variables of a study are likely to threaten the internal validity only where sensitivity is proportional to the absolute rate of the dependent variable (see Skindrud, 1972) or where observers with differing sensitivities are not randomly assigned to treatment conditions. This finding requires replication as it was not documented statistically by Kass and O'Leary (1970) and may be due to a possible selection confound in the Skindrud (1972) study.

Objectives of the Present Study

The present study had three general objectives:

(1) The first was to replicate the findings of Kass and O'Leary (1970) and Skindrud (1972) that informing observers of the predicted outcomes and variables of a study sensitizes observers to the dependent variables across all treatment conditions. It was predicted that informed groups would report higher frequencies of the dependent variables throughout the study than an uninformed control group.

(2) The second and major objective of the present study was to cross-validate the Kass and O'Leary (1970) study with a different coding system, population of observers, and rationale for the manipulation of observer expectancies. The present study also attempted to control for certain deficiencies in previous studies by careful definition of code categories, extensive observer training, and monitoring observer accuracy during data collection. All were hypothesized to control observer drift and minimize observer bias. It was predicted that a powerful research design and manipulation of observer expectancies would produce bias in spite of the controls for the observer drift confound in the Kass and O'Leary (1970) study instituted above.

(3) Assuming that evidence of observer bias is obtained, a third objective was to examine possible correlates of observer bias (e.g., observer accuracy, strength of the expectancy manipulation, behavioral specificity of the code categories, etc.) and develop a theory predicting the circumstances under which observer bias is maximized and minimized.

Design of the Present Study

Twenty-eight mature women were recruited from the community and trained in a behaviorally-defined, 21-category code of family interaction Observers were assigned randomly to three experimental groups and given different expectancy rationales about the outcomes of the study for which they would be collecting data. All groups were told they would be observing a family under a father-present and father-absent condition. However, one group was led to expect an increase, another a decrease, and a third no change in the rate of deviant behavior for the family members as conditions changed from father-present to father-absent None of the groups was told they would be observing identical videotape recordings of family interaction permitting comparison of observer data across groups. To control actual changes in deviant behavior, videotapes had been edited to match rates across father-present and father-absent conditions. To control selection effects due to observer differences in sensitivity to deviant behavior prior to expectancy manipulation, observers were matched on reported rates and randomly assigned to groups. To control observer drift from code definitions, overt random checks of observer accuracy were made throughout data collection. To control sequence effects due to observer fatigue or practice, order of presentation of the father-present and father-absent videotapes was counterbalanced within groups. A two-way analysis of variance (expectancy groups x treatment conditions) was used to test for differential observer sensitivities (main effects) and observer bias (interactions).

CHAPTER II

METHOD

Subjects (Observers)

Recruitment and selection. An advertisement was placed in the help wanted section of the local newspaper which read:

> WOMEN OBSERVER ASSISTANTS NEEDED for interesting research project in child psychology. Requires 3 weeks training and 2-3 weeks work. Must be over 21 years, married, preferably with children.

After initial screening to ensure satisfaction of the age, marital status, and scheduling requirements, 48 applicants were given a battery of aptitude tests designed to select those easiest to train for the observation task. Tests administered included the Minnesota Clerical Test (Psychological Corporation), the numerical reasoning subtest of the Employee Aptitude Survey (Psychological Services), a work sample of the observation task developed by the investigator,[2] and the Bendig (1956) short form of the Taylor Manifest Anxiety Scale. The last test was not used for selection but given for a separate study.[3] The 30 applicants scoring highest on the three selection measures were hired as "observer trainees."

All trainees agreed to a contract requiring them to complete the study within a limited time in order to receive payment. Two dropped out during the first week. The remaining 28 constituted the subjects of the study.

Observer training. The trainees were divided into three groups for optimum training size. Ninety-minute training sessions were scheduled four days a week for three weeks. Sessions were held in a 13' x 20' room containing an Ampex 6000 videotape recorder, a Setchell-Carlson 23" TV monitor, and chairs and clipboards for the trainees. A shelf with two dozen videotapes was prominently displayed to support the illusion of participation as data collectors in a large scale research project.

The observation task involved coding videotape recordings of family interaction according to a 21-category family interaction code based on the system developed by Patterson, Ray, Shaw, and Cobb (1969). One family member was designated the subject of the observation. Observers focused on the behavior of the subject, recorded his behavior and the reactions of other family members to his behavior. They repeated this cycle every six seconds so that a sequence of encoded interactions between the subject and other family members was obtained. Every 30 seconds observers heard a tone to signal them to move down a line on their protocol sheets. Each sheet was designed to accommodate five minutes of family interaction.

The training program[4] consisted of the following steps:

(1) Each of the trainees was given a manual (see Appendix A) and set of flashcards for the 21-category family interaction code. Trainees were told to familiarize themselves with the code definitions so they could correctly repeat the elements of each definition upon presentation of all 21 flashcards prior to the first training session.

(2) The first three training sessions began with written tests on the code definitions. Trainees watched playback of a five-minute recording of simple family interaction while the trainer read (modeled) the correct coding of the behavior of one of the family members. Then they practiced

coding simply that one subject's behavior. Trainees scored their protocol sheets from feedback on the correct coding of the training segment produced by two trainers who had repeatedly coded the segment until both agreed 100% on the code entries. At the end of each session trainees attempted coding typical family interaction from a set of five-minute "test recordings."

(3) The next seven training sessions involved about an hour's practice coding a five-minute segment of family interaction. The trainer usually modeled the correct coding. Then trainees coded the same tape and were given feedback on the "standard criterion coding" for that segment of videotape. The final half-hour of the session was again a test of the trainees' progress at coding interaction with a new "test tape" and feedback on accuracy.

(4) The final two days of training, observers were asked to code yet another family. In actual fact, it was the same family to be used in the present study, but none of the observers was aware of this. They coded 25 minutes of videotape each session with no feedback regarding accuracy. These data provided a stable baseline measurement of the dependent variable on which to match experimental groups. Again, the last half-hour of both sessions was devoted to coding tests with feedback.

Training was conducted by the investigator and two observers from the Social Learning Project at the Oregon Research Institute experienced in the Patterson et al. (1969) family interaction code. The investigator and one of the experienced observers were present at all of the training sessions. One trainer operated the videotape equipment and the other modeled the coding of interaction, provided feedback on accuracy, and clarified code definitions as needed. Working in overlapping pairs tended to ensure that the trainers remained consistent on their definition of code categories and did not "drift apart" over the training period.

Observer accuracy during training. The objective of the training program was 70% accuracy with the trainers' standard criterion coding of the test recordings. The accuracy measure was a stringent one. Observers had to record the same code as the criterion within a 12-second limit of the corresponding criterion code entry without breaking the "stream of behavior" to score one agreement. Percent accuracy was computed by dividing total agreements by total disagreements. During the last week of training, mean observer accuracy was 64% and ranged from 5?% to 70%. On the final day of training two coding tests were administered with a mean accuracy of 68%.

The reader should make a clear distinction between the observer accuracy measure used in the present study and observer agreement between pairs of observers commonly used in experimental-field studies.[5] Observer agreement is frequently higher than observer accuracy, especially where the standard criterion coding of the videotape recording is a "fine-grained" one. The staff of professional observers employed by the Social Learning Project using the Patterson et al. (1969) code average 84% agreement on their field observations but only 64% accuracy when compared to standard codings of videotaped interaction.[6]

Preparation of Videotape Recordings

An intact family known from a research project on "normal" families was contacted to obtain permission for videotape recording of family interaction in the home. The family consisted of both parents and boys aged three, seven, and nine years. None of the family members had undergone psychiatric or psychological treatment of any kind. They were known to be a relatively relaxed family with three active boys and were considered good subjects for the recording of natural family interaction to be used in the present study. They readily consented to the recording with compensation at the rate of $7.50 per hour.

Four videotape recordings were obtained with all family members present and four with all <u>except the father</u> present. Since the three-year-old boy, Craig, served as the subject for all of the observations, the tapes will be generally referred to as the "Craig family tapes." The two sets will be specifically referred to as the "father-present" (FP) and "father-absent" (FA) tapes, respectively. All videotaping was done during and right after the dinner hour so that setting differences (dining vs. living room) were controlled across the two sets of tapes. Extra videotape recording was obtained in both settings so that it would be possible to match both FP and FA tapes on the class of behaviors to be used as the dependent variable in this study.

During a pilot study of the design, the entire set of eight videotapes was coded by six observers trained in the same coding system used in the present study. Analysis of the results of the pilot study indicated that the two sets of tapes were not matched on the dependent variable.[7] Consequently, five-minute segments from the FA tapes were juggled with extra FA segments until the mean rates of the dependent variable were matched across the two sets of FP and FA tapes ($t = 0.31$; df = 38; n.s. at .50 level).

Procedures

The schedule for training and data collection sessions and general design of the study is illustrated in Figure 1.

Variables controlled by the design. The present design controls for (1) actual changes in rates of deviant behavior from FP to FA tapes (p. 18), (2) selection confounds due to observer differences in sensitivity to the dependent variables prior to expectancy manipulation, (3) observer drift from standard code definitions (p. 30-31), (4) sequence effects from the order in which FP and FA tapes were coded due to observer fatigue, boredom, or practice. Unequal group Ns were used to increase the power of that part of the design assessing observer bias.

Control of the potential selection and sensitivity confounds was achieved by rank ordering observers on their coding of the "deviant behaviors" in the baseline tapes.[8] Trios of observers with similar rankings were formed and members randomly assigned to the three experimental groups. Table 1 compares the groups on their baseline observations of deviant behavior in the Craig family.

One of the three expectancy rationales was randomly assigned to each of the experimental groups. Members of the Control group were then reassigned randomly to the Increase and Decrease groups until they numbered 11 each. Six observers remained in the Control group. The purpose of this reassignment was to maximize the possibility of interaction between Increase and Decrease groups and consequently the possibility of finding observer bias.

Possible sequence effects were controlled by counterbalancing the order in which the FP and FA tapes were presented within each expectancy

Figure 1

Design of the Observer Bias Study

Three-week training period 10-21-71 to 11-9-71	Random assignment of subjects to experimental groups 11-11-71	Two-week data collection period for FP-FA subgroups 11-15-71 to 11-30-71		Two-week data collection period for FA-FP subgroups 12-2-71 to 12-14-71	Post-test questionnaire and debriefing 12-15-71
N = 28	Increase (N = 11)	N = 6 →	Retraining for FA-FP subgroups 12-1-71	N = 5 →	
	Control (N = 6)	N = 3 →		N = 3 →	
	Decrease (N = 11)	N = 6 →		N = 5 →	
Sessions 1 2 3 4 5 6 7 8 9 10 11 12		1 2 3 4 5 6 7 8		1 2 3 4 5 6 7 8	
Presentations of the expectancy rationale		* *		* *	

Table 1

A Comparison of the Three Experimental Groups on a Baseline Measure of the Dependent Variable

	Increase	Control	Decrease
Mean rate of deviant behaviors per 5 minutes*	5.63	6.40	5.75
Standard deviation	1.48	1.73	1.60

* $F = 0.44$; df = 2, 25; N.S. at the .25 level

group. Half the members coded the tapes in an FP-FA order and the remaining half in an FA-FP order two weeks later.

The three-week "layoff" for the FA-FP observers between training and data collection could have produced differences in observer accuracy due to inactivity and consequent deterioration of coding skills. To counteract such a trend, one extra training session was scheduled for the FA-FP observers just prior to their two-week data collection period. A comparison of mean observer accuracies for the two sets of counterbalanced subgroups on 11-9-71 and again with the FA-FP subgroups on 12-1-71 resulted in no significant differences ($F = 0.33$; df = 2, 38; N.S. at .25 level).

Manipulation of the Independent Variable

Four one-hour observation sessions were held each week. FP tapes were coded one week and FA tapes another to simulate the collection of consecutive baseline and treatment observations in the field. Different expectancy rationales were presented to each group at the beginning of the first week and repeated at the beginning of the second week. The groups were led to expect different experimental outcomes and to believe they were each collecting data on different sets of FA tapes. Increase, Control, and Decrease subgroups were scheduled on the same day but with a half-hour between so that members from different expectancy groups did not "run into" each other entering and leaving the observation room. There was no evidence either from the post-test questionnaires or from informal conversation that any of the observers from different expectancy groups were aware they were viewing identical FA tapes.

To arouse observer interest in the outcome of the pseudo-study and lend credibility to the rationales presented, one of the three principal investigators from the Social Learning Project (G. R. Patterson, J. B. Reid, or L. A. Hamerlynck) accompanied the investigator during each of the presentations. The accompanying visitor was introduced as "one of the child psychologists at the Oregon Research Institute interested in the outcome of the project."

The following rationales were presented to the FP-FA order subgroups:

Week One

All three groups were told: "You will be coding two sets of videotapes of family interaction: a set made with the father present and, next week, a set with the father absent.

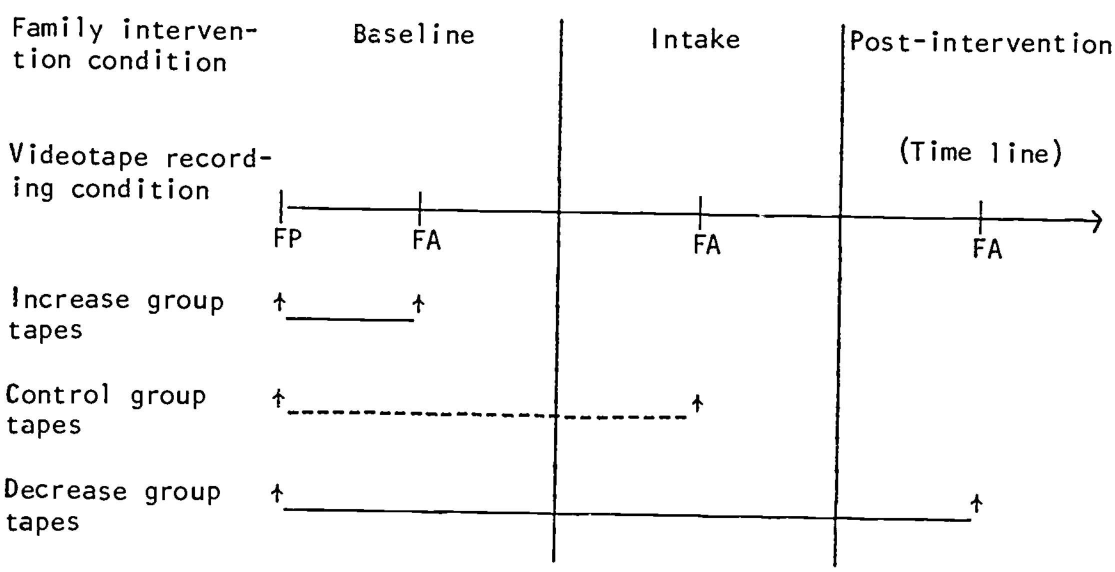
In addition, the Control group was told: "The purpose of this study is to determine the effect of the father's presence upon family interaction."

In addition, both the Increase and Decrease groups were told:

"These videotapes were made of a family referred for the treatment of their boys' behavior problems. Both parents were specifically concerned about Craig's increasingly disruptive behavior--his high rate of yells, whines, and generally aversive behaviors. They wished to bring them under control before Craig entered school and became a behavior problem there. A series of videotape recordings were made in their home at various stages of the treatment program under two conditions: with the father present and absent. All of the groups of observers helping collect data for this study will see a set of FP tapes made before the family received any kind of treatment. However, each group of observers will see a second set of tapes with the FA, made at different times in the treatment program." (A phony research design corresponding to the above description was sketched on the board during the presentation to the Increase and Decrease groups. See Figure 2.)

Figure 2

The Phony Research Design Presented to the Increase and Decrease Groups as Part of the Expectancy Rationale



In addition, the Increase group was told: "This group will see a set of FA tapes next week which were made prior to any treatment. You will be helping us evaluate the effect of the father's absence upon the rate of deviant behaviors, particularly Craig's, the child with the most behavior problems. By 'deviant behaviors,' I mean those listed in the family interaction coding system which are generally undesirable, specifically: crying (CR), threatening commands (CN), disapprovals (DI), dependent requests (DP), destructiveness (DS), high rate behaviors (HR), humiliations (HU), noncompliances (NC), hitting (PN), teases (TE), whines (WH), and yells (YE). We predict that with only one parent there to monitor his behavior, Craig's rate of deviant behavior will significantly increase. Furthermore, some preliminary data from a field study by one of our research assistants strongly suggests that this is the effect of the father's absence. Such an effect will be easier to document from our intensive study of videotapes than we could obtain in the field. We are giving you this information as we have found that observer morale is improved by informing observers of the purpose of the study for which they are collecting data."

In addition, the Decrease group was told: "This group will see a set of FA tapes next week which were made after both parents had been interviewed by the treatment staff and undergone intensive training in child management procedures. We know from considerable research that such training greatly improves parents' ability to manage the behavior problems of their children. We are so confident of these child management procedures we are predicting that after treatment one parent will be able to manage Craig's behavior better than both parents could before.

We expect to see a significant drop in Craig's deviant behaviors as we move from coding the FP to the FA tapes in this group. By 'deviant behaviors,' I mean those codes in the family interaction coding system which are generally undesirable, specifically: crying (CR), threatening commands (CN), disapprovals (DI), dependent requests (DP), destructiveness (DS), high rate behaviors (HR), humiliations (HU), noncompliances (NC), hitting (PN), teases (TE), whines (WH), and yells (YE). We are giving you this information as we have found that observer morale is improved by informing observers of the purpose of the study for which they are collecting data."

Both Increase and Decrease groups were told: "At several points during data collection, a count of the number of deviant behaviors you record on a particular five-minute segment will be made and recorded by one of the trainers. This data will not be shared with the other observers. We wish to get a random sampling of the data to see if there are trends supporting our predictions."

All three groups were told: "Each day your trainers will randomly select one of the five segments you have coded for an accuracy check against a criterion coding of the same. Do the best job you can."

Second Week

The first week's expectancy rationales were reviewed and elaborated at the beginning of the second week when the observers returned to begin coding the set of FA tapes.

The Increase group was told: "As you recall, we predicted a significant increase in Craig's rate of deviant behavior on the FA tapes you'll be coding this week. In fact, we believe there will be an increase in deviant behavior for all family members in the father's absence, including, for example, more teases (TE) from the brothers and more disapprovals (DI) and threats (CN) from the mother. We have some preliminary data from last week's deviant behavior counts for the tapes you coded." (Figure 3 contains the graph of the data sketched on the board for the Increase group.) "As you can see from the preliminary data I've graphed on the board, there was an average of eight deviant behaviors per five minutes with the father present. We predict that the presence of only one parent will bring the average rate up to 12 deviant behaviors per five minutes, about a 33% increase. You may not immediately notice an increase in deviant behavior on the FA tapes as the rates vary tremendously from one segment to another. However, the overall rates should show an increase from the FP to the FA set."

The Decrease group was told: "As you recall, we predicted a significant decrease in Craig's rate of deviant behavior on the set of FA tapes you will be coding this week. In fact, we believe the behavior of all family members will improve as a result of treatment. For example, there should be fewer teases (TE) by the brothers and a smaller number of disapprovals (DI) and threats (CN) by the mother. The intensive training in child management procedures should allow one parent alone to more effectively manage the behavior of the children than both parents could before such training. We have some preliminary data from last week's deviant behavior counts for the tapes you coded." (Figure 4 contains the graph sketched on the board for this Decrease subgroup.) "As you can see from the preliminary data I've graphed on the board, there was an average of nine deviant behaviors per five minutes. We predict that

FIGURE 3

PRE-INTERVENTION FA GROUP (UP)

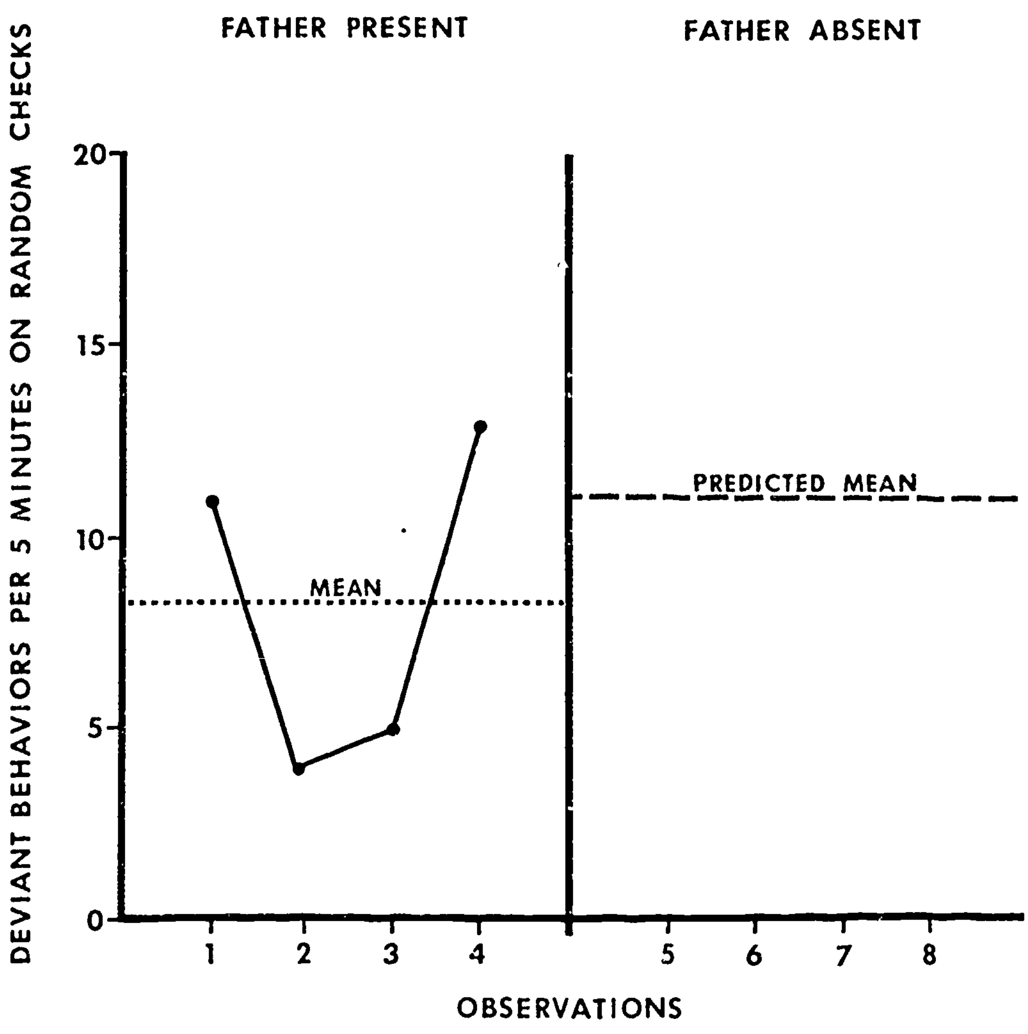
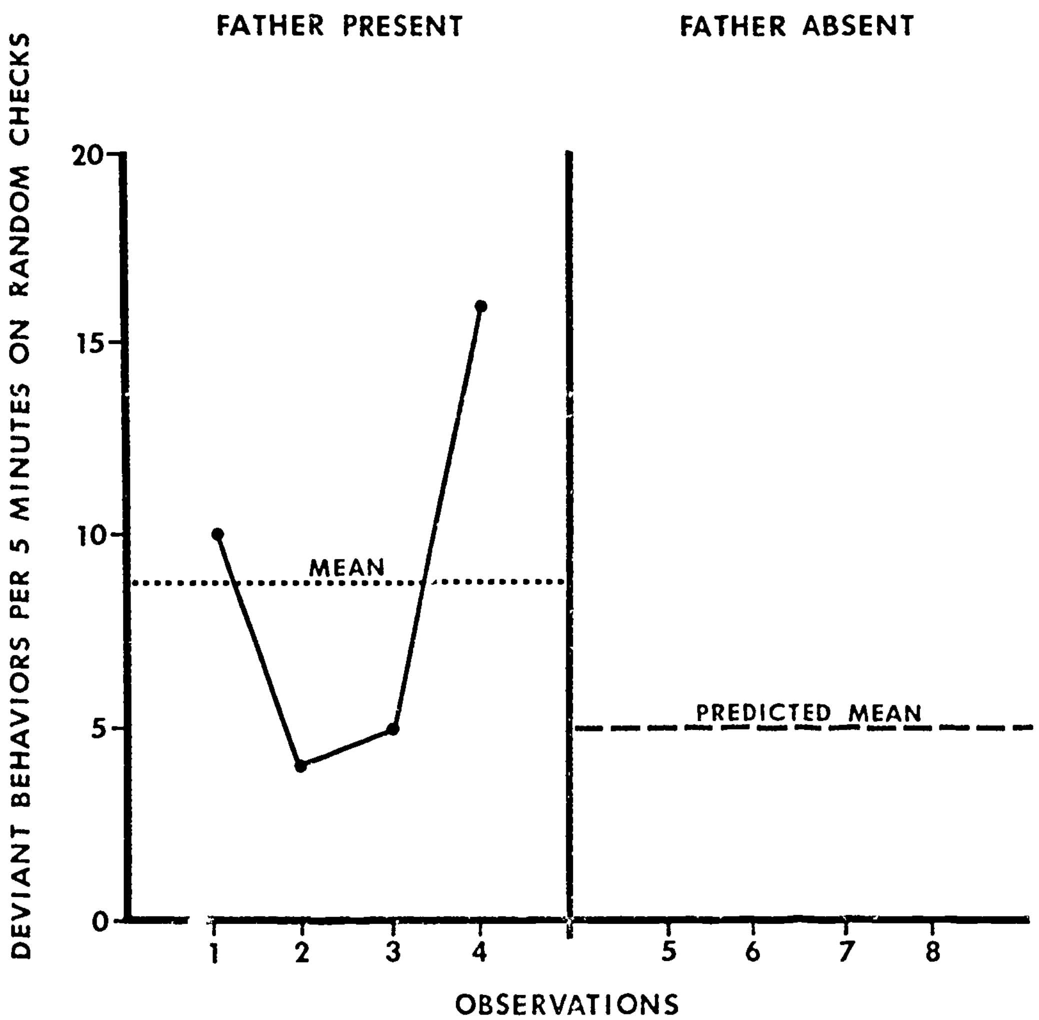


FIGURE 4

POST-INTERVENTION FA GROUP (DOWN)



intensive training in child management procedures will bring the rate down to about six deviant behaviors per five-minute observations, about a 33% decrease. You may not immediately notice a decrease in deviant behavior on the FA tapes as the rates vary tremendously from one segment to another. However, the overall rates should show a decrease from the FP to the FA tapes. Also, we asked the mother not to use the more evident child management procedures while we were videotaping, such as placing Craig in isolation following each behavior problem as this would greatly disrupt videotaping. Consequently, the change in her child management procedures may not be obvious."

(End of rationale.)

The presentation of the expectancy rationales to the FA-FP subgroups two weeks later was identical to that used with the FP-FA subgroups with one exception--it was made clear that they would be seeing the FA tapes first. The direction of change in rate of deviant behaviors was sketched on the board for each group so there would be no confusion about what to expect in spite of the unnatural ordering of the videotapes, viz., post-treatment tapes before pre-treatment tapes, etc.

On the day all group members returned for post-testing and collection of their paychecks, all were debriefed as to the true purpose and the actual design of the study. The need for research on observational methods of data collection in the evaluation of child behavior therapy was also stressed. Several months following their participation in the study all subjects were mailed a summary of the results to comply with ethical requirements ensuring the integrity of the experimenter in studies involving deception of subjects.

Dependent Variables and Evaluative Criteria

Observer expectancies. Two measures were used to determine whether observer expectancies were influenced by knowledge of the experimenter's hypothesis. In view of the possible reactive effects of such measures, only unobtrusive or post-test measures were administered.

(1) Observer Assistant Inventory (Item 7). Throughout the training program an "Observer Trainee Inventory" had been administered to assess the morale of the trainees. The original inventory was slightly revised by the addition of item 7, dealing with the experimenter's prediction (see Figure 5), and relabeling it the "Observer Assistant Inventory." The revised version was administered on the seventh day of data collection as an unobtrusive measure of observer expectancy. A copy of the complete inventory can be found in Appendix B.

Figure 5

Item 7 of the Observer Assistant Inventory

The experimenter's prediction for the set of FA tapes seen by this group of observers was that the rate of deviant behavior in the family would:

+75%	+50%	+25%	0% or have unknown effects	-25%	-50%	-75%

(2) Observer Assistant Questionnaire. A questionnaire was administered following all data collection to assess the observers' comprehension of all elements of the rationale, their personal expectations for change, and any suspicions they had about the true purpose of the study. The investigator concealed all identifying data on each of the completed

questionnaires. Five graduate students acquainted with the design and expectancy rationales used in the study were asked to sort the shuffled, anonymous questionnaires into three categories according to their judgements of expectancy group membership based on responses given to the following questionnaire items:

1. In several sentences give your understanding of what this study was about.

2. What were you told about the history of the Craig family?

3. As far as you can, indicate the variable being manipulated, the specific variables (behaviors) of interest to the investigator, and the investigator's prediction about the variables measured by <u>your</u> group's observations of the videotapes.

4. What evidence or arguments were presented by the investigator to support the prediction given your group?

5. Did you have any personal expectations regarding the outcome of the study? If so, what were they and were you more motivated to see the investigator's prediction or your own confirmed by the results of this study?

A copy of the complete questionnaire can be found in Appendix C and the instructions to the five judges in Appendix D.

<u>Observations of deviant behavior on the FP and FA tapes</u>. Observers coded the behavior of the subject, Craig, and the responses of family members to his behavior every six seconds according to the procedures outlined in the section on observer training above. Twelve of the 21 codes were regarded as deviant codes. The mean rate of the 12 deviant behaviors reported by the observers for all family members on the FP and FA tapes was the major dependent variable of the study.

<u>Observer Accuracy during Observation of the FP and FA Tapes</u>

Observers were told prior to the collection of observation data from

Craig tapes that their accuracy would be randomly spot checked during each of the eight data collection sessions (as included in the rationales presented to all three groups). One of the five segments of recorded family interaction was randomly selected from the 25-minute tape and carefully coded and recoded by two of the trainers as outlined in the section on the preparation of the videotape recordings above. This standard coding serves as a common criterion against which observer accuracy for all three expectancy groups could be measured. The mean observer accuracies for the three groups are outlined in Table 2. An F test across the three expectancy groups suggests that there were no group differences during the two-week data collection period.

Table 2

Mean Observer Accuracies for the Three Expectancy Groups on Spot Checks Made During Observations of the Craig Family Videotapes

Group	N	Father-present tapes	Father-absent tapes	Grand mean*
Increase	11	59.0%	58.0%	58.5%
Control	6	58.0%	58.9%	58.4%
Decrease	11	57.0%	58.2%	57.6%

* $F = 0.20$; df = 2, 25; N.S. at .25 level

Specific Hypotheses and Data Analysis

It was predicted that the presentation of parallel but opposing rationales for the study would result in differing expectancies across experimental groups as measured by the "Observer Assistant Inventory" and the "Observer Assistant Questionnaire."

It was also predicted that such group expectancies would differentially affect the obse vations of the same videotape recordings of the Craig family across the three experimental groups such that:

(1) The Control group would be less sensitive to the dependent variable and report lower frequencies cf deviant behavior across both the FP and FA conditions than the Increase and Decrease groups, and

(2) The frequencies of deviant behavior reported by the Increase and Decrease groups would interact across the FP and FA conditions attributable to confounding observer bias from differing group expectancies.

Given evidence of confounding observer bias, it was predicted that sub-analyses would reveal relationships between the magnitude of observer bias and (a) temporal proximity to the expectancy manipulation, (b) deviant behaviors targeted vs. nontargeted for change in the Increase and Decrease rationales, and (c) observer accuracy.

CHAPTER III

RESULTS

The first section describes the effect of the expectancy manipulation upon two self-report measures of observer expectancy. The second section examines the effects of the expectancy manipulation upon the observations of deviant behavior.

The Effect of the Expectancy Manipulation

Observer Assistant Inventory. Twenty-seven of the observers responded to item 7 describing their expectancies regarding experimental outcome. This reflected their understanding of the research project on day seven of the data collection period. The mean responses of each expectancy group to item 7 are presented in Table 3. The means obtained roughly approximate the 33% increase, 0% change, and 33% decrease predictions included in the expectancy rationales presented to the Increase, Control, and Decrease groups, respectively.

Table 3

Observers' Recall of the Investigator's Prediction Regarding Experimental Results

	Group		
	Increase	Control	Decrease
Group's mean response	+26% (N = 11)	-4% (N = 6)	-42% (N = 10)

$F = 6.88$; df = 2, 24; $p < .01$

Observer Assistant Questionnaire. Twenty-seven of the observers completed the Observer Assistant Questionnaire (Appendix C) administered as a post-test measure of comprehension and acceptance of the expectancy rationales. A perfect assortment of the 27 questionnaires into appropriate expectancy groups by the five judges would result in 55 correct assortments of the 11 Increase questionnaires, 25 correct assortments of the five Control questionnaires, and 55 correct assortments of the 11 Decrease questionnaires. Of the grand total of 135 judgements, only seven were incorrect. The results are presented in Table 4.

Table 4

Judges' Assortments of 27 Post-test Questionnaires into Three Categories

Judgement	Group		
	Increase	Control	Decrease
Increase	54	0	1
Control	1	22	2
Decrease	0	3	52
	(N = 11)	(N = 5)	(N = 11)

$X^2 = 240$; $df = 4$; $p < .001$

Responses to questionnaire item 5 describing the observers' personal expectations were selected for separate analysis. Again, a perfect assortment of the 27 questionnaire responses by the five judges would result in 55, 25, and 55 correct assortments of the 11 Increase, five Control, and 11 Decrease group responses, respectively. Of the total of

135 judgements of persona' expectations, approximately half (66) were in agreement with the expectancy rationale presented to their group.[9] Only 4% of the judgements (6) suggested personal expectations opposed to the experimenter's as presented. The members of the control group were generally without personal expectations. The results are presented in Table 5. Separate instructions to the judges for sorting the questionnaire responses to item 5 are in Appendix E.

Table 5

Judgements of Observers' Personal Expectations Regarding Experimental Results

Personal expectation	Group		
	Increase	Control	Decrease
Increase	19	5	0
No change	30	20	28
Decrease	6	0	27
	(*N* = 11)	(*N* = 5)	(*N* = 11)

$X^2 = 45.14$; df = 4; $p < .001$

The Effect of Differential Expectations upon Reported Observations

A 3 x 2 analysis of variance with repeated measures (Kirk, 1968, pp. 279-281) permits a test of both of the predictions regarding the effect of the expectancy manipulation upon the reported observations of deviant behavior:

(1) The prediction that knowledge of the specific behavior codes constituting the dependent variable of the study would produce higher

mean rates of deviant behavior of the Increase and Decrease groups than for the Control group, and

(2) The prediction that the differing expectancies among the three groups would result in increases, no change, and decreases in the reports of deviant behavior as conditions changed from FP to FA. A graph of the results across baseline,[10] FP, and FA conditions can be found in Figure 6.

The mean rates of deviant behavior observed per five minutes of family interaction are presented in Table 6.

Table 6

Mean Rates of Deviant Behavior per Five Minutes Reported by Expectancy Group across FP and FA Conditions

Group	FP	FA
Increase (N = 11)	7.273	6.827
Control (N = 6)	8.075	7.842
Decrease (N = 11)	8.198	7.677

The 3 x 2 analysis of variance with repeated measures in Table 7 indicates no significant main effects across groups, failing to support hypothesis (1) above. The lack of a significant interaction between groups and FP and FA conditions fails to support hypothesis (2) above.

Ancillary Hypotheses

A number of predictions regarding the specific conditions under which observer bias may occur were suggested at the end of Chapter II. They

FIGURE 6

RATES OF DEVIANT BEHAVIORS REPORTED
BY THREE GROUPS WITH DIFFERING EXPECTATIONS

(FP-FA & FA-FP COUNTERBALANCED)

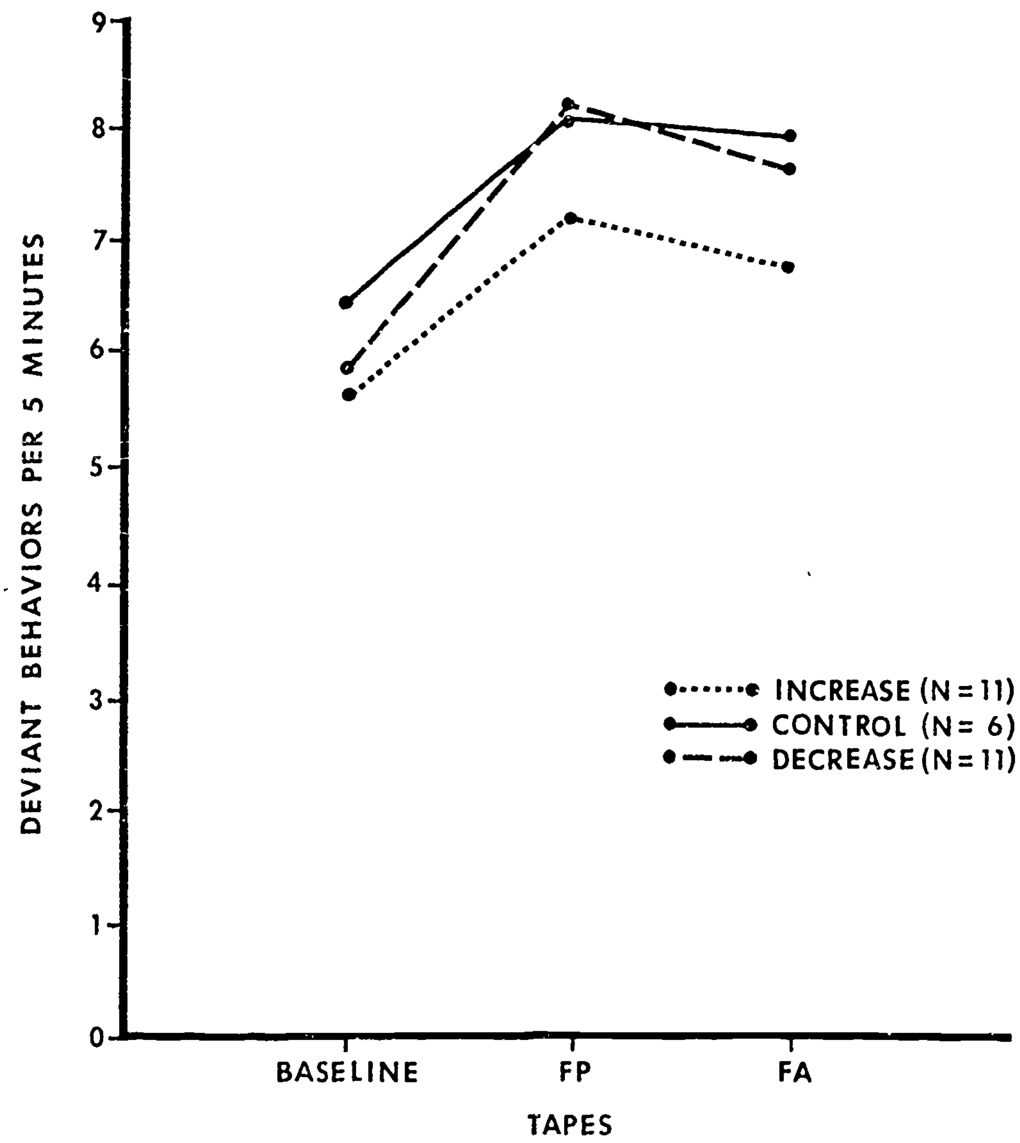


Table 7

Analysis of Variance with Repeated Measures for Mean Rates of Deviant Behavior by Expectancy Group Across FP and FA Conditions

Source	SS	df	MS	F
Between subjects	160.151	27		
Rows	10.703	2	5.351	0.90
Subjects within groups	149.448	25	5.978	
Within subjects	25.461	28		
Columns	2.584	1	2.584	2.84
Rows x columns	0.163	2	0.081	0.09
Columns x subjects within groups	22.714	25	0.909	
Total	185.613	55		

included the possibility that observer bias may be a function of temporal proximity to the presentation of the expectancy rationale, targeted vs. nontargeted deviant behaviors, and/or observer accuracy. Each of these predictions will be examined below.

The prediction that observer bias may occur only on the days when the expectancy rationale was presented (day one of the FP and FA conditions) and "wash out" on subsequent days was tested by plotting the data by days. Visual inspection of the data collected on day one of the FP and FA conditions vs. all other days, presented in Figure 7, does not suggest such an interaction. A repeated measures analysis of variance across the eight days of data presented in Table 8 is summarized in Table 9. No significant

FIGURE 7

RATES OF DEVIANT BEHAVIOR BY DAYS

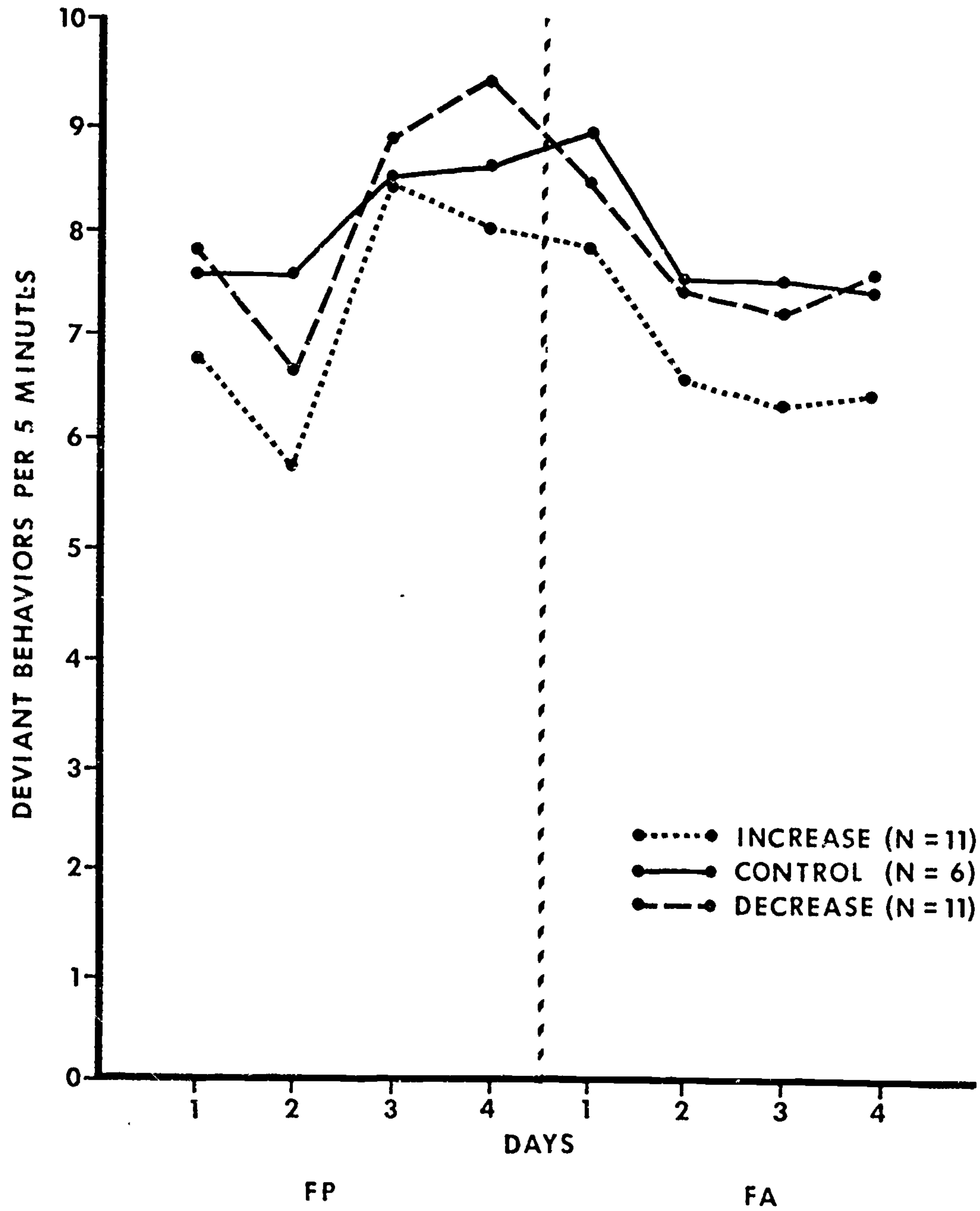


Table 8

Mean Rates of Deviant Behavior per Five Minutes Reported by Expectancy Group across FP and FA Conditions by Days

Group	FP				FA			
	1	2	3	4	1	2	3	4
Increase (N = 11)	6.818	5.745	8.473	8.055	7.873	6.600	6.345	6.491
Control (N = 6)	7.567	7.567	8.500	8.667	8.867	7.533	7.533	7.433
Decrease (N = 11)	7.764	6.691	8.900	9.436	8.400	7.436	7.255	7.618

Table 9

Analysis of Variance with Repeated Measures for Mean Rates of Deviant Behavior by Expectancy Group Across FP and FA Conditions by Days

Source	SS	df	MS	F
Between subjects	640.552	27		
Rows	42.799	2	21.399	0.89
Subjects within groups	597.754	25	23.910	
Within subjects	547.769	196		
Columns	138.777	7	19.825	8.73*
Rows x columns	11.552	14	0.825	0.36
Columns x subjects within groups	397.439	175	2.271	
Total	1,188.321	223		

* $p < .0005$

interaction between expectancy group and eight days of data collection were found. The temporal proximity hypothesis remains unsupported. (The significant columns effect was due to the fact that the sets of tapes were matched on rate of deviant behavior across FP and FA conditions, but not by days within conditions.)

The reader will recall that certain deviant behavior codes were "targeted" for change in the presentations of the expectancy rationales to the Increase and Decrease groups. In the first presentation of the rationales to the two groups, the experimenter strongly suggested that yells (YE), whines (WH), and aversive behaviors (HR) would change. In the second presentation to both groups, it was suggested that changes would also be observed in teases (TE), disapprovals (DI), and aversive commands (CN). Visual inspection of the data for these six deviant behaviors presented in Figure 8 reveals no trends for crossed, sprayed, or monotonic interactions between expectancy group and FP-FA conditions.

The data for targeted deviant behaviors are presented in Table 10.

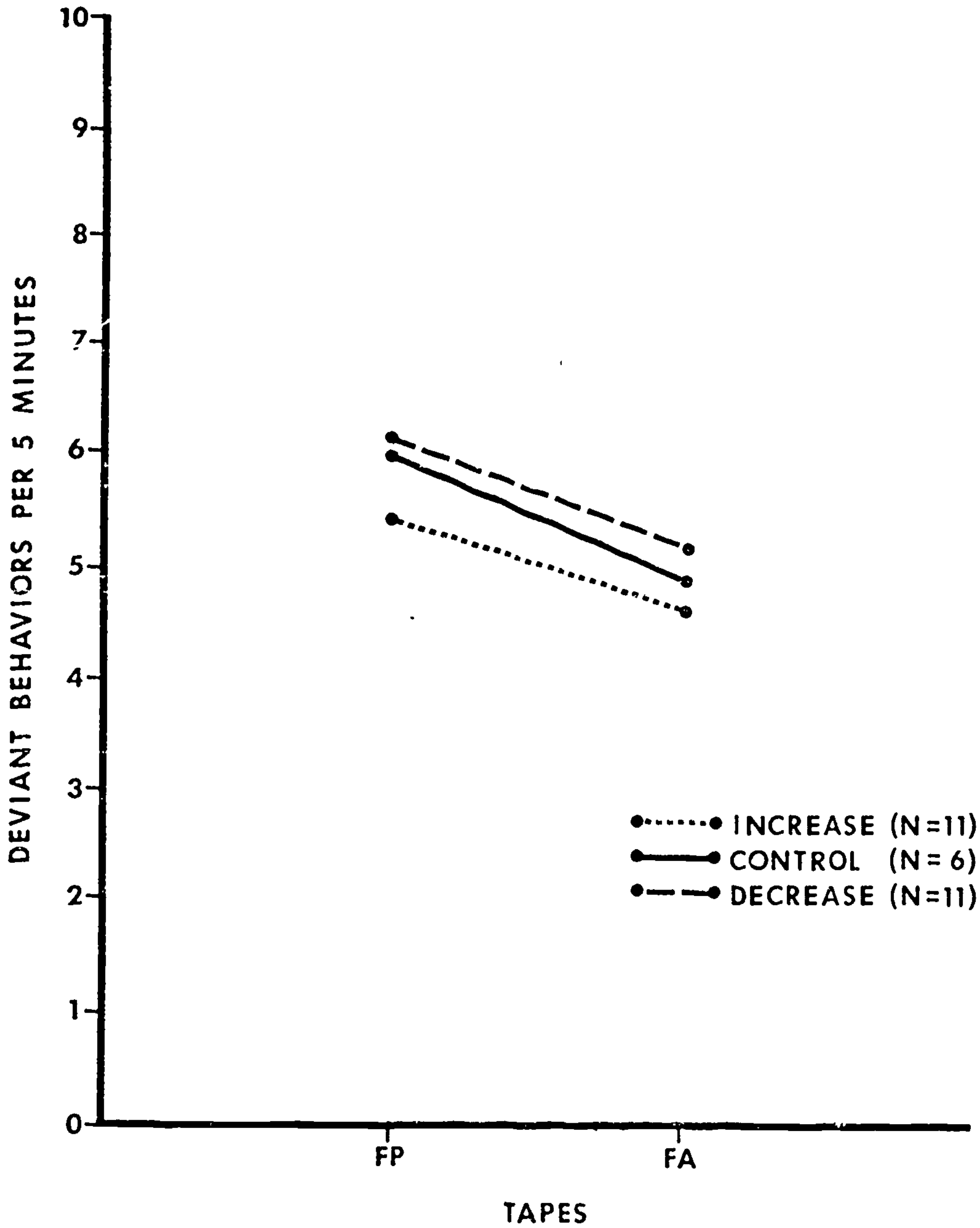
Table 10

Mean Rates of Six Targeted Deviant Behaviors per Five Minutes by Expectancy Group across FP and FA Conditions

Group	FP	FA
Increase (_N_ = 11)	5.444	4.723
Control (_N_ = 6)	5.943	4.901
Decrease (_N_ = 11)	6.155	5.241

FIGURE 8

RATES OF TARGETED DEVIANT BEHAVIORS



Further analysis of the data requires that the possibility of sequence effects resulting from the order in which the FP and FA tapes were presented be ruled out. Consequently, the data from the original analysis presented in Table 7 above were regrouped according to first and second presentation rather than by FP and FA tape sets. Visual inspection of this regrouping of the data presented in Figure 9 suggests no main effects due to the order of presentation such as could be attributed to practice, instrument decay, etc.

The data for the three expectancy groups across first and second presentations of videotape sets can be found in Table 11.

Table 11

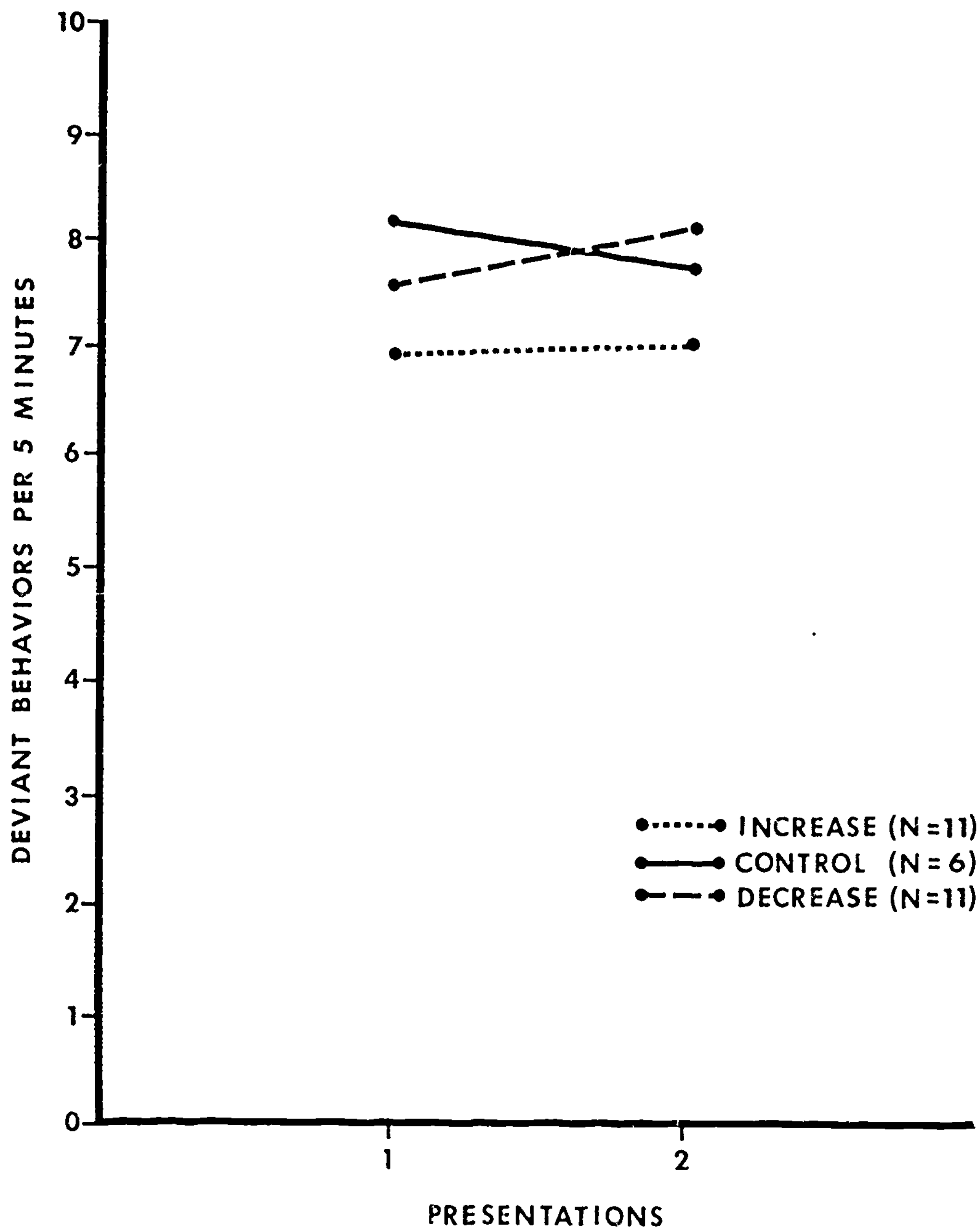
Mean Rate of Deviant Behavior per Five Minutes by Expectancy Group Across First and Second Presentations of Tape Sets

Group	First set	Second set
Increase (N = 11)	6.940	7.040
Control (N = 6)	8.200	7.725
Decrease (N = 11)	7.545	8.295

We may now regroup observers regardless of order in which they saw the FP and FA tapes, permitting a test of the prediction that observer bias may be a function of observer accuracy. All the observers within each expectancy group were ranked according to their mean observer accuracy obtained during the third week of the observer training program. The lowest one-half of each of the expectancy groups was selected

FIGURE 9

ANALYSIS OF SEQUENCE EFFECTS



out. Visual inspection of the low observer accuracy data presented in Figure 10 suggests the presence of observer bias, especially in the Increase group. An analysis of variance with repeated measures was run on the data presented in Table 12 to test the low observer accuracy hypothesis. The results are summarized in Table 13. The trends noted in Figure 10 were not statistically reliable, failing to support the low observer accuracy hypothesis.

Table 12

Mean Rates of Deviant Behavior per Five Minutes by Expectancy Group Across FP and FA Conditions for Low Accuracy Observers

Group	FP	FA
Increase (N = 6)	7.467	7.217
Control (N = 3)	7.600	6.883
Decrease (N = 6)	8.172	7.467

Power Analysis of a Related Design

Inferences regarding null hypotheses have been generally discouraged in the past. However, a number of statisticians (Bakan, 1966; Binder, 1963; Grant, 1962; La Forge, 1967; Natrella, 1960; Nunnally, 1960; Rozeboom, 1960) have suggested the use of confidence intervals and/or power analysis when inferences about a null hypothesis are of interest. In view of the failure to reject the null hypothesis of no observer bias in the original analysis presented in Table 8 above, it was decided

FIGURE 10

RATES REPORTED BY
LOW ACCURACY OBSERVERS

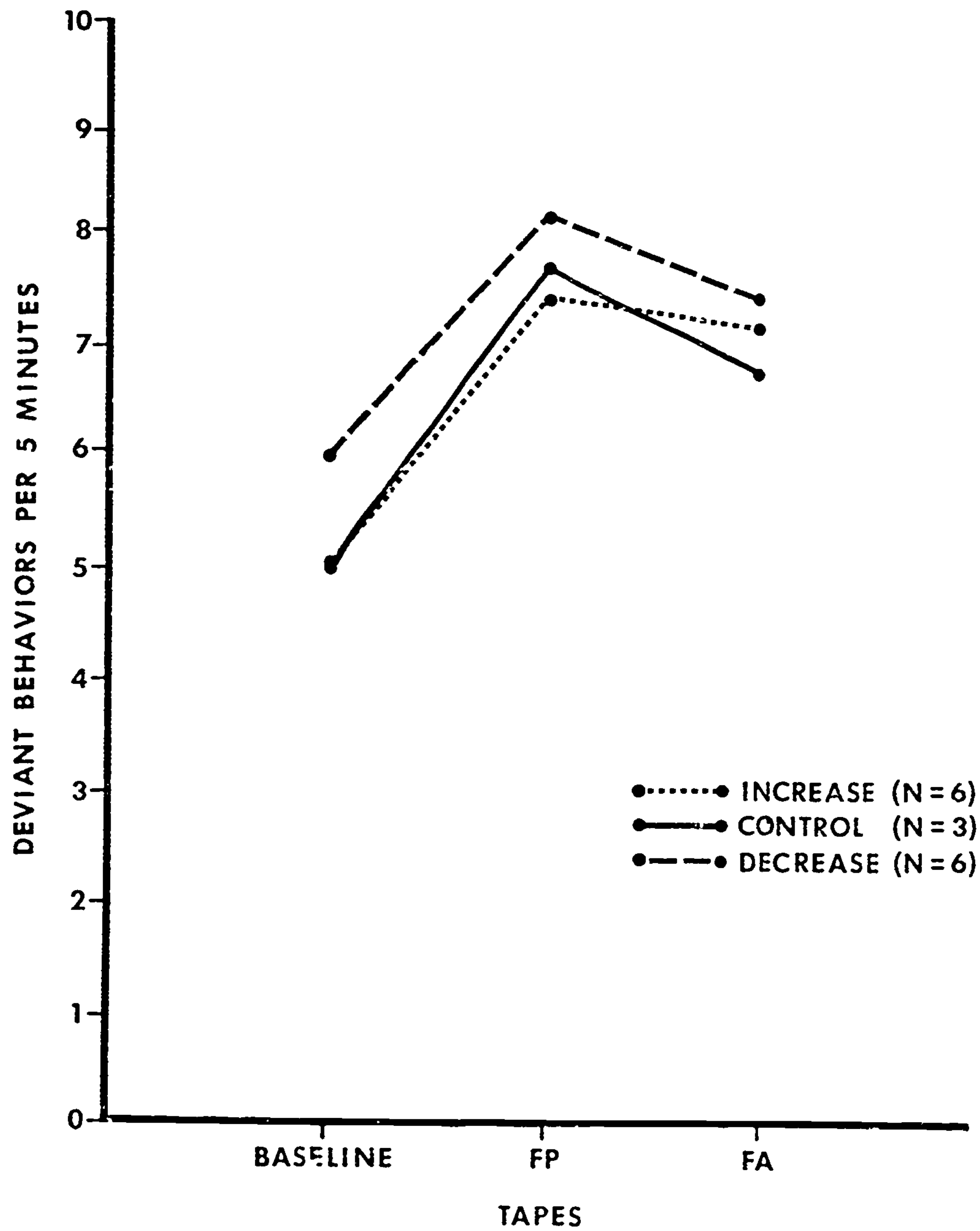


Table 13

Analysis of Variance with Repeated Measures for Mean Rates of Deviant Behavior by Expectancy Group Across FP and FA Conditions for Low Accuracy Observers

Source	SS	df	MS	*F*
Between subjects	102.531	14		
Rows	1.919	2	0.959	0.11
Subjects within groups	100.612	12	8.384	
Within subjects	12.918	15		
Columns	2.070	1	2.070	2.37
Rows x columns	0.379	2	0.190	0.22
Columns x subjects within groups	10.469	12	0.872	
Total	115.449	29		

to conduct a power analysis. Generation of a confidence interval around the null hypothesis of no observer bias would suggest what magnitude of observer bias would first be detected as significant by the present design.

Unfortunately, too many parameters are unknown to permit a power analysis with a repeated measures analysis of variance design. However, a related design that achieves a crude test of statistical interaction between the Increase and Decrease groups across the FP and FA conditions is the *t*-test of differences between differences (Walker & Lev, 1953, pp. 158, 166). Assuming no difference between Increase and Decrease groups in the FP condition, *N*s for various alternate differences in the FA condition such as 0.8, 1.5, and 2.3 deviant behaviors per five minutes could be computed. Such "sprayed interactions" would suggest mean observer biases in the two groups of five, 10, and 15%, respectively, when compared to a mean of 7.74 deviant behaviors obtained in the FP condition. Trial and error runs with $s_p^2 = 1.687$, $\alpha = .10$ and $\beta = .10$, indicated that a minimal bias of 9% for each of the expectancy groups would be required before the t-test would detect the bias as significant with an N of 11 in each group.

By extrapolation, it can be inferred that the repeated measures analysis of variance design used in this study would detect minimal biases of 5 - 10% with the relatively large N employed. In the judgement of this investigator, the design used was relatively sensitive to observer bias.

CHAPTER IV

DISCUSSION

Conclusions

The highly significant differences obtained between groups on the two measures of observer expectancy strongly support the hypothesis regarding manipulation of group expectations. Observers not only comprehended and accurately recalled the predictions given their group but differed significantly across groups in their personal expectations of experimental outcomes.

Differences in sensitivity to deviant behavior between the two informed (Increase and Decrease) and one uninformed (Control) groups were not obtained as hypothesized. These results fail to replicate the findings of Kass and O'Leary (1970) and Skindrud (1972) that observers informed of the dependent variables of a study report higher frequencies of those specific code categories throughout the study. This failure to replicate can be explained on the basis of much less extensive training given the Kass and O'Leary (1970) observers (only six training sessions vs. 12 in the present study) and the absence of the possible selection confound in the Skindrud (1972) study.

Repeated analyses of the data failed to produce any evidence that the successful expectancy manipulation biased the observations. Whether only observation sessions in temporal proximity to the presentation of the expectancy rationale, deviant behaviors targeted for change, or low accuracy observers were considered, no evidence for observer bias was obtained. While it is impossible to state that no biasing of the observations occurred, it is possible with the extrapolation from the power analysis of a related design reported earlier to rule out observer bias of greater than 5 - 10% per group in the present study.

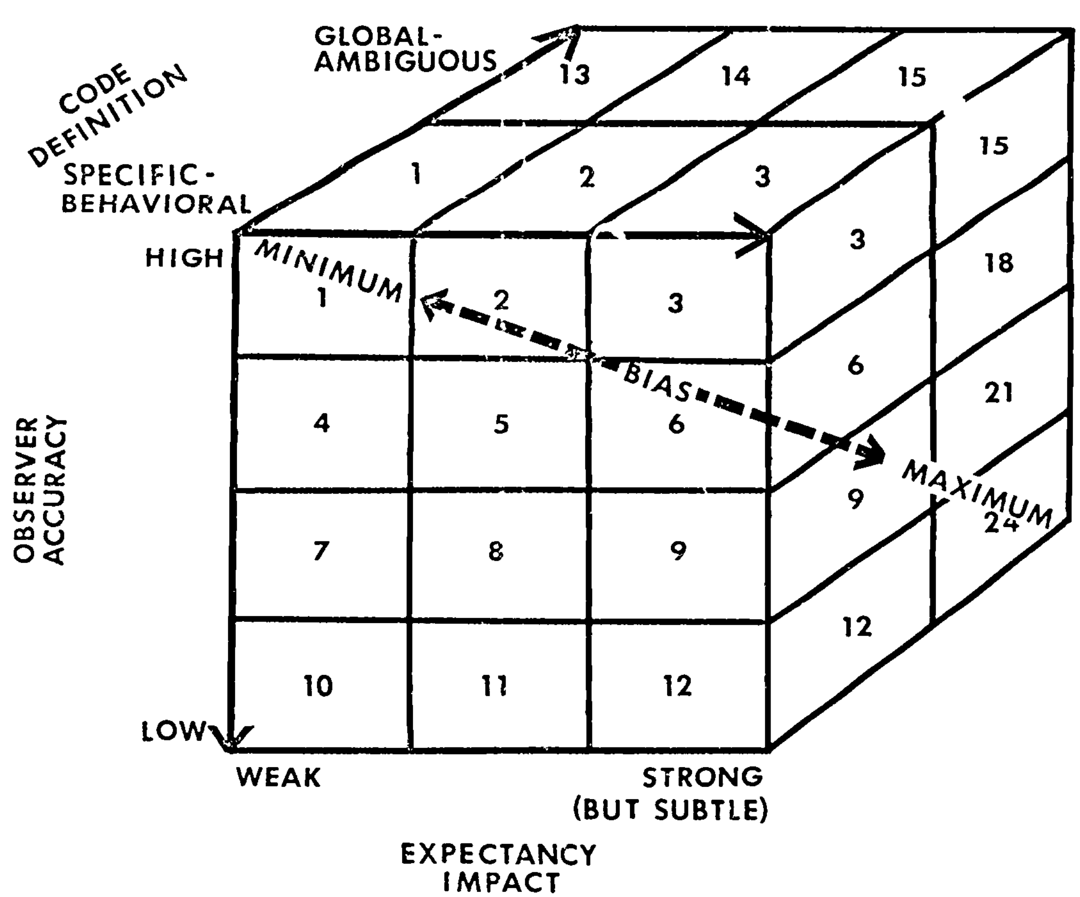
A Conceptual Model

The results of the present study, the review of published studies on observer bias in Chapter I, and recent studies made available to this investigator since data collection for the present study, suggest a three-dimensional model of conditions contributing to observer bias in experimental-field studies. To the extent that such a model is valid, it could be used to predict the conditions under which observer bias are likely to occur and allow alternatives for the control of this confounding variable. The dimensions of the model can be seen in Figure 11.

Impact of the experimenter's expectancies upon observers. The first dimension deals with the degree to which observers are influenced by the experimenter's expectancies. In only a small minority of studies of child behavior therapy (4% according to Pawlicki, 1970), observers are kept totally uninformed of the experimenter's hypothesis, placing them at the "weak" end of the expectancy impact continuum. It is more likely that observers in most studies evaluating behavior therapy are well aware of the intended outcomes and have received some feedback regarding successful "early data returns," placing them in the middle column along

FIGURE 11

A CONCEPTUAL MODEL FOR PREDICTING OBSERVER BIAS



the continuum. The present study which employed a moderately strong manipulation of observer expectancies would be located here. A strong (but subtle) impact would occur where observers would additionally be exposed to experimenter reactions during data collection which selectively reinforce the reporting of confirming data. Displays of interest and comments on the part of the therapist-experimenter (e.g., "Ummm...that data confirms our hypothesis," or, "Ah ha! Deviant behavior is beginning to drop as the family enters treatment.") are almost unavoidable where experimenter-observer interaction is not prevented. Such an impact is found in the extreme right column of the expectancy continuum.

Two small N studies suggest that the location on the expectancy impact dimension is critical and does not interact with the other dimensions in the production of observer bias. Skindrud (1972), in a pilot study for the present investigation, trained six observers and ran them through the same design with only two procedural differences. The pilot study observers were trained on videotapes of the Craig family, resulting in higher levels of observer accuracy than in the present study and, most critically, in addition to the expectancy manipulation, observers were exposed to experimenter reactions selectively reinforcing confirming data reports at eight points throughout data collection. (Deviant behavior counts were public rather than private in the Increase and Decrease groups and especially attended to when they supported the experimenter's predictions.) A significant interaction was obtained in the predicted direction, suggesting the presence of observer bias ($F = 6.45$; df = 2, 3; $p < .10$). O'Leary and Kent (1972) report a study with a multiple baseline design employing four behaviorally specified code categories, high levels of observer agreement, an expectancy manipulation, and differential experimenter reactions to confirming and disconfirming reports by the observers. The results indicated a significant effect consistent for the two categories of behavior subject to the expectancy manipulation (Orienting and Vocalization) and absent for the two unmanipulated categories (Play and Noise). These independent investigations (O'Leary & Kent, 1972; Skindrud, 1972) reveal the powerful biasing effect of experimenter reactions during data collection in spite of high levels of observer accuracy and agreement.

Rosenthal (1966, Chapter 13) reports diminishing or negative bias where attempts to reward reports of confirming data are excessive and obvious. Studies by Rosenthal and his colleagues found that graduate student experimenter-observers offered $5.00 per hour for doing a "good job" to obtain confirming data from their subjects produced less confirming data than those offered a standard rate of $2.00 per hour. Furthermore, there was a tendency for the highly rewarded experimenter-observers' data to correlate negatively with the expectancy manipulation. Post-experimental group discussion with the graduate student experimenters offered excessive rewards revealed that they were upset by the apparent attempt to bribe them to produce confirming data and were bending over backwards to be uninfluenced.

In the case of the present study, the fact that post-test questionnaire results did not produce responses indicating awareness of attempts to manipulate results and that pilot study findings demonstrated observer bias when experimenter reactions were <u>added</u> to the expectancy manipulation suggest that the expectancy manipulation in the present study was neither obvious nor excessive.

Observer accuracy. A second dimension of the conceptual model is the accuracy with which observers encode behaviors relative to standard definitions of the code categories used. Many factors affect observer accuracy. Major determinants of observer accuracy are the methods of assessing observer reliability. These methods can be roughly ordered on a continuum from those associated with high to low observer accuracy as follows:

(1) Observer accuracy is maximized by comparing observer reports to standard criterion codings of videotapes of the same interaction coded and recoded by a pair of well-trained observers anchored to standard code definitions until 100% agreement is obtained. Such a measure of observer accuracy is classified in the top row of the observer accuracy dimension.

(2) Experimental-field studies using the practice of assessing inter-observer agreement by pairing observers and comparing reports will be classified within row two. Since high levels of observer agreement within groups do not guarantee high levels of agreement between groups due to observer drift from standard code definitions over time (O'Leary & Kent, 1972; Romancyzk et al., 1972), this method is judged less rigorous than the preceeding one.

(3) A common practice has been to train observers to acceptable levels of observer reliability and then send them on data collection assignments unmonitored for observer accuracy or agreement. As reported earlier, such a method produces immediate and dramatic drops in levels of agreement when overt monitoring terminates (Reid, 1970; Romancyzk et al., 1972). Consequently, this method is classified in the third row from the top.

(4) A fourth method completely without rigor would be the use of observers who were haphazardly trained without observer accuracy or agreement checks and unmonitored for observer reliability throughout data collection. Studies employing such a method would be found at the bottom of the continuum.

The present study, together with a series of three studies conducted at the State University of New York at Stony Brook by O'Leary and his colleagues, shed some light on the relationship between observer accuracy and bias. The first of the Stony Brook studies (Kass & O'Leary, 1970), which has already been reviewed, did not involve overt monitoring of observer agreement during data collection and did produce significant observer bias. Since the drop in observer agreement upon termination of overt monitoring is a well-replicated finding, it can be presumed that observer accuracy and agreement were low during data collection for the Kass and O'Leary (1970) study. The second study in the Stony Brook series was a dissertation by Kent with a total N of 40 observers (O'Leary & Kent, 1972). Observers were broken into two groups of five within each expectation condition during the final three days of training and for the duration of the study. Throughout observers computed interobserver agreement across rotating, randomly-formed pairs within each group. Subsequent analyses of data from these observers (O'Leary & Kent, 1972) revealed a drifting apart on the definitions of seven of the nine code categories for the groups training and working independently. A third investigation by the O'Leary group with a total N of 20 observers (O'Leary & Kent, 1971) attempted to control for observer drift by dividing observers into pairs and assigning five pairs to each expectancy condition by chance.

Assuming that observer drift is a random phenomenon (O'Leary & Kent, 1972), it was felt that observer drift would be statistically controlled and unconfounded with the effects of the expectancy manipulation. According to Kent (1972), the behavioral ratings of the observers in both the second and third studies in the Stony Brook series were totally unbiased by the expectancy manipulations.

Of the studies reviewed thus far, the present study would be classified in the top row along the observer accuracy continuum, the second and third studies in the Stony Brook group in the second row, the first of the Stony Brook group (Kass & O'Leary, 1970) and the Scott, Burton, and Yarrow (1967) study in the third row, and the Rapp (1965) and Azrin et al. (1961) studies along the bottom. As can be seen, observer bias first appears along the continuum at the point where no method of overtly monitoring observer agreement is used during data collection (row three from the top) in the Kass and O'Leary (1970) and Scott, Burton, and Yarrow (1967) studies. One can conclude that for the type of coding systems and expectancy manipulations used, overt monitoring of observer agreement during data collection may be a critical variable in the prevention of observer bias. One study (Skindrud, 1972) which attempted an indirect comparison of observer bias under overt and covert monitoring of observer agreement did not find bias under either condition. However, the study was limited by unequal error variances in the two observation conditions and relatively small Ns. Replication of the Skindrud (1972) study on a large scale with error variance in the covert monitoring condition controlled would be a strong test of the validity of the model presented here.

Observation code. The third dimension of the conceptual model deals with the degree to which code definitions in observationa' studies are specific and behaviorally defined vs. global or ambivious. Global judgements are to be distinguished from specific code definitions in that they may contain a great number or variety of specific behavioral events and require some degree of inference not based on direct observation. Global categories such as "aggressive," "dependent," and "immature" would be common examples.

Classification of codes as specific and behavioral is more difficult as it is possible to have a relatively specific code such as Yell or Whine without clear behavioral definition. The problem with the observation of such code categories is they fall on continua with other categories such as Talk. How loud does a Yell or how nasal does a Whine have to be to qualify as a deviant behavior rather than Talk? Such categories are more difficult to define behaviorally (without the aid of instrumentation such as a decibel meter) than categories such as Hit or Command which appear more discrete.

If code categories are either global or ambiguous they are classified within the row to the rear of the model. If they are both specific and behaviorally defined they are classified in the row to the front of the code definition dimension.

Studies employing global judgement clearly are influenced by observer expectancies. O'Leary and Kent (1972) asked observers to report their "perceptions" of change at the conclusion of the second study in the Stony Brook series. Within each of the two expectancy groups, half of the observers were shown videotapes that contained the predicted change and half saw tapes with no actual change in rates of disruptive behavior. Surprisingly, the observers' ratings of their global perceptions of change were significantly associated with the expectancy rationales presented their group but not with the actual changes occurring on the tapes coded. It will be recalled that when specific code categories were used these same observers collected data with high levels of agreement and no evidence of observer bias.

Observers in the studies by Rapp (1965), cited by Rosenthal (1966), and by Azrin et al. (1961) employing global categories such as "above and below par" and "expression of personal opinions," respectively, were very susceptible to observer bias.

No studies are available which used exclusively ambigious-specific code categories. The present study, the three Stony Brook studies, and the Scott, Burton, and Yarrow (1967) study all used relatively behaviorally defined code categories. Had the Stony Brook studies used less behaviorally defined code categories, this investigator would have predicted the occurrence of significant observer bias for all three studies. With less behaviorally defined code categories, monitoring observer agreement during data collection may not have been sufficient to prevent observer drift in the direction of the expectancy manipulation in the second and third studies of the Stony Brook series. O'Leary and Kent (1972) report some observer drift in all three of the studies. However, in the second and third studies apparently the drift from standard code definitions was not great or was uninfluenced by observer expectancies due to the overt monitoring. Dropping out overt monitoring in the first

(Kass & O'Leary, 1970) study opened the flood gates to observer drift and significant bias resulted.

Limitations of the Present Study

The model above suggests the major limitation of the present study. Most experimental-field studies fall into rows two and three along with the observer accuracy dimension. They attempt to maintain quality data by the pairing of observers and checking observer agreement throughout data collection or omit observer agreement checks altogether at the conclusion of observer training. The present study took precautions to prevent observer drift through the use of random observer accuracy checks against standard criterion codings of observed interaction throughout data collection. Consequently, generalization of the results to the bulk of experimental-field studies employing naturalistic methods of data collection is questionable.

It does not appear that the external validity was curtailed significantly by the use of observations of videotaped rather than *in vivo* interaction. Observations were of relatively unstructured family interaction recorded around the dinner hour in the home of the Craig family. Also, O'Leary and Kent (1972) report comparisons of three observational media (*in vivo*, behind a one-way mirror, and via closed-circuit TV) where awareness of observation was held constant, controlling subject reactivity. No significant differences occurred across media for an *N* of three observers per group except for one of the nine code categories, Vocalization. The latter was attributed to inadequate audio pick-up by the sound system used in the one-way mirror and closed-circuit TV conditions. Frequencies of Vocalization were lower for those media. In the present study, subject reactivity to the videotape cameras on the part of the Craig family was controlled by the matching of FP and FA tapes on rates of deviant behavior, mentioned earlier.

Recommendations for the Control of Observer Bias

(1) Uninformed observers without culturally determined expectations regarding experimental outcome appears to be the best control against observer bias. If well trained and monitored for observer accuracy, observer sensitivity to the behaviors of interest should be unaffected. Where it is impractical to keep an entire observer staff uninformed, it may be possible to periodically add to the staff one uninformed observer freshly trained and anchored to standard code definitions. Such a "calibrating observer" can be used as a standard against which observer agreement and objectivity can be assessed while providing an incentive for the informed observers to avoid bias (Skindrud, 1972).

(2) If the problem being investigated is obvious and expected outcomes known as in evaluations of child behavior therapy, placebo control groups can be used (Campbell & Stanley, 1966; Walter & Gilmore, 1972). Observers can be kept uninformed regarding group membership controlling expectancy effects.

(3) Where convincing placebo groups are impractical, observers can be assigned to each subject on a rotating basis so that no observer collects data on the same family or classroom more than once. It then becomes difficult for observers to infer the treatment status within subjects. Potentially confounding expectancies can be assessed by

having observers guess the treatment status following each observation as Johnson and Bolstad (1972) have done. If observer "guesstimates" are unrelated to actual treatment status, the investigator may conclude his design is unconfounded by observer bias.

(4) Where impossible to keep observers uninformed of treatment status, videotape recordings during baseline and treatment conditions may be necessary. Observers can be kept totally uninformed of treatment conditions by removal of all identifying data and randomizing the order in which the tapes are observed and coded.

(5) Where the above alternatives are impractical, it is recommended that every effort be made to use specific, behaviorally defined codes with well trained observers anchored to standard code definitions throughout data collection. O'Leary and Kent (1972) and Johnson and Bolstad (1972) provide a number of practical suggestions for the control of observer drift and the maintenance of observer accuracy. Presumably, observers who are "locked in" to well defined code categories will be less susceptible to experimenter expectancy effects unless they intentionally distort the data as in the Azrin et al. (1965) study. The latter appears to be a major problem only where undergraduate observers are used (Rosenthal, 1966, Chapter 3).

BIBLIOGRAPHY

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

Bakan, D. The test of significance in psychological research. Psychological Bulletin, 1966, 66, 423-437.

Barber, T. X., & Silver, M. J. Fact, fiction, and the experimental bias effect. Psychological Bulletin Monograph, 1968, 70 (No. 6, Part II), 1-29.

Bendig, A. W. The development of a short form of the manifest anxiety scale. Journal of Consulting Psychology, 1956, 20, 384.

Binder, A. Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. Psychological Review, 1963, 70, 107-115.

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

Grant, D. A. Testing the null hypothesis and the strategy and tactics of investigating theoretical models. Psychological Review, 1962, 69, 54-61.

Hathaway, S. R. Some considerations relative to nondirective counseling as therapy. Journal of Clinical Psychology, 1948, 4, 226-321.

Johnson, S. M., & Bolstad, O. D. Methodological issues in naturalistic observations: Some problems and solutions for field research. In F. W. Clark & L. A. Hamerlynck (Eds.), Critical issues in research and practice: Proceedings of the Fourth Banff International Conference on Behavior Modification. Champaign, Illinois: Research Press, 1972, in press.

Johnson, S. M., & Lobitz, G. Demand characteristics in naturalistic observation. Unpublished manuscript, University of Oregon, Eugene, 1972.

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

Kennedy, J. L., & Uphoff, H. F. Experimenters on the nature of extrasensory perception: III. The recording error criticism of extra-chance scores. Journal of Parapsychology, 1939, 3, 226-245.

Kent, R. N. Personal communication. March 1972.

Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, California: Brooks/Cole, 1969.

La Forge, R. Confidence intervals or tests of significance in scientific research. Psychological Bulletin, 1967, 68, 446-447.

Laszlo, J. P., & Rosenthal, R. Subject dogmatism, experimenter status, and experimenter expectancy effects. Unpublished manuscript, Harvard University, 1967.

Natrella, M. G. The relation between confidence intervals and tests of significance: A teaching aid. American Statistics, 1960, 14, 20-22, 38.

Nunnally, J. The place of statistics in psychology. Educational and Psychological Measurement, 1960, 20, 641-650.

O'Leary, K. D., & Kent, R. N. Behavior modification for social action: Research tactics and problems. Paper presented at the Fourth Banff International Conference on Behavior Modification, Banff, Alberta, Canada, March 1972.

Orne, M. T. Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow (Eds.), Artifact in behavioral research. New York: Academic Press, 1969.

Patterson, G. R., Cobb, J. A., & Ray, R. S. A social engineering technology for retraining the families of aggressive boys. In H. Adams & L. Unikel (Eds.), Issues and trends in behavior therapy. Springfield, Illinois: Chas. C. Thomas, 1972, in press.

Patterson, G. R., Ray, R. S., Shaw, D. A., & Cobb, J. A. Manual for coding family interactions. Sixth revision. Available from ASIS National Auxiliary Publications Service, c/o CMM Information Service, Inc., 909 Third Avenue, New York, N. Y. 10022. Document #01234.

Pawlicki, R. Behaviour-therapy research with children: A critical review. Canadian Journal of Behavioural Science, 1970, 2 (3), 163-173.

Pfungst, O. Clever Hans: A contribution to experimental, animal, and human psychology. (Trans. by C. L. Rahn.) New York: Holt, 1911.

Rapp, D. W. Detection of observer bias in the written record. Unpublished manuscript, University of Georgia, 1965. (Cited by R. Rosenthal, Experimenter effects in behavioral research. New York: Appleton-Century-Crofts, 1966, p. 21.)

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

Romanczyk, R. G., Kent, R. N., Diament, C., & O'Leary, K. D. Measuring the reliability of observational data: A reactive process. Paper presented at the Second Annual Symposium on Behavioral Analysis in Education, Lawrence, Kansas, May 1971.

Rosenthal, R. Interpersonal expectations: Effects of the experimenter's hypothesis. In R. Rosenthal & R. L. Williams (Eds.), Artifact in behavioral research. New York: Academic Press, 1969.

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

Rosenthal, R., Friedman, C. J., Johnson, C. A., Fode, K. L., Schill, T. R., White, C. R., & Vikan-Kline, L. L. Variables affecting experimenter bias in a group situation. Genetic Psychology Monographs, 1964, 70, 271-296.

Rosenthal, R., & Hall, C. M. Computational errors in behavioral research. Unpublished data, Harvard University, 1968.

Rosenthal, R., & Lawson, R. A longitudinal study of the effects of experimenter bias on the operant learning of laboratory rats. Journal of Psychiatric Research, 1964, 2, 61-72.

Rostand, J. Error and deception in science. New York: Basic Books, 1960.

Rozeboom, W. W. The fallacy of the null hypothesis significance test. Psychological Bulletin, 1960, 57, 416-428.

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

Skindrud, K. Field evaluation of observer bias under overt and covert monitoring of observer reliability: Two preliminary studies. Oregon Research Institute Research Monograph, 1972, 12, No. 7.

Smith, H. L., & Hyman, H. H. The biasing effect of interviewer expectations on survey results. Public Opinion Quarterly, 1950, 14, 491-506.

Snow, E. E. Unfinished Pygmalion. Contemporary Psychology, 1969, 14, 197-199.

Thorndike, R. L. Pygmalion in the classroom: A review. Teacher College Record, 1969, 70, 805-807.

Walter, H. I., & Gilmore, S. K. Placebo vs. social learning effects in parent training procedures designed to alter the behaviors of aggressive boys. Behavior Therapy, 1972, in press.

White, G. The effect of observer presence on mother and child behavior. Unpublished doctoral dissertation, University of Oregon, 1972, in preparation.

FOOTNOTES

1. The work for the present investigation and reported herein was performed pursuant to Contract OEC-X-72-0001 (057) with the United States Department of Health, Education and Welfare, Office of Education, Karlton Skindrud, project director. The pilot work and project director were supported by United States Public Health Service (National Institute of Mental Health, Center for Studies of Crime and Delinquency) Grant #MH 15985-02 to Gerald R. Patterson.

2. The work sample of the observation task consisted of two short videotape segments of the alpha-numeric code presented at different rates. Applicants were asked to copy the alpha-numeric (e.g., "1 PN - 4 CR") from the TV monitor as accurately as they could onto a protocol sheet. They heard a tone every 30 seconds cuing them to move down one line on their sheet.

In the first work sample, the alpha-numeric code was presented at the rate normally coded in the field. The second work sample presented the interactions at twice normal rate. The two videotape segments lasted five and 2 1/2 minutes, respectively. The first was presented to orient the naive applicant to the observer task--most could manage it without difficulty. The second was presented rapidly to screen out those who might encounter difficulty coordinating observing and rapid handwriting. Mean observer accuracy across the 48 applicants during the second work sample was .73 with a range from .44 to .95.

3. A separate study was designed to develop a selection battery which would reliably predict observer accuracy. A multiple regression analysis of the four predictor variables employed in the present study resulted in three significant predictors of observer accuracy at the end of a three-week training program--the Employee Aptitude Survey Numerical Reasoning subtest (+.42), the work sample of the observation task (+.40), and the Bendig short form of the Taylor Manifest Anxiety Scale (+.30). The multiple correlation of these three predictors with the criterion was +.61. Analysis of the residuals suggests that these measures are best at selecting <u>out</u> applicants who find such a complex observation task difficult, particularly where the applicant is handicapped by low handwriting speed. Such a significant multiple correlation is surprising in view of the truncated distribution of applicants. Selection had eliminated the lower one-third of the original distribution.

4. Two observer training methods were compared in a pilot study prior to the present investigation--a "trial by fire" method where observers were expected to code complex family interaction throughout, and a "shaping" method where observers began by coding the behavior of <u>simply one</u> family member and, when proficient, moved on to coding simple and then complex <u>interaction</u>. Evaluation of both training methods on a common criterion suggested no differences at the end of the three-week training program. However, the "shaping" method resulted in fewer complaints from the trainees and faster progress for those trainees with less aptitude for the observation task. Consequently, the "shaping" method was selected for training observers in the present study.

5. Johnson and Bolstad (1972) were the first to make this distinction between observer accuracy and agreement clear.

6. A significant relationship between the simplicity of social interaction and observer accuracy was noted in both the pilot and present studies. Simplicity of interaction was measured by the percent of social interactions repeated consecutively within each five-minute segment of videotape coded. Correlations between simplicity and observer accuracy across 12 test tapes of the Craig family with a mean of 26% and a range from 11 to 36% repeated interactions in the pilot study was +.53. The same correlations across 11 test tapes of the Ross family in the present study was +.65.

This relationship explains the different standards for observer accuracy in observations of videotape recordings and observer agreement in field operations--70% and 85%, respectively, with the Patterson et al. (1969) and related family interaction codes. The difference between the standards for field and videotape observations is attributed to the greater complexity obtained in criterion codings of videotapes. Field observations do not permit the coding of behavior in as much detail as do repeated re-codings of videotape. This difference in simplicity is reflected in the mean number of consecutive repeated interactions reported for criterion codings of videotapes (26%) noted in the representative sampling of videotape protocols above and (41%) from a representative sampling of field observations by paired observers from the staff of the Social Learning Project. A representative sampling of field observations by solo observers not monitored for observer agreement resulted in even higher simplicity ratings (47%).

The conclusion was that a criterion of 70% observer accuracy during coding of videotaped family interaction was equivalent to a criterion of 85% observer agreement in the field. This assumption is supported by a subsequent study with the same observers coding relatively simple family interaction (50% repeated interactions) in the field. Under field conditions they obtained 83% agreement (White, 1972).

7. In the pilot study coding of the FP and FA tapes the mean rate of deviant behaviors for all family members on the FA tapes increased 60% over the rate on the FP tapes. FP and FA days were alternated during the five days of videotaping in the home of the Craig family to control for subject reactivity. These results are in essential agreement with some preliminary field observations under FP and FA conditions which suggested that the father's absence influences child management in "normal" but not "deviant" families.

8. The mean rate for the 12 deviant behavior codes reported by the 28 trainees during their coding of the 50 minutes of the Craig family baseline tapes was 6.1 with a range from 3.2 to 9.1 deviant behaviors per five minutes across the 28 trainees. It was to control these rather large individual differences in observers that the matching procedure with random assignment to groups was used.

9. The possibility existed that there are individual differences in response to an expectancy manipulation (Rosenthal, 1966, Chapter 13, p. 218). Consequently, those 13 observers judged in agreement with the experimenter's predictions were hypothesized to be the most susceptible to observer bias. Visual inspection of the rates of deviant behavior

reported by these 13 observers ($\underline{N}$ = 4 Increase, 4 Control, and 5 Decrease observers) revealed a minimal trend for observer bias among the Decrease observers. However, an analysis of variance with repeated measures showed no significant interactions, failing to support the hypothesis of bias for these observers.

10. Rates of deviant behavior on the baseline tapes were not matched with rates on the FP and FA tapes explaining the "main effect" that would result in a 3 x 3 ANOVA for treatment conditions.

APPENDIX A

Experimental revision for
FP-FA Study
October 1971

MANUAL FOR CODING OF FAMILY INTERACTIONS

Adapted from the Manual by

G. R. Patterson, R. S. Ray, D. A. Shaw, and J. A. Cobb

Oregon Research Institute and University of Oregon

The behavioral coding system described in this manual is designed to provide an accurate running account of social interaction among family members. The behavioral codes are intended to cover all events that occur in a household. By weeding out old categories and adding new ones, it is now feasible to code each behavior that occurs in a home under one of the 29 categories presented in this manual. Three main considerations have determined the current status of the coding system: very few behavioral categories should be used in order to develop a flexible code that would be relatively easy to learn; the behavioral categories should be distinct from one another; and the behavioral categories should require very little inference on the part of the observer, i.e., that the behavior be observable. The success of the system will be shown in the ease with which new observers learn the code and the ease with which the code applies to new families as they join the project.

The manual is divided into four main sections. The first will state some of the general rules regarding observation procedures, the second will provide a definition for each behavioral category, the third will provide more specific rules and use of specialized symbols that are applicable during an observation, and the fourth section will describe verbally a typical family situation and give a complete coded sheet that accurately records the behaviors of the family members in the situation.

General Rules

At any given time one family member is designated as the subject of the observer's attention. During that period of time the subject's behavior is coded alternately with other family members with whom the subject interacts. The simplest way to learn to code alternately is to observe the subject, code his behavior, look at the reactions to that behavior from other family members, code those behaviors, then look at the subject and begin the process again. Often the family member other than the subject may be coded first as the behavior of the family member precedes the subject's behavior, not only in time, but also "naturally." For example, if the son is the subject and the father tells him to empty the garbage, the command of the father would precede whatever behavior the son did in relationship to the command. Therefore, there is no hard rule that the subject behavior must precede the other family member's behavior.

A sequence of behavior is arbitrarily defined as one interaction between the subject and one or more family members. A sequence is made up of two parts, one involving the subject, and the other involving the other family member or members. A part consists of one or more units.

A unit is defined as the identifying number for the person plus the behavioral code or codes. To illustrate the sequence and its component parts, note the diagram.

SEQUENCE

PART I → SUBJECT UNIT

PART II → OTHER FAMILY MEMBER UNIT; OTHER FAMILY MEMBER UNIT

As stated previously, the subject unit need not precede the other family member unit in a sequence. The symbol Z is used when more than one family member is involved in a part of a sequence. Family members are given numbers from 1-8; 1 is reserved for the deviant child; 2 is for the father; 3 is for the mother; 4 is for the oldest child; 5 is for the next oldest child, etc. The number 9 is reserved for those cases in which three or more family members are doing a similar behavior; in that case, instead of listing each family member, the number 9 is used. The number 0 is reserved for the designation of the observer; in some cases a subject will direct behavior to the observer. The subject's behavior must be coded as well as the behavior of the observer, so the 0 is used to differentiate the observer from the family members. The codes consist of two capitalized letters that are mnemonic devices for the behaviors which are to be coded. An example of a subject and his behavior, which is one unit as well as one part of a sequence, is 1LA, which states that the deviant child laughed. A sequence which involves <u>two</u> parts, i.e., the subject's behavior and the other family member's behavior, could be 1LA 2AP. Translated, this means the deviant child laughed and the father approved of the laughter. Sometimes it is necessary to code more than one behavior to describe a unit; this procedure is acceptable up to two behaviors. Thus, 1LA PP means that the deviant child laughed and physically touched another person in a positive manner, e.g., a hug or a kiss. Likewise, it is permissible to code two behaviors for other family members: 1LA PP 2AP PP indicates that after the deviant child had laughed and given some positive physical contact, the father had approved and also had given some positive physical contact. Thus, for each person coded, it is necessary to identify him or her by number and give one behavior that he or she was doing; two is the maximum number of behaviors that can be attributed to each person in a sequence. Also two is the maximum number of persons that can be <u>individually</u> identified within a <u>part</u> of a sequence. 1LA PP 2AP PP Z 3LA PP means that not only did father approve and provide positive physical contact to the son, but mother laughed and approved as well. The example given is the longest sequence that is to be used with the behavioral code. If more than two persons are involved in part of a sequence, use the number 9 code <u>if</u> they are doing the same kind of behavior. In the example cited, if the mother and father as well as one of the other children all smiled approvingly after the child laughed, then the sequence is 1LA 9AP. When more than two persons are involved in a part of a sequence and they are exhibiting different behaviors, the observer must choose the most relevant behaviors to code. This will be explained in a later section. One additional rule in coding a sequence is that the subject is always coded separately, i.e., he is never coded with another individual as other family members are. For example, when 1 is the subject, 1LA 2AP Z 3AP is permissible, but 1LA Z 2AP 3AP is incorrect.

From a sequence the observer will build up longer streams of behaviors. Usually for each member of the family, 10 minutes of data will be collected at each observation. Two non-consecutive five-minute periods of observation will be collected for each family member. The five-minute segments are divided into 30-second intervals, each 30 seconds representing one line of data on the Behavior Coding Sheet. In each 30-second interval the observer is expected to average five sequences. The number of sequences will be determined by the behaviors of the family as well as the speed of the observer.

Behavioral Codes

This section is divided into two main sections, First Order Behaviors and Second Order Behaviors. The reason for the division into two sections is for the observer to have a knowledge of priorities in coding behaviors. It is impossible to code every behavior emitted, and many times a person will emit three or four of the behaviors listed in the manual. In order to resolve the problem and keep the number of behaviors attributable to one individual down to two per sequence, some behaviors are designated Second Order Behaviors, which means that they are never coded when a First Order Behavior can be coded. It is up to the discretion of the observer what behaviors to choose among several behaviors within the same order. Since the observer can code only two, she must pick those behaviors that best describe the social interaction that is occurring.

Not only have behaviors been divied on a priority basis, but also on whether they are verbal, non-verbal, or a combination. This is to aid the observer in cataloguing the codes, and perhaps learning them with greater ease. Behaviors are listed alphabetically within each subarea.

First Order Verbal Behaviors

CM (COMMAND): This category is used when a direct, reasonable, and clearly-stated request or command is made to another person. The statement must be sufficiently specific as to indicate clearly the behavior which is expected from the person to whom the command is directed. The command need not require immediate compliance, e.g., father tells the son that he has to mow the lawn on Saturday. However, the observer is always to indicate whether the command is complied with. In the example cited, the son could indicate verbally that he is or is not going to comply with the father's request. In those instances where the compliance will not follow directly, but is likely to occur before the observer is finished coding on the subject's observation sheet, the immediate response should be coded and when compliance or non-compliance occurs, that should be coded. For example, mother tells the child, who is the subject, to wash his hands before coming to dinner. The child tells his mother that he will and continues whatever he was doing, but in a minute he goes to the sink and washes his hands. The response to the mother's command would be the child's talking and compliance would be coded when he began washing his hands. Note that many questions are most appropriately coded as talk (TA) rather than CM. For example, "What s for dinner?" or "What time is it?" would be coded TA, while "Would you go into the living room and tell your father that dinner is ready?" or "Will you help me lift this table?" would be coded as CM.

CN (COMMAND NEGATIVE): This is a command which is very different in "attitude" from the reasonable command or request described above. This kind of command has some of the following characteristics: (1) Immediate compliance is demanded. (2) Aversive consequences are implicitly or actually threatened if compliance is not immediate. (3) It is a kind of sarcasm or humiliation directed to the receiver. An example of the implicit use of aversive consequences is indicated by the tone of voice as well as the statement: Mother tells Johnny to shut the door in a normal tone of voice; he does not comply; she then raises her voice and says, "You'd better shut that door, young man." He shuts the door. The sequence would be coded 3CM 1NC 3CN 1CO.

CR (CRY): Use this category whenever a person cries. There are no exceptions.

HU (HUMILIATE): This category should be used when the agent makes fun of, shames, or embarrasses the subject intentionally. Examples: laughing in a derisive manner at the subject when he attempts to tie his shoe; telling the subject in a firm tone of voice, "Boy, you are really stupid"; when the subject is playing a game and someone says quite strongly, "You are a cheater." The observer must be careful to differentiate between playful verbal statements or nicknames and humiliate, e.g., some people call each other "stupid" but more in terms of endearment than of humiliation. The tone of voice, as well as the language used should be considered by the observer before a decision is made to code HU or some other appropriate code.

LA (LAUGH): Used whenever a person laughs in a non-humiliating way. For example, a person tells a joke and the other people laugh at the joke. However, if one of the people who heard the joke laughed in a derogatory manner at he person for the way he told the joke, that would be coded as HU and not as LA.

WH (WHINE): Use this category when a person states something in a slurring, nasal, high-pitched, falsetto voice. The content of the statement can be of an approving, disapproving, or neutral quality; the main element is the voice quality.

YE (YELL): This category is to be used whenever the person shouts, yells, or talks loudly. The sound must be intense enough that if carried on for a sufficient time, it would be extremely unpleasant.

Non-Verbal Behaviors of the First Order

DS (DESTRUCTIVENESS): Use of this category is applicable to those behaviors by which the person destroys, damages. or attempts to damage any object; attacks on people are covered by PN. The damage need not actually occur, but the potential for damage must exist, e.g., the child starts to throw a glass, but is stopped by the father. The value of the object is of no consideration nor is the actual damage done.

HR (HIGH RATE): This category is applicable to any behavior not covered by other categories that if carried on for a long period of time would be aversive, e.g., running back and forth in the living room, jumping up and down on the floor, "rough housing." If the behaviors can be covered by other categories, e.g., YE, PN, DS, then HR is not to be used. It may happen that in a sequence of behaviors HR will be coded intermittently with more specific behaviors, e.g., the children are playing leap-frog in the house and at times one of them gives out with a scream; the coding would be the following: 1HR 4HR 1YE 4HR 1HR 4HR 1YE 4YE 1HR 4HR, etc.

PN (PHYSICAL NEGATIVE): Used whenever a subject physically attacks or attempts to attack another person. The attack must be of sufficient intensity to potentially inflict pain, e.g., biting, kicking, slapping, hitting, spanking, and taking an object rougly from another person. The circumstances surrounding the act need not concern the observer, only the potential of inflicting pain. For example, children may be playing and part of the play involves wrestling. If during the wrestling one child hits the other child or pins him down to the point where pain could result, then the act of hitting or pinning down should be coded PN.

PP (PHYSICAL POSITIVE): Use this category whenever a person touches another person in a friendly or affectionate manner, e.g., hut, pat, kiss, arm around shoulders, holding hands, ruffling hair, etc.

First Order Behaviors that may be Verbal or Non-Verbal

AP (APPROVAL): Used whenever a person gives a clear gestural or verbal approval to another person. Approval is more than attention, in that approval must include some clear indication of positive interest or involvement. Examples of approval are smiles, head nods, phrases such as, "That's a good boy," "Thank you," and "That's right."

CO (COMPLIANCE): Use this category when a person does what is aske of him in a CM, CN, or DP. Remember, compliance need not follow the previously mentioned behaviors immediately, as other behavioral sequences can intervene: 3CM 1CR 3PP 1CO.

DI (DISAPPROVAL): Use this category whenever the person gives verbal or gestural disapproval of another person's behavior or characteristics. Shaking the head or finger are examples of gestural disapproval. "I do not like that dress," "You didn't pick up your clothes again this morning," "You're eating too fast," are examples of verbal disapproval. In verbal statements it is essential that the content of the statement explicitly states disapproval of the subject's behaviors or attributes, e.g., looks, clothes, possessions, etc. DI can be coded simultaneously with CM, but never with CN, as CN always implies disapproval. An example of DI and CM being coded together is when father says to the child, "Put on a shirt before you come to the dinner table. I don't like you wearing T-shirts to dinner."

DP (DEPENDENCY): Behavior is coded DP when Person A is requesting assistance in doing a task that he is capable of doing himself. For example, mother is reading the newspaper in the evening and a child who is in junior high school requests her to look up a word in the dictionary; or a child, age 10, asks his mother to tie his shoes. Everyday requests should not be coded as DP; they must meet two criteria: that the person is capable of doing the act himself and it is an imposition on the other person to fulfill the request. For example, asking someone to pass the newspaper which is very close to the individual to who the request is directed would not be considered DP, since the person would be able to hand the newspaper to the other individual without any undue amount of effort. If the paper were across the room from where the person is to whom the request has been made, and the person would have to move to get the paper, thus unduly interrupting whatever he were doing, then the request is coded DP.

NC (NON-COMPLIANCE): This code is used when a person does not do what is requested of him by CM, CN or DP. The non-compliance can be of a verbal or non-verbal nature. If the request is not to be complied with until some later time and the person says he will not comply, then the appropriate code is NC. Care must be taken to distinguish DI from NC. For example, mother tells daughter to do the dishes; daughter says that mother is always making her work; daughter goes to the sink and begins to do the dishes; the proper coding is 3CM 4DI CO.

TE (TEASE): Use this category when a person is teashing another person in such a way that the other person is likely to show displeasure and disapproval or when the person being teased is trying to do some other behavior, but is unable to because of the teasing. For example, a child is trying to do homework and another child keeps tickling him in the ribs or turns the pages of the book that the child is using for studying. Another example would be two parents teasing a young child by saying, "You're not my boy; go away from me," and when the child goes to the other parent, he hears the same remarks. This category should be distinguished from PL, LA, HU, and PN. Many cases of teasing will fall into the PL category.

Verbal Behaviors of the Second Order

The following are lists of behaviors that should be considered by the observer as secondary in coding. If it is possible to code behaviors using the First Order behaviors, the Second Order codes should not be employed.

TA (TALK): This code is used if none of the other verbal codes are applicable.

Non-Verbal Second Order Codes

AT (ATTENTION): This category is to be used when one person listens to or looks at another person, and the categories AP or DI are not appropriate. Sometimes when listening is used as a reason for coding AT, it may be difficult to tell if the person is listening. The situation will generally resolve the question, as the person who has been "listening" may make some comment and the content of the comment will indicate that he has been listening.

NR (NO RESPONSE): This category is to be used when a person does not respond to another person. This category is applicable when a behavior does not require a response, or when behavior is directed at another person, but the person to whom the behavior is directed failes to perceive the behavior. There is a clear differentiation between NR and IG. IG is intentional non-responding and NR may be accidental, e.g., there could be a great deal of noise in the house so the person cannot hear the behavior to which a response is expected, or the person may be attending to something else in the environment, e.g., mother may be feeding the baby when an older child comes in and asks a question. Whenever behavior is specifically directed toward another person and the person does not respond it is necessary to code either NR or IG.

Specific Rules and Specialized Symbols

This section will cover the specifics that are involved in observing family interaction. The necessity of spelling out rules in detail is to ensure that every observer does exactly the same procedure as another observer so that data are comparable from one observer to another observer. Another reason for specifying rules is to make the work of the keypuncher easier; if all observers use the same symbols and follow the same technique, a keypuncher is able to read any sheet in the same manner and keypunch faster and more efficiently.

One observation sheet is used for each family member for each five-minute segment. From the observation sheet the data are then punched onto cards so the data can be entered into the computer and analyzed. At the top of the Behavior Coding Sheet are several blanks to be filled in regarding the family and the observation. The "Family Number" is a numerical number given to each family on the basis of their entry into the research project in comparison to other families that have already been in the project. The ID number is the number that is punshed in IBM cards for computer purposes. It consists of 10 digits. The "Phase" is another blank for the observer to fill. In the phase the numbers can be 1-5 inclusive. The numbers mean the following:

1. Regular baseline (meaning the family is seen for 10 consecutive week days)
2. Split baselin (the family is observed for five consecutive week days, a week or more intervenes, and the family is observed for an additional five days)
3. Intervention (the family is being seen by a therapist)
4. Follow-up (the family has already been through intervention and is now on their own)

5. A control condition

In the "subject" blank the observer puts down the appropriate number of the person who is the current focus of observation. In the "observer" blank the observer puts her initials; in the "Date" blank, the numbers for the current month, day, and year. In the "No" blank, the observer records the number of the sheet she has used for the subject for that particular observation, e.g., if it is the first five minutes that the subject is being observed, the number "1" is placed in the blank. Most of the preceding information is then coded into a 10 digit number which is placed in the "ID Number" section. The "ID Number" must always come out to be nine digits; so, if space is available for two digits and only one digit is used the other space, always to the left, is supplied with a "0." Following is a list of the spaces and the information to be filled in by use of numbers:

1-2 spaces	- Family Number
3	- Subject Number
4	- Phase Number
5, 6, 7, 8	- Month, Day, Year
9	- Sheet Number

For the month and day four spaces are available so the observer must be careful to place zeros in the appropriate places if they are needed. For example, the date is March 4th. The four spaces reserved for month and day would be 0304.

Following the blanks the observer has the codes listed alphabetically with the appropriate symbol. Then the main body of the sheet begins with a line split into five segments. Each segment is to contain a sequence of behavior. The five segments should not be considered as constricting; the observer should code more than five sequences if more than five occur.

Each line represents 30 seconds of data. And with 10 lines, five minutes of data are collected on each Behavior Coding Sheet. During the data collection there are several events that can occur; and, because of these events, a series of symbols besides the behavioral categories has been devised to record in a simple fashion what those events are. If the observer takes a break while coding a line, then the point at which the break occured is coded with the letter "U." If the subject takes a break while he is being coded within a line, the observer writes the letter "K" at that point where the subject took a break. Examples of a subject taking a break are leaving the room during an observation or being out of view of the observer for one reason or another. If the observer break occurs at the end of a line, then the letter "B" is used; if the subject break occurs at the end of a line, the letter "A" is used.

The symbols "A, B, K," and "U" are to be circuled on the Behavioral Coding Sheet so that they are clearly discernible from the behavioral codes. Sequences that are repeated on a line need not be coded using number and letters over and over again. Instead, simply put a dash and a slash, i.e., "-/." Never use these symbols at the beginning of a line even though the sequence is the same as those in the preceding lines. Always write out the first sequence of a line and then the use of the dash and slash is appropriate.

The observer is aware of each 30-second period because a timing device in the clipboard which is used to hold the Behavior Coding Sheets gives off auditory signals through an earplug every 30 seconds. A reset button on the outer casing of the timing device should be pushed at the beginning of a five-minute segment and at each time when the observer picks up from a break that has been taken at the end of a line. When a break has occurred in the middle of a line, the observer resumes coding on the same line after the break is over, until there are a total of five sequences on the line; at that point, the observer is to push the reset button and go on to the next line.

In coding the observer must have complete sequences, i.e., an identifying number for the subject, the subject's behavior, an identifying number for another person and his behavior. Parts of sequences are meaningless in this coding system. Additionally, the observer should attempt to begin each line of coding with a behavior of his current subject, if possible.

Sometimes during coding a subject, the observer will find that the subject is not interacting with other family members. Thus, the code that is proper for the response of the other family members to the subject is NR. There is a temptation to code the behavior of the other persons in the room; however, there is the danger that to do so would provide a non-random sample of the behavior of these other persons. NR should be used unless it is clear that the subject being observed is actually interacting with (or attending to) other persons.

Finally, at the end of the coding sheet lines are provided for the observer to record the situation that was going on while the subject was being observed. The observer should write terse statements or simply one-word descriptions of what was occurring, e.g., dinner, working in the kitchen, reading a newspaper, etc. When more than one descriptive statement is used, the observer should put in the line number or numbers appropriate to each statement. Also, any event that occurred that was difficult to code should be included so the observer can obtain clarification on how the action should be coded.

Sample Observation

To illustrate the use of the behav ral code, a fictional description of a family is provided. It is unlikely that all the events and behaviors that are described would occur in five minutes. In order to illustrate examples of all the rules and the use of every behavior, the time element has been speeded up considerably, e.g., dinner takes only one minute. The deviant child is Kevin, age 9; he is the family member who is being observed and is coded as number 1. His sister, Freida, age 8, is number 4. Mother is number 3 and father is number 2. The observation takes place on March 6, 1969 during baseline, and the family is the fiftieth family accepted into the research project. The sheet to be coded is the first sheet for Kevin this night.

1 Kevin is in the living room playing alone with his soldiers. Father is also in the living room and he is reading the newspaper. No interaction is going on between them. The pattern persists for approximately 20 seconds, at which time mother calls into the living room and says in a conversational tone of voice to father, "How was your day today?" Father answers, "It was fine, although it was a little hot driving home from the office tonight." Since Kevin does not attend to this interaction and is not involved in this interaction and since the parents are obviously not providing a consequence for Kevin's playing, this is coded 1PL 2NR Z 3NR.

(2) Kevin is still playing with the soldiers and Freida enters from the kitchen saying, "OK if I play with you?" Kevin says, "No, I don't like the way you play." Freida replies, "Oh, you're such a stupid idiot; you don't know if I were playing right or wrong." Father says, "Kevin, let Freida play with you." Kevin says, "No." Father says in a stern tone of voice, "Kevin, if you know what's good for you, you'll let your sister play with you." Kevin replies in a negative tone of voice, "OK." Kevin and Freida play soldiers.

(3) They continue to play soldiers for another 20 seconds Freida picks up one of the soldiers and pulls an arm off it. Kevin hits her. Father tells Kevin to leave the room. Kevin obeys.

(4) A few minutes later Kevin goes into the dining room where the family is having dinner. The family eats in silence for 30 seconds.

(5) Kevin asks his mother, "Mom, will you mash my potatoes?" Mother does. Kevin thanks her and father says, "Kevin, I think you are old enough to mash your own potatoes now." The observer takes a break. Freida tells a joke that she had heard in school. Kevin laughs. Kevin tickles Freida. Freida says in a high-pitched, sing-songy voice, "Kevin is picking on me again." Kevin stops and eats.

(6) Kevin goes to father's chair and stands alongside it. Father puts his arm around Kevin's shoulders. Kevin says to mother as Freida looks at Kevin, "Can I go out and play after supper?" Mother does not reply. Kevin raises his voice and repeats the question. Mother says, "You don't have to yell; I can hear you." Father says, "How many times have I told you not to yell at your mother?" Kevin scratches a bruise on his arm while mother tells Freida to get started on the dishes, which Frieda does.

Kevin continues to rub and scratch his arm while mother and daughter are working at the kitchen sink.

(7) Mother tells Kevin to empty the rubbish. Kevin takes the garbage out of the house. He returns crying, while mother and Freida look at him. He says, "A bee stung me." As mother says, "Where did it bite you?", Freida and father watch Kevin. The father says, "What a big sissy, crying over a little bee sting." Kevin replies, "You're so mean, you don't know how much it hurts." Mother hands Kevin a tube of salve. The observer takes a break.

(8) Kevin is doing his homework on the living room floor while Freida, mother, and father engage in a conversation. This continues for about 20 seconds. Mother gets up from the knitting she has been doing, and comes over to Kevin and says, "You look so hot, dear, let me take your sweater off for you." Kevin allows her to take off the sweater. She then says, "I'll do these arithmetic problems for you," and takes the pencil out of his hand and works the problems for Kevin while he looks on.

(9) She continues for a few seconds to do his problems for him. Kevin says, "I think I know what the answer to the next problem is." Mother doesn't seem to hear. Kevin gets up from his homework and begins running around the room. Mother and father say loudly to him, "Why do you have to make so much noise?" Father then says, "If you don't stop this nonsense immediately, you're not going on the picnic Saturday." Kevin stops. Kevin returns to his homework, while mother and father are discussing plans for the picnic.

(10) Kevin says to Freida, "Hand me that book on the table, will you please." She does. Kevin thanks her. Freida says in an amused tone of voice,

"You're just sooo polite." Kevin continues to work while father reads the newspaper, mother knits, and Freida listens to the phonograph. Then the other three family members begin a discussion of the picnic and Kevin looks at them as they talk. He returns to work and Freida watches him.

Family Number 50

ID Number 50 11 0406 1

BEHAVIOR CODING SHEET

SUBJECT 1 OBSERVER JC DATE PHASE 1 NO 1

Behavior Codes

AP	Approval	DP	Dependency	NO	Normative	TA	Talk
AT	Attention	DS	Destructiveness	NR	No Response	TE	Tease
CM	Command	HR	High Rate	PN	Negative Physical Contact	WH	Whine
CN	Command (negative)	HU	Humiliate			YE	Yell
CO	Compliance	LA	Laugh	PP	Positive Physical Contact		
CR	Cry	NC	Non-compliance				
DI	Disapproval						

1	1NO-2NR	—1	—1	1NO-2NR Z3NR	—1	
2	1NO-4TA	1DI-4HU	2CM-1NC	2CN-1CO	1NO-4NO	
3	1NO-4NO	—1	—1	4DS-1PN	2CM-1CO	A
4	1NO-9NO	—1	—1	—1	—1	
5	1DP-3CO	1AP-2DI	u 4TA-1LA	1TE-4WH	1NO-9NR	
6	1NO-2PP	1TA-3NR Z4AT	1YE-2/3DI	1NO-9NR	1NO-9NR	
7	3CM-1CO K	1CR-3/4AT	1TA-3TA Z2AT	2HU-1DI	3NO-1NO	B
8	1NO-9NR	—1	—1	3TA-1AT	3TA-1AT	
9	1NO-3NO	1TA-3NR	1HR-2/3YEDI	2CN-1CO	1NO-9NR	
10	1CM-4CO	1AP-4TE	1NO-9NR	1AT-9NR	1NO-4AT	

Description Playing before Dinner (1-3), Dinnertime (4-5), Doing homework after dinner (8-10).

APPENDIX B

FP-FA Study
November 1971

OBSERVER ASSISTANT INVENTORY

Date _______________ Group (circle one): 9:30

11:00

1:00

Check the point on each of the following scales which reflects your feelings or understanding of this research project right now.

For me, this observation task is:

1	2	3	4	5	6	7
Easy						Very difficult

I __________ assisting with the data collection for this research project.

1	2	3	4	5	6	7
Like						Dislike

I feel the work of an observer is:

1	2	3	4	5	6	7
Challenging						Boring

The research assistant supervising this project expects __________ of us:

1	2	3	4	5	6	7
Too much						Too little

This research project is:

1	2	3	4	5	6	7
Interesting						Uninteresting

I feel that the data were are collecting in this research project is:

1	2	3	4	5	6	7
Important						Unimportant

The experimenter's prediction for the set of FA tapes seen by this group of observers was that the rate of deviant behavior in the family would:

+75%	+50%	+25%	0%	-25%	-50%	-75%
Increase			or have unknown effects			Decrease

If I were in charge of this research project I would run it:

1	2	3	4	5	6	7
The same way						Differently

My suggestions for changing or improving the procedures used in this or future research projects of this nature are:

____

____

____

____

____

____

____

____

APPENDIX C

OBSERVER ASSISTANT QUESTIONNAIRE - 1
December 1971

1. In several sentences give your understanding of what this study was about.

2. What were you told about the history of the Craig family?

3. As far as you can, indicate the variable being manipulated, the specific variables (behaviors) of interest to the investigator and the investigator's prediction about the variables being measured by <u>your</u> group's observations of the videotapes.

4. What evidence or arguments were presented by the investigator to support the prediction given your group?

5. Did you have any personal expectations regarding the outcome of the study? If so, what were they and were you more motivated to see the investigator's prediction or your own confirmed by the results of this study?

6. Why were the observers told about the study?

7. If you were the investigator, would you have conducted this study any differently? How would you have conducted it? Why?

OBSERVER ASSISTANT QUESTIONNAIRE - 2

1. At any time during this pilot study (where you were coding FP and FA videotapes), were you suspicious of the rationale given you by the investigator? That is, did you question whether you were being told the truth and that the investigator was studying what he described?

   a. If yes, at what point did you become suspicious and why?

   b. If yes, what do you think the investigator was really studying?

APPENDIX D

Memo to: Judges

From: Karl Skindrud

Re: Instructions for Sorting the Observer Assistant Questionnaires

I am asking you to take 30 minutes to read through the questionnaires completed by a group of 27 subjects. You will be sorting the questionnaires into three groups, as explained below.

The subjects were divided into three groups which were given different instructions regarding a study in which they were participating. The instructions given each of the groups can be summarized as follows: Control group: "You will be observing two sets of videotapes of family interaction--one set with the father present and another of the same family with the father absent. We hope to determine the effect of the father's absence upon family interaction."

The essential components that differentiate the instructions given the control group from at least one of the other groups include:

1) no specific information given about family status

2) no predictions were made

3) no specific behaviors were given special importance

4) no relevant studies were quoted or theories expounded

Increase group: "You will be observing two sets of videotapes of family interaction--one set with the father present and another of the same family with the father absent. The parents of the family were concerned about the behavior of their youngest boy, Craig, particularly his whining, yelling, and high rate behaviors. The tapes you will be observing were made at various stages of a treatment program. In the sets of tapes your group will be observing, we are predicting that <u>during the father's absence an increase in the rate of deviant behaviors will be observed</u>. By "rate of deviant behaviors" we mean the rate with which the CN, CR, DI, DS, DP, HU, HR, NC, PN, TE, WH, and YE codes are observed. We are quite certain that this prediction will be confirmed in this study because other studies have shown these trends as a result of the father's absence and we have theoretical reasons for predicting an increase. We are conducting this large-scale study to confirm these trends."

The essential components that differentiate the instructions given the <u>increase</u> group from at least one of the other groups include:

1) given information about the status of the family observed (about o enter a treatment program)
2) given the experimenter's prediction regarding an <u>increase in deviant behaviors</u> during the father's absence
3) the specific does predicted to change were listed
4) the experimenter's prediction was supported by early data returns from other studies and theory.

<u>Decrease group</u>: "You will be observing two sets of videotapes of family interaction--one set with the father present and another of the same family with the father absent. The parents of the family were concerned about the behavior of their youngest boy, Craig, particularly his whining, yelling, and high rate behaviors. The tapes you will be observing were made at various stages of a treatment program. In the sets of tapes your group will be observing, we are predicting that <u>during</u>

the father's absence a decrease in the rate of deviant behaviors will be observed. By "rate of deviant behaviors" we mean the rate with which the CN, CR, DI, DS, DP, HU, HR, NC, PN, TE, WH, and YE codes are observed. We are quite certain that this prediction will be confirmed in this study because other studies have shown these trends as a result of the father's absence under the same treatment conditions. We also have theoretical reasons for predicting a decrease. We are conducting this large-scale study to confirm these trends."

The essential components that differentiate the instructions given the decrease group from at least one of the other groups include:

1) given information about the status of the family observed (in treatment)
2) given the experimenter's prediction regarding a decrease in deviant behaviors during the father's absence
3) the specific codes predicted to change were listed
4) the experimenter's prediction was supported by early data returns from other studies and theory.

Instructions to the sorters: Your task is to read each one of the questionnaires and to sort them into three piles according to the information given you. If sorted correctly, you should finish with the following distribution:

Control Group	Increase Group	Decrease Group
5 questionnaires	11 questionnaires	11 questionnaires

If you do not finish with this distribution, re-sort borderline cases until you achieve the above distribution. (Please note that some subjects saw the father-absent tapes following and some before the father-present tapes. Consequently, the "increases" and "decreases" described should always be relative to the father-absent condition.) Please place each pile in one of the labeled envelopes provided as appropriate. Return to Karl Skindrud for recording of your sortings as soon as possible. Thank you for your help.

APPENDIX E

Memo to: Judges

From: Karl Skindrud

Re: Instructions for Sorting Questionnaire Responses to item 5

Please take 10 minutes to read each observer's response to item 5 of the Observer Assistant Inventory. I am interested in your judgement as to the observer's personal expectations of change in rate of deviant behavior from the FP to the FA condition. Did the observer expect an increase, no change, a decrease, or admit no personal expectations for the Craig family FA tapes?

Sort the attached 27 questionnaires into three piles according to item 5 responses as follows:

Increase	No Change or No Personal Expectations	Decrease

If you find responses which only indicate agreement or disagreement with the experimenter's prediction, place them in a separate pile with your judgement attached and I will assign the response to one of the above piles according to actual group membership.

Thank you for your assistance.