DOCUMENT RESUME

ED 071 749 PS 006 291

AUTHOR	Johnson, Stephen M.; Bolstad, Orin D.
TITLE	Methodological Issues in Naturalistic Observation: Some Problems and Solutions for Field Research. Final Report.
SPONS AGENCY	National Inst. of Mental Health (DhEW), Bethesda, Md.
PUB DATE	Mar 72
NOTE	87p.; Presented at the Banff International Conference on Behavior Modification (4th, March 1972)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Behavioral Science Research; Children; Classification; *Data Analysis; *Family Life; Observation; Problem Solving; *Psychology; *Research Methodology; Standards

ABSTRACT

An attempt at defining and describing those factors which most often jeopardize the validity of naturalistic behavioral data is presented. A number of investigations from many laboratories which demonstrate these methodological problems are reviewed. Next, suggestions, implementations, and testing of effectiveness of various solutions to these dilemmas of methodology are steps taken. Research in the paper involves the observation of both "normal" and "deviant" children and families in the home setting. The observation system employed is a modified form of the code devised by Patterson, Ray, Shaw, and Cobb (1969). The observations are made under certain restrictive conditions: (1) All family members must be present in two adjoining rooms; (2) No interactions with the observer are permitted; (3) The television set may not be on; and (4) No visitors or extended telephone calls are permitted. Other later studies are also reviewed in this paper. (CK)

METHODOLOGICAL ISSUES IN NATURALISTIC OBSERVATION: SOME PROBLEMS AND SOLUTIONS FOR FIELD RESEARCH

FINAL REPORT

by

Stephen M. Johnson and Orin D. Bolstad

University of Oregon

Presented at
The Fourth Banff International Conference on Behavior Modification
in March 1972.
To be published in the proceedings of the Banff Conference,
Research Press, 1972, in press.

METHODOLOGICAL ISSUES IN NATURALISTIC OBSERVATION: SOME PROBLEMS AND SOLUTIONS FOR FIELD RESEARCH[1]

Stephen M. Johnson and Orin D. Bolstad

University of Oregon

Encapsulated schools of thought have occurred in all sciences at some stage in their development. They appear most frequently during periods where the fundamental assumptions of the science are in question. Manifesto papers, acrimonious controversy, mutual rejection, and isolation of other schools' strategies are hallmarks of such episodes [David . Krantz, The separate worlds of operant and non-operant psychology. Journal of Applied Behavior Analysis, 1971, 4 (1), p. 61].

History may well reveal that the greatest contribution of behavior modification to the treatment of human problems came with its emphasis on the collection of behavioral data in natural settings. The growth of the field will surely continue to produce greater refinement and proliferation of specific behavior change procedures, but the critical standard for assessing their utility will very likely remain the same. We will always want to know how a given procedure affects the subject's relevant behavior in his "real" world.

If a behaviorist wants to convince someone of the correctness of his approach to treating human problems, he is generally much less likely to rely on logic, authority, or personal testimonials to persuade than are proponents of other schools of psychotherapeutic thought. Rather, it is most likely that he will show his behavioral data with the intimation that this data speaks eloquently for itself. Because he is aware of the research on the low level of generalizability of behavior across settings (e.g., see Mischel, 1968), he is likely to be more confident in this data as it becomes more naturalistic in character (i.e., as it reflects naturally occurring behavior in the subject's usual habitat). As a perusal of the behavior modification literature will indicate, these data are often

extremely persuasive. Yet, the apparent success of behavior modification and the enthusiasm that this success breeds may cause all of us to take an uncritical approach in evaluating the quality of that data on which the claims of success are based. A critical review of the naturalistic data in behavior modification research will reveal that most of it is gathered under circumstances in which a host of confounding influences can operate to yield invalid results. The observers employed are usually aware of the nature, purpose and expected results of the observation. The observed are also usually aware of being watched and often they also know the purpose and expected outcome of the observation. The procedures for gathering and computing data on observer agreement or accuracy are inappropriate or irrelevant to the purposes of the investigation. There is almost never an indication of the reliability of the dependent variable under study, and rarely is there any systematic data on the convergent validity of the dependent measure(s). Thus, by the standards employed in some other areas of psychological research, it can be charged that much behavior modification research data is subject to observer bias, observee reactivity, fakability, demand characteristics, response sets, and decay in instrumentation. In addition, the accuracy, reliability and validity of the data used is often unknown or inadequately established.

But, the purpose of this paper is not to catalogue our mistakes or to argue for the rejection of all but the purest data. If that were the case, we would probably have to conclude with that depressing note which makes so many treatises on methodology so discouraging. Although dressed in more technical language, this purist view often expresses itself as: "You can't get there from here." We can get there, but it's not quite as

simple as perhaps we were first led to believe. The first step in getting there is to define and describe those factors which most often jeopardize the validity of naturalistic behavioral data. To this end, we will review a host of investigations from many laboratories which demonstrate these methodological problems. The second step is more constructive in nature: to suggest, implement, and test the effectiveness of various solutions to these dilemmas of methodology. Because behavioral data has become the primary basis for our approach to diagnosing and treating human problems, the endeavor to improve methodology is perhaps our most critical task for strengthening our contribution to the science of human behavior.

We will argue that the same kinds of methodological considerations which are relevant in other areas of psychology are equally pertinent for behavioral research. At least with respect to the requirements of sound methodology, the time of isolation of behavioral psychology from other areas of the discipline should quickly come to an end.

Throughout this paper, we will rely heavily on the experience of our own research group in meeting, or at least attenuating, these problems. We take this approach to illustrate the problems and their possible solutions more precisely and concretely. Most of our solutions are far from perfect or final, but it is our hope that a report based on real experience and data may be more meaningful than hypothetical solutions which remain untested. Thus, before beginning or the outline of methodological problems and their respective solutions, it will be necessary for the reader to have a general understanding of the purposes and procedures of our research. This research involves the observation of both "normal" and "deviant" children and families in the home setting. The observation

system employed is a modified form of the code devised by Patterson, Ray, Shaw, and Cobb (1969). This revised system utilizes 35 distinct behavior categories to record all of the behaviors of the target child and all behaviors of other family members as they interact with this child. The system is designed for rapid sequential recording of the child's behavior, the responses of family members, the child's ensuing response, etc. Observations are typically done for forty-five minutes per evening during the pre-dinner hour for five consecutive week nights. The observations are made under certain restrictive conditions: a) All family members must be present in two adjoining rooms; b) No interactions with the observer are permitted; c) The television set may not be on; and, d) No visitors or extended telephone calls are permitted. Obviously, this represents a modified naturalistic situation.

On the average, these procedures yield the recording of between 1,800 and 1,900 responses and an approximately equal number of responses of other family agents over this time period of 3 hours and 45 minutes. This data is collected in connection with a number of interrelated projects. These include normative research investigations of the "normal" child (e.g., Johnson, Wahl, Martin & Johansson, 1972); research involving a behavioral analysis of the child and his family (e.g., Wahl, Johnson, Martin & Johansson, 1972; Karpowitz, 1972; Johansson, Johnson, Martin, & Wahl, 1971); outcome research on the effects of behavior modification intervention in families (Eyberg, 1972); comparisons of "normal" and "deviant" child populations (Lobitz & Johnson, 1972); and studies of methodological problems (Johnson & Lobitz, 1972; Adkins & Johnson, 1972; Martin, 1971). These latter studies will be reviewed in detail in the body of this paper. More recently, we have begun

to investigate the generality of children's behavior across school and home settings, and to document the level of generalization of the effects of behavior modification in one setting to behavior in other settings (Walker, Johnson, & Hops, 1972). Research is also in progress to relate naturalistic behavioral data to parental attitudes and behavioral data obtained in more artificial laboratory settings. With all of these objectives in mind, it is most critical that the behavioral data collected is as valid as possible and it is to this end that we explore the complex problems of methodology presented here.

Observer Agreement and Accuracy I: Problems of Calculation and Inference

The most widely recognized requirement of research involving behavioral observations is the establishment of the accuracy of the observers. This is typically done by some form of calculation of agreement between two or more observers in the field. Occasionally, observers are tested for accuracy by comparing their coding of video or audio tape with some previously established criterion coding of the recorded behavior. For convenience, we will refer to the former procedure as calculation of observer agreement and the latter as calculation of observer accuracy. In general, both of these procedures have been labeled observer reliability. We will eschew this terminology because it tends to confuse this simple requirement for observer agreement or accuracy with the concept of the reliability of a test as understood in traditional test theory. As we shall outline in section three, it is quite possible to have perfect observer agreement or accuracy on a given behavioral score with absolutely no reliability or consistency of measurement in the traditional sense. Generally, the classic reliability requirement involves

a demand for consistency in the measurement instrument over time (e.g., test-retest reliability) or over-sampled item sets responded to at roughly the same time (e.g., split-half reliability). An example may help clarify this point. If two computers score the same MMPI protocol identically, there is perfect "observer agreement" but this in no way means that the MMPI is a reliable test which yields consistent scores.[2] Although the question of reliability as traditionally understood has been largely ignored in behavioral research, we will argue in section three that it is a critical methodological requirement which should be clearly distinguished from observer agreement and accuracy.

There is no one established way to assess observer agreement or accuracy and that is as it should be, because the index must be tailored to suit the purposes of each individual investigation. There are three basic decisions which must be made in calculating observer agreement. The first decision involves the stipulation of the unit score on which the index of agreement should be assessed. In other words, what is the dependent variable for which an index of accuracy is required as measured by agreement with other observers or with a criterion? An example from our own research may help clarify this point. We obtain a "total deviant behavior score" for each of the children we observe. This score is based on the sum output of 15 behaviors judged to be deviant in nature. An outline of the rationale and validity of this score will be given in a later section. Suffice it to say, whenever two observers watch the same child for a given period, they each come up with their own deviant behavior score. These scores may then be compared for agreement on overall frequency. It is obvious that the same deviant behaviors need not be observed to get high

indexes of agreement on the total number of deviant behaviors observed. Yet, for many of our purposes, this is not important, since we merely want an index of the overall output of deviant behavior over a given period. The same procedure is, of course, applicable to one behavior only, chains of behavior, etc. The point is that the researcher must decide what unit is of interest to him for his purposes and then compare agreement data on that variable. In complex coding systems, like the one used in our laboratory, it has been customary to get an overall percent agreement figure which reflects the average level of agreement within small time blocks (e.g., 6-10 seconds) over all codes. In general, we would argue that this kind of observer agreement data is relatively meaningless. It has limited meaning because it is based on a combination of codes, some of which are observed with high consensus and some which are not. Furthermore, the figure tends to overweight those high rate behaviors which are usually observed with greater accuracy and underweight those low frequency behaviors which are usually observed with less accuracy. Patterson (personal communication) has reported that the observer agreement on a code correlates .49 with its frequency of use. Since it is often the low base rate behaviors which are of most interest to researchers, this overall index of observer agreement probably overestimates the actual agreement on those variables of most concern.

The second question to be faced involves the time span within which common coding is to be counted as an agreement. For most purposes of our current research, score agreement over the entire 225 minutes of observation is adequate. Thus, when we compute the total deviant behavior score over this period, we do not know that each observer sees the same deviant

behavior at the same time. But, good agreement on the overall score tells us that we have a consensually validated estimate of the child's overall deviancy. For some research purposes, this broad time span for agreement would be totally inadequate. For conditional probability analysis of one behavior (cf. Patterson & Cobb, 1971), for example, one needs to know that two observers saw the same behavior at the same time and (depending on the question) that each observer also saw the same set or chain of antecedents and/or consequences. This latter criterion is extremely stringent, particularly with complex codes where low rate behaviors are involved, but these criteria are necessary for an appropriate accuracy estimate.

Once one has decided on the score to be analyzed and the temporal rules for obtaining this score, one must then face the problem of what to do with these scores to give a numerical index of agreement. The two most common methods of analysis are percent agreement and some form of correlational analysis over the two sets of values. Both methods may, of course, be used for observer agreement calculation within one subject or across a group of subjects. Once again, neither method is always appropriate for every problem and each has its advantages and disadvantages. The most common way of calculating observer agreement involves the following simple formula:

$$\frac{\text{number of agreements}}{\text{number of agreements + disagreements}}$$

What is defined as an agreement or disagreement has already been solved if one has decided on the "score" to be calibrated and the time span involved.

Use of this formula implies, however, that one must be able to discriminate the occurrence of both agreements and disagreements. This can

only be accomplished precisely when the time span covered is relatively small (e.g., 1-15 seconds) so that one can be reasonably sure that two observers agreed or disagreed on the same coding unit. It has been common practice for investigators to compare recorded occurrences of behavior units over much longer time periods and obtain a percent agreement figure between two observers which reflects the following:

$$\frac{\text{smaller number of observed occurrences}}{\text{larger number of observed occurrences}}$$

The present authors would view this as an inappropriate procedure because there is no necessary "agreement" implied by the resulting percent. If one observer sees 10 occurrences of a behavior over a 30-minute period and the other sees 12, there is no assurance that they were ever in agreement. The behavior could have occurred 22 or more times and there could be absolutely no agreement on specific events. The two observers did not necessarily agree 84% of the time. Data of this kind can be more appropriately analyzed by correlational methods if such analysis is consistent with the way in which the data is employed for the question under study. Although the same basic problem mentioned above can, of course, occur, the correlational method is viewed as more appropriate because; a) The correlation is computed over an array of subjects or observation time segments and b) The correlation reflects the level of agreement on the total obtained score and it does not imply any agreement on specific events.

Whenever using the appropriate method of calculating observer agreement percent, (i.e: $\frac{\text{number of agreements}}{\text{number of agreements + disagreements}}$) the investigator should be particularly cognizant of the base rate problem. That is, the obtained percent agreement figure should be compared with the amount of agreement that could be obtained by chance. An example will clarify this point. Suppose two coders are coding on a binary behavior coding system (e.g., appropriate vs. inappropriate behavior). For the sake of illustration, let us suppose that observers have to characterize the subject's behavior as either appropriate or inappropriate every five seconds. Now, let us suppose, as is usually the case, that most of the subject's behavior is appropriate. If the subject's behavior were appropriate 90% of the time,

two observers coding randomly at these base rates (i.e., .90-.10) will obtain 82% agreement by chance alone. Chance agreement is computed by squaring the base rate of each code category and summing these values.[3] In this simple case, the mathematics would be as follows: $.90^2 + .10^2 = .82$. The same procedure may, of course, be used with multi-code systems.

The above .90-.10 split problem may be reconceptualized as one in which the occurrence or nonoccurrence of inappropriate behavior is coded every five seconds. If, for purposes of computing observer agreement, we look at only those blocks in which at least one of two observers coded the occurrence of inappropriate behavior, the chance level agreement is drastically reduced. The probability that two observers would code occurrence in the same block by chance is only $.10^2$ or one percent. It would not be theoretically inappropriate to count agreement on nonoccurrence but, in the present example and in most cases, this procedure is associated with relatively high levels of chance agreement.

Whenever percent agreement data is reported, the base rate chance agreement should also be reported and the difference noted. Statistical tests of that difference can, of course, be computed. As long as the base rate data is reported, the percent agreement figure would always seem to be appropriate. For obvious reasons, however, it becomes less satisfactory as the chance agreement figure approaches 1.0.

The other common method of computing agreement data is by means of a correlation between two sets of observations. The values may be scores from a group of subjects or scores from n observation segments on one subject. This method is particularly useful when one is faced with the high chance agreement problem or where the requirement of simple similarity in ordering subjects on the dependent variable is sufficient for the research. As we shall illustrate, the

correlation is also particularly useful in cases where one has a limited sample of observer agreement data relative to the total amount of observation data. In general, correlations have been used with data scores based on relatively large time samples. In other words, they

tend to be used for summary scores on individuals over periods of 10 minutes to 24 hours. There is no reason why correlation methodology could not be applied to data from smaller time segments (e.g., 5 seconds), but this has rarely been done. So, studies using correlation methods have generally been those in which one cannot be sure that the same behaviors are being jointly observed at the same time. In using correlation methods for estimating agreement, one should be aware of two phenomena. First, it is possible to obtain high coefficients of correlation when one observer consistently overestimates behavioral rates relative to the other observer. The difference can be rather large, but if it is consistently in one direction, the correlation can be quite high. For some purposes this problem would be of little consequence but for other purposes it could be of considerable importance. The data can be examined visually, or in other more systematic ways, to see to what extent this is the case. This problem can be virtually eliminated if one uses many observers and arranges for all of them to calibrate each other for agreement data. Under these circumstances, one will obtain a collection of regular observer figures and a list of mixed calibrator figures for correlation. This procedure should generally correct for systematic individual differences and make a consistent pattern as outlined above extremely unlikely. The second problem to be cognizant of in using correlations is that higher values become more possible as the range on the dependent variable becomes greater. This fact may lead to high indexes of agreement when observers are really quite discrepant with respect to the number of a given behavior they are observing. An illustration may clarify this point. Let us suppose we are observing rates of crying and whining behavior in preschool children over a five-hour period. Some

particularly "good" children may display these behaviors very little and, given a true occurrence score of 7, two observers may obtain scores of 5 and 10 on this behavior class. This would be only 50% agreement. Other children display these behaviors with moderate to very high frequency. For a child with high frequency, we may find our two observers giving us scores of 75 and 125 respectively. This would be equivalent to 60% agreement and, of course, represents a raw discrepancy of 50 occurrences. Yet, if these examples were repeated throughout the distribution of scores and if there were little overlap, a high correlation would be obtained. This would be even more true, of course, if one observer consistently overestimated the rates observed by the other. Yet, even this possibility does not necessarily jeopardize the utility of the method. It must merely be recognized, examined and its implication for the question under study evaluated. In our own research we want to catalogue the deviancy rates of normal children, compare them with deviant children, and observe changes in deviancy rates as a result of behavior modification training with parents. For these purposes, general agreement on levels of deviant responding is quite good enough.

In our research on the normal child, we have had 47 families of the total 77 families observed for the regular five-day period by an assigned observer. On one of these days an additional observer was sent to the family for the purpose of checking observer agreement. The correlation between the deviant behavior scores of the two observers was .80. But, in a purely statistical sense, this figure is an underestimate of what the agreement correlation would be for the full five days of observation. Since we are using a statistic based on five times as much data, we want to know the expected

observer agreement correlation for this extended period. Adding time to an observation period is analogous to adding items to a test. The problem we are faced with here is very similar to that dealt with by traditional test theorists who have sought, for example, to estimate the reliability of an entire test based on the reliability of some portion of the test. In our case, we want to know the expected correlation for the statistic based on five days when we have the correlation based on one day. The well-known Spearman-Brown formula (Guilford, 1954) may be applied to this end (as in Patterson, Cobb, & Ray, 1972; Patterson & Reid, 1970; Reid, 1967).[4]

$$r_{nn} = \frac{nr_{tt}}{1 + (n-1)r_{tt}}$$

where r_{tt} = reliability of the test of unit length

n = length of total test.

With the Spearman-Brown correction, the expected observer agreement correlation for the deviant behavior score is .95. This same procedure has also been applied to other statistics of particular interest in this research including: a) the proportion of the parent's generally "negative" responses (correct agreement = .97), b) the proportion of the parent's generally positive responses (corrected agreement = .98), c) the median agreement coefficient of the 29 behavior codes observed for five or more children (corrected agreement = .91), d) the median corrected agreement of the 11 out of 15 deviant behavior codes used (r = .91), e) the number of parental commands given (corrected agreement = .99), and f) the compliance ratio (i.e., compliances/compliances plus noncompliances) of the child (corrected agreement = .92). As our research is completed, we will be presenting observer agreement data using different statistics, computed in different ways, and evaluated by different criteria.

The primary point of this section is to indicate that there are many ways of calculating observer agreement data and there is no one "right way to do it." The methods differ on three basic dimensions: a) the nature and breadth of the dependent variable unit, b) the time span covered, and c) the method of computing the index. Each investigator must make his own decisions on each of these three points in line with the purposes of his investigation. But, the investigator should be guided by one central prescription--the agreement data should be computed on the score used as the dependent variable. It makes no sense to report overall average agreement data (except perhaps as a bow to tradition) when the dependent variable is "deviant behavior rate." In addition, it makes little sense to make the agreement criteria relative to time span more stringent than necessary. If the dependent variable is overall rate of deviant behavior for a five-day period, then this is the statistic for which agreement should be computed. It is not necessary for this limited purpose that both observers see the same deviant behavior in the same brief time block.

Before closing this section on the computation of observer agreement, we should address the somewhat unanswerable question of the minimum criteria for the acceptability of observer agreement data. In other words, how much agreement is sufficient for moving on to consider the results of a particular study. When using observer agreement percent, it would seem reasonable, at the very minimum, to show that the agreement percent is greater than that which could be expected by chance alone. When dealing with correlation data, one should at least show the obtained correlation to be statistically significant. These criteria are, of course, extremely minimal and certainly far below those criteria commonly used in traditional testing and measurement

to establish reliability (e.g., see Guilford, 1954). Yet, these criteria do provide a reasonable lowest level standard and there are some very good reasons why we should not be overly conservative on this point. In the first place, very complex codes, which may provide us with some of our most interesting findings, are very difficult to use with complete accuracy. On the basis of our experience, and that of G. R. Patterson (personal communication), we see an overall agreement percent of 80% to 85% as traditionally computed as a realistic upper limit for the kind of complex code we are using.

Furthermore, to the exten. that less than perfect agreement represents only unsystematic error in the dependent variable, it cannot be considered a confounding variable accounting for positive results. Any positive finding which emerges in spite of a good deal of "noise" or error variance is probably a relatively strong effect.

Low observer agreement does, however, have very important implications for negative results. This gets us back to the fundamental principle that one can never prove the null hypothesis. The more error in the measurement instrument, the greater the chance for failing to discover important phenomena. Thus, just as with traditional test reliability, the lower the observer accuracy, the less confidence one can have in any negative findings from the research.

Observer Agreement and Accuracy II: Generalizability of Observer Agreement Data

All of the preceding discussion on the calculation of observer agreement data relies on the assumption that the obtained estimates of agreement are generalizable to the remainder of the observers' data collection.

In most naturalistic behavioral research, however, this assumption cannot go unchallenged and this brings us to our next, and largely [illegible] methodological problem. To illustrate this problem, let us take the not untypical case of an investigator who trains his observers on a behavioral code until they meet the criterion of two consecutive observation sessions at 80% agreement or better. After completing this training, the investigator embarks on his research with no further assessment of observer agreement. There are three basic problems with this methodology which make the generalizability of this agreement data extremely questionable. These problems are a) the nonrandomness of the selected data points, b) the unrepresentativeness of the selected data points in terms of the time of the assessment, and c) the potential for the observer's reactivity to being checked or watched. The first two problems may be rather easily solved in all naturalistic research, but the third problem represents quite a challenge to some forms of naturalistic observation. Let us explore these problems in more detail. The nonrandomness of selecting the last two "successful" observation sessions in a series for establishing a true estimate of agreement should be very obvious. It is not unlikely that, had the investigator obtained several additional agreement sessions, he would find the average agreement figure to be lower than 80%. It is quite possible that our observers had, by chance, two consecutive "good days" which are highly unrepresentative of the days to come. One can almost visualize our hypothetical investigator, after the first day of highly accurate observation, saying to his observers, "That was really a good one; all we need is one more good session and we can begin the study." But, now we are getting into problems two and three.

The second problem of unrepresentativeness in terms of time has previously been discussed by Campbell and Stanley (1966) and labeled instrument decay. That is, estimates of observer accuracy obtained one week may not be representative of observer accuracy the next week. The longer the research lasts, the greater is the potential problem of instrument decay. In the case of human observers, the decay may result from processes of forgetting, new learning, fatigue, etc. Thus, because of instrument decay, our investigator's estimate of 80% agreement is probably an exaggeration of the true agreement during the study itself. The problem of instrument decay is also often compounded by the fact that during observer training, there is usually a great deal of intense and concentrated work with the code, coupled with extensive training and feedback concerning observer accuracy. This intensity of experience and feedback is usually not maintained throughout the course of the research, and, as a result, the two time periods are characterized by very different sets of experiences for the observers. The third problem of generalizability of this agreement data involves the simple fact that people often do a better, or at least a different, job when they are aware of being watched as opposed to when they are not. Campbell and Stanley (1966) have labeled this problem reactive effects of testing. It is likely that, when observers are being "tested" for accuracy, they will have heightened motivation for accuracy and heightened vigilance for critical behaviors or for the coding peculiarities of their calibrator. This point has been brought home dramatically to us on more than one occasion by the tears of an observer after earning a particularly low agreement rating. Thus, because of the reactivity problem, estimates of observer agreement obtained with the awareness of the observer are likely to over-

estimate the true agreement level which would be obtained if the observer were not aware of such calibration.

Fortunately, all of the preceding logical arguments have been investigated in some recent research largely contributed by John Reid of the Oregon Research Institute. In his first published study on this problem, Reid (1970) designed a study which, from the observer's point of view, was almost identical to the hypothetical example given earlier. In this study, observers were trained for a maximum of eight days on complex behavioral coding of videotape material in a laboratory setting. They were trained until they obtained 70% agreement on two consecutive days for 15 minutes of observation per day. In each training session, feedback was given to the observers concerning their accuracy, but in the immediately following assessment session no such feedback was given. After training was completed, the observers were led to believe that they would be the only ones seeing the new tapes because criterion protocols had not been previously constructed and the tapes would have to be erased immediately after coding so that they might be reused. Observers were urged, therefore, to be very careful in their subsequent coding. On each of four consecutive days after reaching criterion, the observers were asked to complete 30 minutes of observation on similar stimulus materials. The first 5 minutes of each session served as a practice period in which the experimenter gave feedback to the observer when necessary. Although the observers were not aware of it, criterion protocols had been prepared on these videotape materials, thus providing a covert assessment of observer accuracy. Throughout the study, observer accuracy was computed by comparing the data entry with the standard protocol. Both errors of commission and omission were counted. This is the type of overall and generally stringent accuracy criteria which we

have earlier recommended against for most research purposes. In any event, all observers reached the 70% agreement criterion at or before eight days of training and the average agreement during the two criterion days was 75.50%. The average agreement during the following four days, however, was only 51.25%, representing a drop of nearly 25 percentage points. Now it is obvious that, in this first study, all three confounding problems were present--a) unrepr entativeness of the criterion days, b) instrument decay due to passage of time, etc., and c) reactivity to being tested. This is the procedure, however, which is representative of many investigations in the field. Reid's study certainly demonstrates the potentially powerful impact of these confounding influences. Because of the relatively short time span involved in this study, it is likely that most of this effect is attributable to the reactivity problem.

A more recent study by Taplin and Reid (in preparation) sheds greater light on the contribution of each of these confounding influences. Although some of the details of this latter study differed from the earlier investigation, the basic procedures and concepts involved were the same. After training to criterion of 70% agreement, the observers were assigned to one of three groups: a) No-check group--the procedures in this group essentially replicated those of the former study such that the observers were unaware of agreement assessment beyond the criterion period. b) Spot-check group--in this group, observers were told that their accuracy would be checked from time to time during the experimental phase, but that they would be aware of when these checks would be made. This group controls for the problem of unrepresentativeness of criterion day performance as well as instrument decay, but does not control for reactivity to testing. c) Random-check group--in this group, the observers were told that their

accuracy would be checked randomly in 20% of the sessions but that they would not be aware of when these checks occurred. This procedure not only controls for the first two problems, but also controls for the reactivity problem by making it constant. In other words, it is likely that the knowledge of a "potential" assessment of accuracy will increase the observer's motivation and vigilance. Yet, this is not a confounding influence since any random assessment of accuracy in the series will not be unrepresentative of the other points in the series. Thus, whenever practically possible, this strategy of random assessment without observer awareness or assessment on every observation occasion is most consistent with rigorous research requirements. The results of the Taplin and Reid (in preparation) study are presented in Figure 1 and are of particular interest here for the way in which they illustrate the operation of the three confounding

---
Insert Figure 1 about here
---

problems. The discrepancy between percent agreement on the criterion days and on the days of data collection is obvious and significant for all groups. Thus, the use of this kind of previously collected criterion data for generalization to subsequent data collection appears highly inappropriate, regardless of the method of data collection or the rationale provided observers during the data collection period. In all groups, the confounds of unrepresentativeness of criterion day data and instrument decay from the criterion days are obvious. Now, the observer accuracy data produced on the spot check days are not unrepresentative, in the sense that these days were presumably not chosen initially as particularly "good" ones, as were the criterion days. In addition, this procedure controls for the

problems of instrument decay due to passage of time since the check were randomly throughout the period of data collection. It is clear that the "spot check" accuracy assessment produced less overestimation of true accuracy than did the criterion day estimates. The data in this "spot check" group is very clear, however, in illustrating the problem of observer reactivity. The agreement percents in the two spot check periods are clearly overestimates of agreement percents obtained under covert assessment. Thus, the "spot check" method represents a decided improvement over the criterion score method, but it is an imperfect solution.

In some naturalistic settings, the "spot check" method may have to remain the optimal solution for reasons of cost and the potential for the reactive effects inherent in using multiple observers continuously (e.g., in observations in homes or other confined quarters). As previously stated, the methods of total accuracy assessment (e.g., as used by Wahler, 1967, and Browning & Stover, 1971) or random accuracy assessment without awareness (as in Taplin & Reid, in preparation) are always preferable when possible. These methods are, of course, particularly simple to apply with video or audio tape materials or in natural settings where two or more observers are, for whatever reason, employed simultaneously and continuously. In classrooms, for example, it is often the case that two or more observers record the behaviors of two or more children. Under these circumstances, the investigator can arrange the observers' recording schedules so that their observation of subjects overlap at random times. In this way, two observers can record the behavior of the same subject at the same time without either having knowledge of the ongoing calibration for agreement which is occurring at that specific time. This procedure would replicate the "random check

group" of Taplin and Reid (in preparation) in a field setting. It would probably be difficult, if not impossible, to keep the fact of random calibration a secret from the observers for any extended period, but, as stated earlier, this is no real problem, because the randomly collected data without specific awareness is representative of accuracy at other times. The Taplin and Reid (in preparation) data would suggest that the motivational effects of informing observers of the random checks slightly increases the level and stability of their accuracy scores. (Compare the three groups' accuracy level and stability in the data collection period in Figure 1.)

In more recent research, Reid and his colleagues have directed their efforts to finding ways of eliminating the instrument decay or "observer drift" observed in all previous studies regardless of the method of monitoring. In several long-term research projects, including our own (e.g., Johnson, Wahl, Martin & Johansson, 1972), the one directed by G. R. Patterson (e.g., Patterson, Cobb, & Ray, 1972) and the one reported by Browning and Stover (1971), continuous training, discussion of the coding system, and accuracy feedback are provided for the observers. It is possible that this kind of training and feedback could eliminate, or at least attenuate, observers' accuracy drift as well as the problem of the unrepresentativeness of "spot check" accuracy assessments. To test this hypothesis, DeMaster and Reid (in preparation) designed a study in which three levels of feedback and training during data collection were compared on a sample of 28 observers. The observers were divided into 14 pairs and all subsequent procedures were carried out in the context of these fixed pairs. The three experimental groups were as follows: Group I--Total Feedback--In this group observers a) discussed their observation performance together while reviewing their coding of the previous day's video tape, b) discussed their previous

day's observation with the experimenter in terms of their agreement with the criterion coded protocol, and c) received a daily report of their accuracy with respect to the criterion protocol. Group II--Pair Agreement Feedback--In this group, observers were given the opportunity to discuss their performance as in a above and b)were given a daily report on the extent to which each observer's coding protocol agreed with the protocol of the other observer. Subjects in this group were deprived of a discussion or report of their level of agreement with the criterion protocols. Group III--No Feedback--Subjects in this group were deprived of the kinds of feedback given in the previous two conditions and were instructed not to discuss their work among themselves to eliminate a possible "bias of the data." This group was similar in concept to the random-check group in the Taplin and Reid (in preparation) study in that they were told, as were all other subjects, that their accuracy would be checked at random intervals in the data collection period. The dependent variables were a) the agreement scores between pairs of observers and b) the "accuracy" scores reflected by the percent agreement with the criterion protocols. The results showed that the intra-pair observer agreement scores were significantly higher than were scores reflecting agreement with the criterion. These results tend to corroborate the hypothesis forwarded by Baer, Wolf, and Risley (1968) and Bijou, Peterson and Ault (1968) that high intra-pair agreement does not necessarily reflect proper use of the coding system. We shall call this problem "consensual observer drift." It is very important to note, however, that the design of this study which placed observers in fixed and unchanging pairs would tend to maximize this effect. In the field studies referred to above, observers typically meet in larger groups for training and feedback and observers rotate in calibrating each other's observations. Under these

circumstances, the effects of consensual drift would logically be expected to be less potent. Indeed, further data from the DeMaster and Reid (in preparation) study lends support to this argument. On those video-tape materials where more than one pair of observers had coded the sequence the investigators compared the fixed pair agreement with the agreement between observers in other pairs. In all cases, the fixed pair agreed more with one another than they did with the observers in the other pairs. Thus, this idiosyncratic drift of fixed pairs may be greater than drift experienced under currently employed field research procedures. Yet, a recent study by Romanczyk, Kent, Diament, and O'Leary (1971) showed that during overt agreement assessment observers would change their coding behavior to more closely approximate the differential coding styles of their calibrators. Thus, it is possible for observers to produce one kind of consensual drift with some calibrators and an opposite consensual drift with others to yield artifically high observer agreement data.

The manipulations in the Romanczyk et al. (1971) study were quite powerful, however, and one can question the generalizability of these artificially induced conditions to real field studies. Nevertheless, this study does demonstrate the potential for powerful and differential consensual drift. In spite of these considerations, one must realize that it is impossible in an ongoing field observation to have a "pure" criterion protocol, since one cannot arbitrarily designate one observer's protocol as the "true" criterion and the other as the imperfect approximate. But, one can attenuate this problem considerably by having frequent training sessions with observers on pre-coded video-tape material or on pre-coded behavioral scripts which may be acted out live by paid subjects. The importance of

this recommendation is underlined by DeMaster and Reid's (in preparation) second important finding. Analysis of the data indicated a significant main effect for feedback conditions, with the total feedback group doing best, followed by the intra-pair feedback group and the no feedback group, respectively.

It may be of interest to review briefly how our own project stacks up with regard to these considerations and to suggest ways in which it and similar projects might be improved in this area. Initial observer training in our laboratory consists of the following program: a) reading and study of the observation manual, b) completion of programmed instruction materials involving precoded interactions, c) participation in daily intensive training sessions which include discussion of the system and coding of precoded scripts which are acted out live by paid but nonprofessional actors, d) field training with a more experienced observer followed immediately by agreement checks. Currently, when an observer obtains five sessions with an average overall percent agreement of 70% or better, she may begin regular observation without constant monitoring. All observers continue to participate in continuous training and are subject to continuous checking with feedback. This is accomplished in two ways. First, each observer is subject to one spot-check calibration for each family she observes. This calibration may come on any one of the regular five days of observation. Both observers figure their percent agreement in the traditional way immediately after the session and discuss their disagreements at this time. If they cannot resolve their disagreement on a particular or idiosyncratic problem, they call the observer trainer immediately who serves as sort of an imperfect criterion coder. From time to time, idiosyncratic problems arise which cannot be resolved by the coding manual alone. Decisions on how to

code these special cases are made by the group and the trainer and are entered in a "decision log" which is periodically studied by all observers. These special circumstances are unfortunate and provide an opportunity for consensual drift, but are part of the reality with which we must deal. The "decision log" helps attenuate the drift problem on these decisions, and most of them tend to be idiosyncratic to one or two families. The second aspect of continual training involves a minimum of one 90-minute training session per week for all observers involving discussion and live coding experience. We have been negligent in our procedures in not retaining our precoded scripts over time and recoding these from month to month and year to year. On the basis of our review of Reid's excellent work, we have now begun to correct this error by retaining these scripts and subjecting them to recoding periodically to check the problem of "consensual observer drift." As will be obvious, we use the imperfect method of "spot check" calibration for observer agreement, but Reid's data is encouraging in that it indicates that the kind of intensive and continual training outlined here may attenuate the problems associated with this method. Furthermore, our observers are convinced that calibration scores obtained on a single day of observation are probably lower than would be obtained over two or more days of observation. The reason for this belief is that the calibrator would logically have more difficulty in adapting to each new home environment and identifying the subjects of observation on the first day in the home than on subsequent days. Unfortunately, we have no hard data to prove this hypothesis, but we have begun to do more than one day of calibration on families in order to test it.

The problem of consensual drift is also attenuated in this project by

the practice of having each observer calibrate all other observers. We recently began to employ only one calibrator for reasons of convenience and cost, but this review has persuaded us to return, at least partially, to multiple calibration among all observers.

As stated earlier, the problems associated with reactivity to testing for observer agreement could largely be solved by procedures which involved coding of audio or video tapes. This is true because one could arrange calibration on a random basis without observer awareness. Because procedures of this kind could also solve or attenuate problems of observer bias and subject reactivity, we are beginning to consider procedures of this type more seriously for future research and are now involved in pilot work on the feasibility of these methods. Short of this, we must be content with the "spot check" method as outlined and attempt to attenuate the problems associated with this method by use of extensive training and feedback as suggested by DeMaster and Reid (in preparation).

Reliability of Naturalistic Behavioral Data

One must look long and hard through the behavior modification literature to find even an example of reliability data on naturalistic behavior rate scores. In classical test theory, the concept of reliability involves the consistency with which a test measures a given attribute or yields a consistent score on a given dimension. Theoretically, a test of intelligence, for example, is reliable if it consistently yields highly similar scores for the same individual relative to other individuals in the sample. There are several approaches to measuring reliability including split-half measures, equivalent forms, test-retest methods, etc. Each method has a somewhat different meaning, but the basic objective of each is an estimate

of the consistency of measurement. It is difficult to tell whether behaviorists have simply neglected, or deliberately rejected, the reliability requirement for their own research. The concept comes out of classical test theory and is obviously allied to trait concepts of personality. Behaviorists may feel that the concept is irrelevant to their purposes. After all, we know that there is often very little proven consistency in human behavior over time and stimulus situations (e.g., see Mischel, 1968), so why should we require a consistency in our measurement instruments that is not present in real life? Behaviorists may feel that reliability is an outmoded concept and belongs exclusively to the era of trait psychology. If this is, in fact, the reason for the neglect of the reliability issue in behavioral research, it represents a serious conceptual error and a clear misapplication of the meaning of the data on the lack of behavioral consistency so eloquently summarized by Mischel (1968). It is true, of course, that behaviorists employ more restricted definitions of the topography of the relevant response dimensions (e.g., hitting vs. aggression) and that they often include more restrictive stimulus events in defining these dimensions (e.g., child noncompliance to mother's commands vs. child negativism). Yet, the fact remains that we are still dealing with <u>scores</u> that reflect behavioral dimensions. If the word "trait" offends, then another label will do as well. Furthermore, the scores are obtained for the same purposes that trait scores are obtained--to correlate with some other variable. Generally, behavior modifiers "correlate" these scores with the presence or absence of some treatment procedure but certainly our data is not limited to this one objective. In our own research, for example, we are currently comparing children's deviant behavior rates in their homes with their deviancy in the school classroom (Walker, Johnson, & Hops, 1972) and comparing the deviancy

rates of normal children with those observed in referred or "deviant" children (Lobitz & Johnson, 1972). The most elementary knowledge of the concept of reliability tells us that some minimal level of behavior score reliability is necessary before we can ever hope to obtain any significant relationship between our behavioral score and any external variable. Thus, the requirement of score reliability is just as important in research employing behavioral assessment as it is in more traditional forms of psychological assessment, but with only a few exceptions (e.g., Cobb, 1969; Harris, 1969; Olson, 1930-31; Patterson, Cobb, & Ray, 1972) behaviorists have ignored this important issue.

As a consequence of the reasoning presented above, we have been particularly cognizant of the reliability of the scores used in our research. We were quite encouraged to find, for example, that the odd-even-split-half reliability of our "total deviant behavior score" in a sample of 33 "normal" children was .72. This reliability was computed by correlating the total deviant behavior score obtained on the first, third, and first half of the fifth day with the same score obtained from the remainder of the period. After applying the Spearman-Brown correction formula, we found that the reliability of this score for the entire five-day observation period was .83. This relatively high level of reliability indicates that this score should, at least in a statistical sense, be quite sensitive to manipulation or to true relationships with other external variables (e.g., social class, or educational level of the parents). Other behavioral scores which are important to our research include: a) the proportion of generally negative responses of the parents (corrected reliability = .90), b) the proportion of generally positive responses of parents (corrected reliability = .87),

c) the median reliability of the 35 individual codes (r = .69), d) the corrected median reliability of the deviant codes = .66, 3) the number of parental commands during the observation (corrected reliability = .85), and f) the compliance ratio (i.e., compliances/compliances + noncompliance) of the child (corrected reliability = .49). The reliability of the compliance ratio is not as high as we might have wished, but it may still be high enough to be sensitive enough for powerful manipulations. We have been less fortunate in obtaining good reliability scores on some other statistics important to our research efforts. For example, the compliance ratios to specific agents (i.e., to mothers or fathers) have yielded rather low reliabilities. The reasons for this are two-fold: First, ratio scores are always less reliable than are their componant raw scores, because they combine the error variance of both components. Second, and of more general importance, these scores are based on relatively few occurrences. On the average, for example, fathers give only 36 commands over the five-day period. These occurrences must then be divided for the compliances and noncompliances and further split in half for the odd-even reliability estimate. By the time this erosion takes place, there are few data points on which to base reliability estimates. This problem is even more profound when we use one day of compliance ratio data to compute observer agreement on this statistic, since, on the average, fathers give only 7.2 commands per day. Thus, when we are dealing with behavioral events of fairly low base rate, observer agreement correlations and reliability coefficients may often not be "fairly" computed because there is simply not enough data. In classical test theory terminology, there may often not be enough "items" on the behavioral test to permit an accurate estimation of the reliability of the

score. What should we do with cases of this kind? A methodological purist might argue that we should throw out this data and use only scores with proven high reliability and observer agreement. We would argue that this course would be a particularly unfortunate solution for several reasons. First, low base rate behaviors are often those of special importance in clinical work. Second, if low reliability reflects nothing more than random, unsystematic error in the measurement instrument, it cannot jeopardize or provide a confounding influence on positive results (i.e., it cannot contribute to the commission of Type I errors). But, either low reliability or low observer agreement does have profound implications for the meaning of negative results (i.e., the commission of Type II errors). Fortunately, the effects of many behavior modification procedures are so dramatic that they will emerge significant in spite of relatively low reliability or observer accuracy.

In one of the other few examples of reliability data in the behavior modification literature, Cobb (1969) found that the average odd-even reliability of relevant behavioral codes used in the school setting was only .72. Yet, Cobb (1969) found that the rates of certain coded behaviors showed strong relationships to achievement in arithmetic. Thus, relatively low reliability or observer agreement jeopardizes very little the meaning of positive results, but leaves negative results with little meaning. There is, however, one very critical qualifying point to this argument. It is that the error expressed in low reliability or observer accuracy must be random, unsystematic, and unbiased. With this consideration in mind, we now move to what are perhaps the most important methodological issues in naturalistic research--observer bias and observee reactivity to the observation process.

The Problem of Observer Bias in Naturalistic Observation

Shortly after the turn of the century, O. Pfungst became intrigued with a mysteriously clever horse named Hans. By tapping his foot, "Clever Hans" was able to add, subtract, multiply and divide and to spell, read, and solve problems of musical harmony (Pfungst, 1911). Hans' owner, a Mr. von Osten, was a German mathematics teacher who, unlike the vaudeville trainers of show animals, did not profit from the horse's peculiar talents. He insisted that he did not cue the animal and, as proof, he permitted others to question Hans without his being present. Pfungst remained incredulous and began a program of systematic study to unravel the mystery of Hans' talents.

Pfungst soon discovered that, if the horse could not see the questioner, Hans could not even answer the simplest of questions. Neither would Hans respond if the questioner himself did not know the answer. Pfungst next observed that a forward inclination of the questioner's head was sufficient to start the horse tapping, and raising the head was sufficient to terminate the tapping. This was true even for very slight motions of the head, as well as the lowering and raising of the eyebrows and the dilation and contraction of the questioner's nostrils.

Pfungst reasoned and demonstrated that Hans' questioners, even the skeptical ones, expected the horse to give correct responses. Unwittingly, their expectations were reflected in their head movements and glances to and from the horse's hooves. When the correct number of hoof taps was reached, the questioners almost always looked up, thereby signaling Hans to stop (Rosenthal, 1966).

Some fifty years later, Robert Rosenthal began to investigate the importance of the expectations of experimenters in psychological research.

In his now classical article, Rosenthal (1963) presented evidence suggesting that the experimenter's knowledge of the hypothesis could serve as an unintended source of variance in experimental results. In a prototypical study, Rosenthal and Fode (1963) had naive rats randomly assigned to two groups of undergraduate experimenters in a maze-learning task. One group of experimenters was told that they were working with maze-bright animals and the other group was told that their rats were maze-dull. The group of experimenters which was led to believe that their rats were maze-bright reported faster learning times for their subjects than the group which was told their animals were maze-dull. An extension of this finding to the classroom was offered by Rosenthal and Jacobson (1966). Teachers were led to believe that certain, randomly selected students in their classrooms were "late bloomers" with unrealized academic potential. Pre- and post-testing in the fall and spring suggested that children in the experimental group (late bloomers) had a greater increase in IQ than did the controls.

The purpose of this section will be to examine the problem of experimenter-observer bias with regard to naturalistic observational procedures. The amount of literature which deals directly with observer bias in naturalistic observation is sparse (Kass & O'Leary, 1970; Skindrud, 1972; Kent, 1972). However, Rosenthal has written an extensive review of experimenter bias in behavioral and social psychological research (Rosenthal, 1966). In spite of failures to replicate many of Rosenthal's findings (Barber & Silver, 1968; Clairborn, 1969) and extensive criticisms of Rosenthal's methodology (Snow, 1969; Thorndike, 1969; Barber & Silver, 1968), the massive body of literature compiled and summarized by Rosenthal (1966) remains the

best available resource for conceptualizing the phenomenon of observer bias and for isolating possible sources of bias relevant to naturalistic observation. A brief review of this literature follows with a focus on integrating implications from this literature with naturalistic observational procedures. In addition, we will give consideration to the few experiments which have directly investigated observer bias in naturalistic observation and further consider some proposals for dxperiments yet to be conducted. Finally, suggestions for minimizing observer bias will be outlined and data on this problem from our laboratory will be presented.

Conceptualization of Observer Bias

Rosenthal (1966) has defined experimenter bias "as the extent to which experimenter effect or error is asymmetrically distributed about the 'correct' or 'true' value." Observer errors or effects are generally assumed to be randomly distributed around a "true" or "criterion" value. Observer bias, on the other hand, tends to be unidirectional and thereby confounding.

Sources of Observer Bias

An important distinction should be drawn between observer error and observer effect on subjects. Invalid results may be contributed solely by systematic or "biased" errors in recording by observers. Or, invalid findings may be realized as a result of the effect that the observer has on his subjects (Rosenthal, 1966). First we will consider recording error as a source of observer bias.

Kennedy and Uphoff (1939) illustrate the problem of recording errors in an experiment in extrasensory perception. The observers' task was simply to record the investigator's guesses as to the kind of symbol being "trans-

mitted" by the observer. Since the investigators guesses for the observers had been programmed, it was possible to count the number of recording errors. In all, 126 recording errors out of 11,125 guesses were accumulated among 28 observers. The analysis of errors revealed that believers in telepathy made 71.5 percent more errors increasing telepathy scores than did non-believers. Disbelievers made 100 percent more errors decreasing the telepathy scores than did their counterparts. Sheffield and Kaufman (1952) found similar biases in recording errors among believers and nonbelievers in psychokinesis on tallying the results of the fall of dice. Computational errors in summing recorded rates have also been documented by Rosenthal in an experiment on the perception of people (Rosenthal, Friedman, Johnson, Fode, Schill, White, & Vikan-Kline, 1964).

It is doubtful that these recording and computational errors were intentional. However, as Rosenthal (1966, p. 31-32) notes, data fabrication or intentional cheating is not absent in psychological research, especially where undergraduate student experimenters are employed as data collectors. Rosenthal points out that these students "have usually not identified t a great extent with the scientific values of their instructors." Students may fear that a poor grade will be the result of an accurately observed and recorded event which is incompatible with the expected event. Of two experiments by Rosenthal which were designed to examine intentional erring by students in a laboratory course in animal learning, one revealed a clear instance of data fabrication (Rosenthal & Lawson, 1964) and the other showed no evidence of intentional erring but did show some deviations from the prescribed procedure (Rosenthal & Fode, 1963). Another study employing student experimenters by Azrin, Holz, Ulrich, and Goldiamond (1961) replicated

Verplanck's (1955) verbal conditioning experiment. However, an informal post-experimental check revealed that data had been fabricated by the student experimenters. Later, the authors employed advanced graduate students as experimenters and found that Verplanck's results were not replicated.

The implications for naturalistic observation are obvious. Observer error, whether it be unintentional or intentional, incurred during recording or during computation, must be guarded against by accuracy checks and by carefully concealing the experimenter's hypotheses. Although observer agreement checks do not rule out the possibility of bias among the observers whose data is compared, it at least arouses suspicion where agreement figures are low and disagreements are consistent. Ideally, observers should not be made responsible for the tallying of their own data. Computations should be made by a nonobserver who is removed from knowledge of the observations. Observers should be selected on the basis of their identification with scientific integrity and admonitions against possible biasing effects should be repeated during the course of the experiment. Finally, observers should be encouraged to disclose to the experimenter both the nature and sources of any information they receive that might be relevant to the objectivity of their observations. A questionnaire, filled out after observation sessions, can facilitate this disclosure.

The other source of observer bias, which Rosenthal discusses (Rosenthal, 1966), is the effect of the observer's expectancy on the subject. If an observer has an hypothesis about a subject's behavior, he may be able to communicate his expectations and thereby influence the behavior.

Expectancy effects have previously been alluded to in Rosenthal's study with animal laboratory experimenters (Rosenthal & Fode, 1963) and

teachers in the classroom (Rosenthal & Jacobson, 1966). Rosenthal's first major study in expectancy effects is instructive in its simplicity. Rosenthal and Fode (1963) had 10 experimenters obtain ratings from 206 subjects on the photo person-perception task. All 10 experimenters received identical instructions except that five experimenters were informed that their subjects would probably average a +5 success rating on the ten neutral photos while the other five experimenters were led to expect a -5 failure average. The results revealed that the group given the +5 expectation obtained an average of +.40 vs. the -5 expectation group which yielded a -.08 score. These differences were highly significant and subsequent replications have supported these findings (Fode, 1960; Fode, 1965).

The implications for naturalistic observational procedures of the expectancy effect on the subject's behavior are most discomforting. If, as in the Rosenthal laboratory studies, observers in the natural setting can communicate their expectancies to their subjects such that the subject's behavior falls in line with these expectations, a serious threat to internal validity is posed. Assuming that humans are no less sensitive to subtle cues than Mr. von Osten's Clever Hans, it seems reasonable to infer that observer expectancy effects are operative in the natural setting. Consider the not atypical case of an observer who records selected deviant behaviors of a child in a classroom before, during, and after treatment. Seldom is it not obvious to the observer when treatment begins and ends. Assuming that an observer might infer the expectations of the experimenter in such a setting, how might he communicate these expectations to his subjects? One way of influencing the targeted child is by nonverbal expressive cues. Expressions of amusement by the observer during baseline might inflate deviant behaviors. During intervention, expressions of disapproval or

caution by the observer might reduce the subject's deviant rate. These biasing effects may be systematic and confounding.

Although few studies have systematically assessed the effects of observer bias in the natural setting, many field investigators have taken note of the expectancy phenomenon, and have included procedures to minimize its effect. One such technique is to mask changes in experimental conditions (e.g., Thomas, Becker, & Armstrong, 1968). Another is to keep observers unaware of assignment of subjects to various treatment or control conditions (e.g., O'Conner, 1969). The addition of new observers in the last phase of a study who are naive to previous manipulations is another approach (e.g., Bolstad & Johnson, 1972).

Three studies in the natural setting shed further light on expectancy effects with naturalistic observational procedures. Rapp (1966) had eight pairs of untrained observers describe a child in a nursery school for a period of one minute. One member of each observer pair was subtly informed that the child under observation was feeling "under par" that day and the other that the child was "above par." In fact, all eight children showed no such behaviors. Seven of the eight pairs of observers evidenced significant discrepancies between partners in their description of the nursery children in the direction of their respective expectations. Both recording errors and expectancy effects on the subjects' behavior may have contributed to this demonstration of observer bias.

A second study by Azrin et al. (1961) employed untrained undergraduate observers who were asked to count opinion statements of adults when they spoke to them. The observations of those who had been exposed to an operant interpretation of the verbal conditioning phenomenon under study were the exact opposite of those given a psychodynamic interpretation.

Again, both the expectancy effects of the observer on the subject and recording errors may have accounted for the observer bias. Post experimental inquiries by an accomplice student revealed that recording errors were the main factor. The accomplice learned that 12 of the 19 undergraduates questioned intentionally fabricated their data to meet their expectations.

A third study by Scott, Burton and Yarrow (1967) allows a comparison between the simultaneous observations of hypothesis informed (Scott herself) and uninformed observers. The observers coded behavior into positive and negative acts from an audio-tape recording of the targeted child and his peers. The informed observer's data differed significantly from the others' in the direction of the experimenters' hypothesis.

These three studies strongly suggest that data collected by relatively untrained observers are influenced by observer expectations. Do these findings generalize to the observations of professional observers who are highly trained in the use of sophisticated multivariate behavior codes? As indicated earlier, the amount of available research which directly pertains to this question is limited and somewhat equivocal.

Kass and O'Leary (1970) conducted the first systematic attempt to manipulate observer expectations in a simulated field-experimental situation. Three groups of female undergraduates observed identical videotaped recordings of two disruptive children in a simulated classroom. The observers were trained in nine category codes of disruptive behavior. Group I was then given the expectation that soft reprimands from the teacher would increase the rate of disruptive behavior. Group II was told that soft reprimands would decrease disruptive behavior. And, Group III was given no expectation at all about the effects of soft reprimands. Rationales were

given each group explaining the reasons for each specific expectation. The effects of these expectations were assessed by having the observers watch four days of baseline and five days of treatment data. The interaction between the mean rate of disruptive behavior in the three conditions and the two treatment conditions was significant at the .005 level, indicating the presence of observer bias. Ronald Kent (1972) has suggested that these reported effects of expectation bias were confounded with observer drift in the accuracy of recording. When different groups of raters, who are interreliable within groups, fail to frequently compute agreement between groups, they may "drift" apart in their application of the behavioral code. However, it should be noted that when this drift, comprised of recording errors, is alligned asymmetrically in the direction of the expectation, then the drift is, by definition, observer bias.

Skindrud (1972) attempted to replicate the findings of Kass and O'Leary (1970). Observers were divided into three groups, each group given a different expectation about video-taped family interactions. The first group was given the expectation that when the father was absent there would be more child deviant behaviors than when the father was present. A second group was given the opposite expectation. Appropriate rationales were provided for each of these two groups. An additional control group was added with no expectations provided regarding father-present or father-absent tapes. All observers were checked at the end of training on the rates of deviant behaviors they recorded and subsequently matched on this variable when assigned to conditions. Throughout the study, observer agreement data was collected randomly. During training, reliability was checked daily, and the average observer agreement prior to the beginning of the manipulation was 64%. The results of the study gave no evidence for observer bias. There were no

significant differences between groups and no significant interaction effects. There was little drift in the accuracy with which the code was used. Sequential reliabilities were computed for the increase, decrease, and control groups with average observer accuracy of 58.5%, 57.6%, and 58.4%, respectively. These accuracy figures were computed by comparisons with previously coded criterion protocols. The relatively small and consistent decline in accuracy is consistent with the failure to find bias.

A similar unsuccessful attempt to replicate Kass and O'Leary (1970) was reported by Kent (1972). Kent found that knowledge of predicted results was not sufficient to produce an observer bias effect. However, when the experimenter reacted positively to data which was consistent with the given predictions and negatively to inconsistent data, a significant observer bias effect was obtained.

The available literature dealing with observer bias in naturalistic observation is both sparse and contradictory. Furthermore, the few studies available have focused exclusively on only one source of observer bias, namely, recording errors or errors of apprehension. Thus far, no one has systematically investigated the effects of the observer's expectancies on the subjects' behavior. In the three studies reported above, all observations were made from video-taped recordings. There were no opportunities for the observers to communicate their expectancies to their subjects. Yet, in most studies employing naturalistic observational procedures, observers do have that opportunity.

An important study which needs to be conducted is one which examines the observer's expectancy effects on the subject. First, it would be interesting to determine if observers could nonverbally communicate their

expectancies to subjects such that the subject's behavior changes in the direction of the expectancy. The next step, of course, would be to replicate this same design without specifically asking observers to attempt to influence subjects, but merely to give them an expectation.

Perhaps the most important test of observer bias effects will be the one which combines recording errors and effects of observer expectancy on subjects in the naturalistic setting. One can question the generalizability of highly controlled laboratory studies to live observations and to research projects in which the observers are more invested in the outcome of the research. The generalizability of studies which employ only taped versions of a subject's behavior is further limited by excluding the possible effects of an observer's expectancy on his subject's behavior.

Another variable which seems crucial to observer bias in the naturalistic setting is the observer's responsiveness to admonitions to remain scientific, objective, and impartial in the collection of data. Rosenthal (1966) stresses the importance of the experimenter-observer's identification with science and objectivity. He cites evidence suggesting that graduate students obtain less biased data than undergraduates and interprets this difference as a function of identification with science. Perhaps observers who are repeatedly reminded to be impartial might be less susceptible to the influence of biasing information than observers not given these admonitions.

A dimension which seems important in considering observer bias is the specificity of the code. In most of the Rosenthal literature, the dependent variable is scaled between such global poles as success and failure. Intuitively, it seems logical that the more ambiguous the dependent measure, the greater the possibility for bias. A multivariate coding system, with

well-defined behavioral codes might be expected to restrict interpretive bias. This is an empirical question worthy of examination.

Another variable which might greatly affect observer bias is observer agreement. The greater the observer agreement, the less likely is observer bias, even among observers with the same expectancy.

Until more information is available on observer bias effects in naturalistic observation, it seems very critical to do everything possible to minimize the potential for these effects. Whenever possible, observers should not have access to information that may give rise to confounding consequences and encouraged to reveal the nature and source of any information they do receive. In our research, we are currently observing both families in clinical treatment and "normal" or nontreated families. Knowledge of a family's status might seriously affect the observer's data. Also, knowledge about treatment stages (baseline, mid-treatment, post-treatment, and follow-up) might effect the observers' data. After each observation, it is our policy to have observers fill out a questionnaire concerning the nature and source of any biasing information. Thus far, of 75 observations of referred families, observers have considered themselves informed only 36% of the time. And, in all of these cases, their information was correct. This information usually comes from a member of the family being observed (56%). Other sources of information include information leaks from the therapists (12%), the Child Study Center Clinic generally (16%), and other sources (16%). Of the observer considering themselves informed as to the clinic vs. "normal" status of the families, 29% also considered themselves informed as to treatment stage, but only two-thirds of these observers were correct in their discrimination. In only 20% of the cases did the observer actually know

the status of the case (i.e., clinic vs. normal) and the treatment stage (baseline vs. after baseline). Of the observers considering themselves completely uninformed of the families' status, their guessing rate (clinical or "normal") barely exceeded chance at 51%. Their guesses as to the four stages of treatment were 36% correct and 80% correct on the discrimination between baseline and after baseline.

Of the "normal" families seen, observers have considered themselves informed as to family status only 17% of the time. However, in only 45% of these cases were the observers actually correct in making the discrimination. In the uninformed observations, however, observers were able to guess the family's status correctly 75% of the time.

Not only are these questionnaires beneficial in gauging the amount of potentially biasing information that observers discover, but they are helpful in two other ways as well. First, by revealing sources of information leakage, steps can be made to eliminate these sources. Second, questionnaires, given after each family is observed, serve as a regular reminder for the importance of unbiased, objective recording of behavior.

It is difficult to make any firm conclusions about the presence or absence of observer bias in naturalistic observation. Clearly, more research is needed on this question. However, it should also be clear that the potentially confounding influence of observer bias cannot be ignored and that steps can and should be taken to minimize its possible effect.

The Issue of Reactivity in Naturalistic Observation

In the previous section, we have considered the effects of an observer's bias in naturalistic observation. In this section, we will discuss the effect of the observer's presence on the subjects being observed. Whereas observer bias can potentially invalidate comparisons by confounding influences, the reactive effects to being observed primarily constitute a

threat to the generalizability of the findings. That is, subjects' observed behavior in the natural setting may not generalize to their unobserved behavior. Webb, Campbell, Schwartz, and Sechrest (1966) have defined reactivity in terms of measurement procedures which influence and thereby change the behavior of the subject. Weick (1968) has also referred to reactivity as "interference" or the intrusiveness of the observer himself upon the behavior being observed. Clearly, situations which are highly reactive in terms of "observer effects" are not likely to be generalizable to situations in which such effects are absent.

Reactive effects have been studied with two basic paradigms: a) by the study of behavioral stability over time and b) by comparison of the effects of various levels of obtrusiveness in the observation procedure. In employing the first method, investigators have typically examined behavioral data for change over time in the median level and variance of the dependent variable. In general, it has been assumed that change reflects initial reactivity and progressive adaptation to being observed. This interpretation is particularly persuasive if there is an obvious stability in the data after some initial period of change or high variability. While this is a viable way of checking for reactivity effects, it is a highly indirect method and relies on assumptions concerning the causes of observed change. It is obvious that other processes could account for such change. Furthermore, the lack of change certainly does not indicate a lack of reactive effects. The second method, comparing obtrusive levels of observation, appears less inferential than the first method. The problem with this method is that it only provides a picture of relative degrees of reactivity between obtrusiveness levels; it does not provide a measure of the degree of reactivity relative to the true, unobserved behavior. However, this problem can

be remedied if one of the observational treatments in the comparison is totally unobtrusive or concealed.

To what extent does reactivity occur in naturalistic observation? The literature addressing this question is commonly reported in reviews to be contradictory (Wiggins, 1970; Weick, 1968; Patterson & Harris, 1968). Several studies have been cited as providing evidence for the position that reactive effects may be quite minimal. Others have been cited which suggest that reactive effects are quite pronounced. The purpose of this review is to: a) reconsider the contradictions in the literature on reactivity, b) tease out those factors which seem to account for reactivity, and c) propose further investigations which isolate these factors.

In a number of reviews on reactivity, several studies have been consistently cited which support the position that reactivity does not constitute a major threat to generalizability. One study frequently cited is the timely investigation of a Midwest community by Barker and Wright (1955). In this admirable study, careful naturalistic observations were made of children under ten years of age and their daily interactions with peers and parents. The authors assumed that reactive effects were short lived and that the adults and other members of the families quickly habituated to the presence of the observers. In addition, it was reported that, with the younger subjects in the sample, reactive effects were slight. However, these findings should be interpreted with much cautior What is easily lost sight of in the summaries of this work is that the observers in this study were free to interact with the subjects in a friendly but nondirective manner. In fact, the basis for the authors' conclusion that reactive effects were not pronounced was the

finding that "only" 20% of the children's behavioral interactions were with the observer. Allowing the observer to interact with the subject must certainly have increased the intrusiveness of the observer and provided the opportunity for the observer to influence the subject's behavior. The authors' other conclusion that reactivity, as measured by frequency of interactions, positively correlated with age is also suspect in that children below the age of five were not always informed that they were being observed, whereas children above this age were.

Another study commonly cited in support of the minimal reactivity position is that of Bales (1950). In this controlled laboratory investigation, the behavior of a discussion group was not found to be changed by three levels of observer conspicuousness. This finding, however, may be limited to the laboratory setting.

Two additional studies, frequently mentioned as supportive of the minimal reactivity argument, made use of radio transmitter recording in the naturalistic environment. Goskin and John (1963) had a married couple wear a transmitter the entire time they were on a two-week vacation. Purcell and Brady (1965) outfitted adolescents in a treatment center with a similar recording device for one hour a day. When the protocols in both studies were examined for the frequency of comments about being observed or listened to, it was found that such references declined to a zero level either during the first or second day of recording. This is not to say, of course, that these subjects were not still aware of, and affected by, the recording device; the results only indicate that the subjects talked about the device less after the first day.

A recent investigation by Martin, Gelfand, and Hartmann (1971) can also be interpreted as providing evidence for low levels of reactivity to

observation. This study involved 100 elemntary school children, ages 5 to 7. Equal numbers of male and female subjects were assigned to five observation conditions following exposure to an aggressive model: a) observer absent, b) female adult observer present, c) male adult observer present, d) female peer observer present, and e) male peer observer present. During the free-play session, the subjects' aggressive behavior was recorded by observers behind a one-way mirror. No significant differences in aggressive behaviors were obtained between the observer-present and observer-absent conditions. The absence of differences between these two levels of obtrusiveness in observation suggests little or no reactivity to the presence of an observer. Within the observer-present condition, however, it was found that peer observers significantly facilitated imitative aggressive responding in both boys and girls compared to adult observers. Also, there was more imitative aggression when the observer was the same sex as the subject. The girls, but not the boys, showed significant increases in aggressive output over time when the observer was present but not when the observer was absent. This latter finding suggests that girls manifest initial reactivity to the presence of an observer but later habituate to the observer's presence. It is interesting that both paradigms for measuring reactivity were used in this investigation and that each method supports different conclusions about the degree of reactivity. In considering the generalizability of these findings to naturalistic observation procedures, it should be noted that observers in this study were instructed to not pay attention to the subjects and were either seated facing away from the subjects (adult observers) or given a coloring task to complete (peer observers). With naturalistic observation procedures, on the other hand, observers typically pay very close attention to their subjects.

for the most part, the studies which have been reported to provide evidence for minimal levels of reactivity to observation have been based on data of questionable meaning or restricted to highly specific circumstances (e.g., Bales, 1950; Martin et al., 1971).

Many other studies have been cited as demonstrating considerable reactive effects of observation in naturalistic settings. One such study is that of Polansky, Freeman, Lorwitz, Irwin, Rainis, Rappaport, and Whaley (1949). These investigators observed delinquent children in a study of group emotional contagion phenomenon. The children were informed that the observers were studying their reactions to various aspects of the summer-camp program. The authors report that during the first week of observations, the children essentially ignored the presence of the coders. But, during the second week, many "blow-ups" occurred which were directed against the coders, especially by the older children. The authors speculate that the aggressiveness of the children can be explained as a result of resistance to being observed. They also conceded that the resistance hypothesis was confounded by "the second week syndrome" which they describe as an increasing anti-adult aggressiveness that typically evolves after the children have adjusted to the camp, peaking in the second week. It is unclear as to what to conclude from this study about reactivity. Was reactivity most prevalent when children were aggressive toward the observers in the second week? Or, was reactivity most intensive during the first week when the delinquent children were "suppressing" aggression prior to habituating to the unfamiliar environment. And, more importantly, how much of this pattern of cooperation in the first week and anti-adult aggressiveness in the second would have occurred in the absence of observers? This question is left unanswered by the Polansky et al. (1949) study.

A more sophisiticated study illustrating considerable reactive effects is that of Roberts and Renzaglia (1965). In this study, eight therapists saw two "clients" (students solicited from an introductory psychology course) for three sessions. The three conditions were: a) with a tape recorder in the room, b) microphone only, c) with a concealed microphone and a concealed recorder. While no differences were found as to the number of minutes that subjects talked in each condition, it was found that the clients made more favorable self-references when they knew they were being recorded (conditions a and b, and more unfavorable self-references when they did not realize they were being recorded (condition c). While the generalizability of this study is limited by the small n and structured situation employed, it does provide evidence for reactivity effects.

The Bechtel (1967) study has also been presented in reviews as a study demonstrating considerable reactive effects. Bechtel studied the movements of people in an art museum with an automated recording device, called an hodometer. In one condition, subjects were requested to go through the art exhibit in one room and rank the prints according to preference. Another group was given the same instructions and also informed that they were being observed. For the first group, subjects were found to leave the room with an average latency of 313 seconds, covering an average area of 76 square feet. The second group left after 220 seconds and covered only 58 square feet. Bechtel concludes that being observed constituted an aversive situation which led to escaping the room faster. No statistics were provided to determine whether or not the reported averages were significantly different, nor were the number of subjects in each condition reported. Unfortunately, this kind of study does not provide any information about habituation effects, since subjects were allowed to leave the setting.

[illegible] room environment. In [illegible] normal [illegible] living room environment [illegible]

[illegible] activity level [illegible] activity [illegible] of the laboratory setting.

Few studies have [illegible] observer reactivity in the natural setting [illegible], and [illegible] (19[illegible]) provide evidence for [illegible] observer. [illegible] families [illegible] even after six hours of observation.

A study by Patterson and Harris [illegible] evidence for considerable reactive effects [illegible] observation in a naturalistic environ-

ment. This article is the only study available which was designed specifically to manipulate and measure observer effects in the homes of the families observed. In this study, data obtained from mothers on their own families were compared with the data on the same family collected by an outside observer. There were three conditions, with five families per condition: a) mothers collected the first five ten-minute sessions of observational data and an outside observer collected the second five sessions of data on the child and father only (M-O), b) the observer collected all ten sessions as a test for habituation effects (O-O), and c) the mothers collected all ten sessions as a control for habituation effects (M-M). The dependent variables were the rates of total behaviors and the rate of deviant behaviors. A problem in the research design of this study should be noted. The mother was present in the family as a participant in the second condition (O-O) and the second half of condition a (M-O), but was not a participant when she was an observer in condition c and the first half of condition a. These comparisons are confounded by mother presence and absence. In spite of this confound, which would probably bias in favor of showing group differences, no main effects for groups were found in analysis of variance for either the rate of total interactions or deviant behaviors. Thus, on the initially selected dependent variables, no reactive effects were apparent.

Patterson and Harris also divided their groups into high and low rate interactors on the basis of the first five sessions. On the frequency of total interactions measure, high rate interactors in the first five sessions showed significant reductions in rate during the last five sessions. The authors describe this decline as a "structuring effect" in that the subjects appeared to program some activity together in the first five sessions.

Conversely, the low rate interactors in the first five sessions showed slight increases in rates during the last five sessions. The authors describe this transition as an habituation effect in that subjects initially involved themselves in solitary activities or attempted to escape the observational situation but later adjusted to it and interacted more. In general, there were no changes in deviant behavior from the first set of five observations to the last set of five. The only significant finding was that subjects who displayed low rates of deviant behavior in the first five sessions (under the M-0 condition) increased their rate in the last five sessions. However, it is possible that the mothers were recording less deviant behaviors and more positive behaviors in the first five sessions than were the observers in the second five sessions, thus contributing differentially to main trials effects. An observational study by Rosenthal (1966) supports such a thesis. He found that parents tended to code more positive changes in their children than were actually present. And, Peine (1970) found that parents were less observant of their children's deviant behaviors than were nonparent observers.

Patterson and Harris conclude that "generalization about 'observer effects' should probably be limited to special classes of behavior " (p. 16). A more recent study by Patterson and Cobb (1971) analyzed the stability of each of the 29 behavior codes used in their coding system. If it is assumed that individuals adapt to the presence of an observer over time, then a repeated measures analysis of variance should reveal differences in the mean level of various behaviors. Patterson and Cobb analyzed data for 31 children from problem and nonproblem families over seven baseline sessions. None of the changes in mean level for the codes produced a significant effect over time. The investigators conclude that the observation data were

fairly stable for most code categories. It is possible, of course, that had observations continued over a longer period of time, significant changes in mean level for some behaviors would have been discovered. Given that families were rarely observed on consecutive days by the same observer, it is possible that different observers could have resensitized the families each day, thereby extending the period required for adaptation.

In summary, there are a few well-designed studies which have discovered reactive effects (e.g., Roberts and Renzaglia, 1965; Bechtel, 1967; White, 1972), but there are several others where the meaning of the results is unclear. There can be little doubt that the entire question has been inadequately researched. Any general conclusions about the extent of reactivitv in naturalistic observation would seem premature at this time.

As White (1972) points out, the finding of reactive effects seems to depend on many factors, including the setting (e.g., home, school, laboratory), the length of observation, and the constraints placed on subjects by the conditions of observation (e.g., no television during observations, remain within two adjacent rooms, etc.). Furthermore, it should be realized that reactivity may or may not be discovered depending upon what paradigm of measurement is used (e.g., Patterson & Harris, 1968; Martin et al., 1971) and what variables are analyzed as dependent variables (e.g., Roberts & Renzaglia, 1965; White, 1972). Unless these factors are controlled for in comparing experiments on reactivity, both contradictions and consistencies as to the relative presence or absence of reactivity may falsely appear.

Assuming that reactivity to being observed in naturalistic settings does occur, even if only to some minimal degree, the critical task is to localize the sources of interference so that they can be dealt with more

directly. Four such sources will be discussed and experiments will be proposed to measure the extent of their instrusiveness.

Factor 1: Conspicuousness of the Observer

The literature points to the level of conspicuousness or intrusiveness of the observer as an important factor contributing to reactivity. Presumably, the more novel and conspicuous the agent of observation, the more distracting are the effects upon the individuals being observed. It would also follow that longer habituation periods would be required for more distracting observational agents in order to achieve stability of data.

Bernal, Gibson, William, and Pesses (1971) compared two observation procedures which would presumably vary on obtrusiveness. These investigators compared data collected by an observer with that collected by means of an audio tape recorder which was switched on by an automatic timing device. The family members involved in this study were aware of the presence of the recorder but were unaware of the exact time of its operation. The primary purpose of this study was to explore the feasibility of the audio tape method and to explore the relationship of data collected by the two methods rather than to study reactivity per se. The results indicated that, during the same time interval, there was a high relationship between the mother's command rate as coded by the observer and from the tape (r = .86) but that the observer coded more commands. Similar results were obtained when the observer's data was compared with data based on coding of the audio tapes from different time intervals. The question arises as to how much of this latter discrepancy was due to differences in levels of reactivity and how much was due to differences associated with the source of coding. The authors point out, for example, that the observer could code gestural

commands while the coder using the tape could not. Since the discrepancies at the same time and at different times were of the same general order of magnitude, it is likely that most of the observed difference across time was due to the material on which coding was based rather than to differences in subject reactivity. To study the impact of reactivity effects separately, one might design such a study so that the same stimulus materials would be used for coding.

We are currently completing a study on reactivity which employs this strategy to compare reactivity associated with an observer present in the home carrying a tape recorder vs. the tape recorder alone. This study involves six days of observation for 45 minutes per day with single-child families. The two conditions are alternated so that the observer is present one evening and not present the next. The observer is actually a "bogus" observer. All behavioral coding is done on the basis of the tapes. It is our suspicion that reactivity to the tape recorder will be short lived and minimal compared to the reactivity associated with the observer present.

If these hypotheses are substantiated in this and other research, alternatives to having an observer present in the home should be explored. One solution to be seriously considered would be extended use of portable video or audio tape recording equipment. These recording devices could remain in the homes over an extended observation period to facilitate habituation effects. In addition, the devices could be preprogrammed to turn on and off at different times during the day so that the observed would not know when they are in operation (as in Bernal et al., 1971). This solution, which would, of course, require full knowledge and consent of the parties involved, appears to be a promising one for attenuating reactivity effects as well as solving problems of observer bias.

Factor 2: Individual Differences of the Subjects

Some people might be expected to manifest more reactivity to the presence of an observer than others. A "personality" variable such as guardedness might be correlated with degree of reactivity. For example, scores on the K scale of the MMPI (or other comparable tests) might be related to the effects of being observed in a natural setting.

The literature also suggests that age is correlated with reactivity. Several authors (Barker & Wright, 1955; Polansky et al., 1949) have suggested that younger children are less self-conscious and thereby less subject to reactive effects than older children. The Martin et al. (1971) study also suggests that sex might be an important factor accounting for different levels of reactivity. Experiments are needed which compare these individual difference variables in the natural setting with naturalistic observation procedures.

Factor 3: Personal Attributes of the Observer

Evidence from semi-structured interviews suggests that reactive effects may also be contributed by the unique attributes of the observer. Different attributes of the observer may elicit different roles on the part of the subject, depending upon what might be appropriate given the observer's attribute. Rosenthal (1966) reports several such attributes that have been demonstrated to yield differential effects, including the age of the observer,

sex, race, socio-economic class, and the observer's professional status (i.e., undergraduate observer vs. Ph.D. therapist). Martin et al. (1971) also discovered that both the factors of age and sex of the observer had differential effects on the subjects being observed. Varying any of these dimensions parametrically would be relatively simple in investigating this problem in the natural setting.

Factor 4: Rationale for Observation

Another factor that may be important in accounting for reactivity is the amount of rationale given subjects for being observed. Whereas the Bales (1950) study found no differential reactivity of three levels of observer conspicuousness in a group-discussion setting, Smith (1957) found that nonparticipant observers aroused hostility and uncertainty among participating group members. Weick (1968) suggests that this discrepency may have been a function of different amounts of rationale for the presence of an observer. We hypothesize that a thorough rationale for being observed might be expected to reduce guardedness, anxiety, etc., and thereby reduce the reactivity.

Observer reactivity is a problem that cannot be easily dismissed for naturalistic observation. There is sufficient evidence to suggest that observer reactivity can seriously limit the generalizability of naturalistic observation data. Clearly, factors accounting for reactivity need to be investigated and solutions derived to minimze the effects of the observer on the observed. In the next section, we will describe how reactivity, in addition to posing a problem for generalizability, can also interact with and confound the dependent variable.

Observee Bias:

Demand Characteristics, Response Sets and Fakability

Reactivity to observation will always be a problem for naturalistic research, but it would be a relatively manageable one if we could assume it to be a relatively constant, noninteractive effect. That is, if we knew that the presence of an observer reliably reduced activity level or deviant behavior by 30%, for example, the problem would not be too damaging to research investigations involving groups of subjects. But, what if the observee's reactivity to being observed interacts with the dependent variable under study.

Let us take the example of a treatment study on deviant children in which observations are taken prior to and after treatment. Prior to treatment, the appropriate thing for involved parents or teachers to do is to make their referred child appear to be deviant in order to justify treatment. The appropriate response at the end of treatment, on the other hand, is to make the child appear improved in order to justify the termination, please the therapist, etc. These are the demand characteristics of the situation. In this case, the reactivity to being observed is not constant or unidirectional, but interacts with and confounds the dependent variable. It is possible that any improvement we see in the children's behavior is simply the result of differential reactivity as a consequence of the demand characteristics of the situation. Now, let us suppose we employ a wait list control group and collect observational data twice before beginning treatment and at the same interval as used for the treated group. This procedure provides an excellent pretest-post-test control for our treated group. But, what of the demand characteristics of this procedure? On the first assessment, the involved

parents or teachers will probably behave in the same general way as their counterparts in the treated group, but by the second observation they may be more desperate for help and even more concerned to present their child as highly deviant. Thus, simply as a result of the demand characteristics involved, we might expect our treatment group to show improvement while the control groups would show some deterioration.

We also may wish to compare our referred children with children who are presumably "normal" or at least not referred for psychological treatment. Once again, however, we might anticipate that parents recruited for "normative" research on "typical" families would be more inclined than our parents of referred children to present their wards as nondeviant or good. In other words, a response set of social desirability could be operative with this sample making them less directly comparable to the referred sample.

These arguments would, of course, be even more persuasive if we were dealing with the observed behavior of the adults themselves. The foregoing observations on children assume, however, that the involved adults are capable of influencing children to appear relatively "deviant" or "normal" if they wish to do so (i.e., that observational data on children is potentially fakable by adult manipulation).

We have just completed a study (Johnson & Lobitz, 1972) which was directed at testing this assumption. Twelve sets of parents with four- or five-year-old children were instructed to do everything in their power to make their children look "bad" or "deviant" on three days of a six-day home observation and to make their children look "good" or "nondeviant" on the remaining three days. Parents alternated from "good" to "bad" days in a counterbalanced design.

Four predictions were made regarding the behavior of both children and parents. During the "fake bad" periods, it was anticipated that, relative to the "fake good" periods, there would be:

a) more deviant child behaviors,

b) a lower ratio of compliance to parental commands,

c) more "negative" responses on the part of parents, and

d) more parental commands.

Predictions a, c, and d were confirmed at or beyond the .01 level of confidence. Only the child's compliance ratio failed to be responsive to the manipulation. It will be recalled from the section on reliability that this statistic is by far the least reliable and thus the least sensitive (statistically) to manipulation. These results which demonstrate the fakability of naturalistic behavioral data indicate that this kind of data may potentially be confounded by demand characteristics and/or response sets.

We are aware of only one other study involving naturalistic observation which helps demonstrate this problem (Horton, Larson, & Maser, 1972). This study involved one teacher who was under the instruction of a "master" teacher for the purpose of raising her classroom approval behavior. She was observed, without her knowledge, by students in the class. The results clearly showed that her approval behavior was at a much higher rate when she was being observed by the "master" teacher than when she was not being observed. Generalization from overtly observed periods to periods of covert observation was very minimal indeed. More generalization was found when the "master" teacher's presence in the classroom was put on a more random schedule. This study is not completely analogous to most naturalistic research because, in this case, the observer and trainer were the same person

and the study is limited in generalizability because of the N = 1 design. Yet, in most cases, the observed are aware that the collected observational data will be seen by the involved therapist, teacher, or researcher, and if the problem exists for one subject, it is a potential problem for all subjects. Observee bias is really a special case of subject reactivity to observation. Thus, the potential solutions outlined in the previous section apply here as well. In general, we suspect that observation procedures which are relatively unobtrusive and which allow for relatively long periods of adaptation will yield less reactivity and observee bias.

Validity of Naturalistic Behavioral Data

Just as behaviorists have ignored the requirement of classical reliability in their data, they have also neglected to give any systematic attention to the concept of validity. Most research investigations in the behavior modification literature which have employed observational methods have relied on behavior sampling in only one narrowly circumscribed situation with no evidence that the observed behavior was representative of the subject's action in other stimulus situations. In addition, behaviorists have largely failed to show that the obtained scores on behavioral dimensions bear any relationship to scores obtained or the same dimensions by different measurement procedures. This fact calls into serious question the validity of any of this research where the purpose has been to generalize beyond the peculiar circumstances of the narrowly defined assessment situation. Of course, the methodological problems we have presented thus far all pose threats to the validity of the behavioral scores obtained. But, we would argue that even if all these problems could somehow be magically solved, the requirement for some form of convergent validity would still be essential.

As with reliability, there are many different methods of validation, but as Campbell and Fiske (1959) point out:

> Validation is typically convergent; a confirmation by independent measurement procedures. Independence of methods is a common denominator among the major types of validity (excepting content validity) insofar as they are to be distinguished from reliability. . . . Reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods.

Thus, convergent validity is established when two dissimilar methods of measuring the same variable yield similar or correlated results. Predictive validity is established when the measure of a behavioral dimension correlates with a criterion established by a dissimilar measurement instrument.

With only a few exceptions, behaviorists have restricted themselves to face or content validity. And, of course, it must be admitted that the face validity of narrowly-derined behavioral variables is often quite persuasive. This is particularly true in cases where the behavioral dimension under study has very narrow breadth or "band width." After all, a behaviorist might argue, what can be a more valid measure of the rate of a child's hitting in the classroom than a straight-forward, accurate count of that hitting. While this argument is persuasive, two counter arguments must be considered. First, because of all of the methodological problems which we have presented thus far, we can never be certain that the observed rates during a limited observation period are completely valid or generalizable even to very similar stimulus stiuations. While many of the problems we have outlined can be solved and othersattenuated, it is unlikely that all will ever be completely eliminated. Second, is it not still of consequence to know whetner our behavior rate estimates have any relationship to other important and logically related external variables? Is it not important,

for example, to know whether or not the teacher and classmates of an observed high-rate hitter perceive this child as a hitter? It does seem important to us, particularly for practical clinical purposes, since we know that people's perceptions of others' behavior often have more to do with the way they treat them than does the subject's actual behavior. The need for establishing some form of convergent validation becomes even more profound as the behavioral dimensions we deal with increase in band width. As we begin to talk about such broad categories as appropriate vs. inappropriate behavior (e.g., Gelfand, Gelfand, & Dobson, 1967), deviant vs. nondeviant behaviors in children (e.g., Patterson, Ray, & Shaw, 1969; Johnson et al., 1972), or friendly vs. unfriendly behaviors (e.g., Raush, 1965), we are labeling broader behavioral dimensions. At this level, we are dealing with constructs, whether we like to admit it or not, and the importance of establishing the validity of these constructs becomes crucial. In most cases, these broad behavior categories have been made up of a collection of more discrete behavior categories and, in general, the investigators involved have simply divided behaviors into appropriate-inappropriate or deviant-nondeviant on a purely a priori basis. While the categorizations often make a good deal of sense (i.e., have face validity), this hardly seems a completely satisfactory procedure for the development of a science of behavior.

We have had to face this problem in our own research, where we have sought to combine the observed rates of certain coded behaviors and come up with scores reflecting certain behavioral dimensions. The most central dimension in this research has been the "total deviant behavior score" to which we have repeatedly referred in this chapter. Let us outline here the procedures we have used to explore the validity of this score. Although

we had a pretty good idea of which child behaviors would be viewed as "deviant" or "bad" in this culture, we attempted to enhance the consensual face validity of this score by asking parents of the "normal" children we observed to rate the relative deviancy of each of the codes we use in our research. Thus, in our sample of 33 families of four- and five-year-old children, we asked each parent to read a simplified version of our coding manual and characterize each behavior on a three-point scale from "clearly deviant" to "clearly nondeviant and pleasing." We established an arbitrary cut-off score and characterized any behavior above this cut-off as deviant. This resulted in a list of 15 deviant behaviors out of a total of 35 codes. The second step in validating this score and our implicit deviant-nondeviant dimension was presented in a study by Acking and Johnson (1972). We had already divided our 35 codes into positive, negative, and neutral consequences. This categorization was done on a purely a priori basis with a little help from the data provided by Patterson and Cobb (1971) on the function of some of these codes for eliciting and maintaining children's behavior. We reasoned that behaviors which parents viewed as more deviant would receive relatively more negative consequences than would behaviors viewed as less deviant. To test this hypothesis, we simply rank ordered each behavior, first by the mean parental verbal report score obtained and second by the mean proportion of negative consequences the behavior received from family members. The results of this procedure are presented in Table 1. Not all 35 behaviors are

---
Insert Table 1 about here
---

included in this analysis, but the complex reasons for this outcome can more parsimoniously be explained in a footnote [5] In any case, the Spearman Rank Order Correlation between the two methods of characterizing behaviors

on the deviant-nondeviant dimension was .73. This was an encouraging finding, but we noticed that the most dramatic exceptions to a more perfect agreement between the two methods involved the reasonable command codes (command positive, and command negative). These codes are used when the child reasonably asks someone to do something (positive command) or not to do something (negative command). Naturally, most parents felt that these innocuous responses were nondeviant. But, behaviorally, people don't always do what they are asked to by a four- or five-year-old child, and since noncompliance was coded as a negative consequence, it seemed that this artifact of our characterization might have artificially lowered this coefficient. By eliminating these two command categories from the calculation, the correlation coefficient was raised to .81.

The third piece of evidence for the validity of the deviant behavior score comes from the Johnson and Lobitz (1972) study already reviewed in the previous section. In this study, parents were asked to make their children look "good" and "nondeviant" for half of the observations and "bad" or "deviant" on the other half. They were <u>not</u> told how to accomplish this, nor were they told what behaviors were considered "bad" or "deviant." The fact that the deviant behavior score was consistently and significantly higher on the "bad" days lends further evidence for the construct validity of the score.

While evidence for the convergent or predictive validity of behavioral data is difficult to find in the literature, there are some encouraging exceptions to this general lack of data. Patterson and Reid (1971), for example, found an average correlation of .63 ($p < .05$) between parents' observations of their children's low rate referral symptoms on a given day and

the trained observer's tally of targetted deviant behaviors on that day. Several studies have found significant relationships between behavioral ratings of children in the classroom and academic achievement (Meyers, Attwell, & Orpet, 1968; D'Heurle, Millinger, & Haggard, 1959; Hughes, 1968). The data base of these studies is somewhat different from that currently employed by most behaviorists because they involve ratings by observers on relatively broad dimensions, as opposed to behavior rate counts. For example, dimensions used in these studies included "coping strength," defined as ability to attend to reading tests while being subjected to delayed auditory feedback (Hughes, 1968), or "persistence," defined as ". . . uses time constructively and to good purpose; stays with work until finished" (D'Heurle, Mellinger, & Haggard, 1959). Nevertheless, these studies demonstrate the potential for behavior observation data to provide evidence of predictive validity. Two other studies (Cobb, 1969; Lahaderne, 1968) yield similar predictive validity findings based on behavioral rate data. Lahaderne (1968) found that attending behavior as observed over a two-month period, provided correlations ranging from .39 to .51 with various standard tests of achievement. Even with intelligence level controlled, significant correlations between attentive behavior and achievement were found. Cobb (1969) obtained similar results in correlating various behavior rate scores with arithmatic achievement, but found no significant relationship between these behavior scores and achievement in spelling and reading. These predictive validity studies are very important to the development of the field as they suggest that manipulation of these behavioral variables may well result in productive changes in academic achievement.

In our own laboratory, we are exploring the convergent validity of naturalistic behavioral data by relating it to measures on similar dimensions

in the laboratory which include: a) parent and child interaction behavior in standard stimulus situations similar to those employed by Wahler (1967) and Johnson and Brown (1969), b) parent behavior in response to standard stimulus audio tapes similar in design to those used by Rothbart and Maccoby (1966) and parent behavior in standardized tasks similar to those used by Berberich (1970), and c) parent attitude and behavior rating measures on their children. Unfortunately, at this writing, most of this data has not been completely analyzed, but an overall report of this research will be forthcoming. A recent dissertation by Martin (1971), however, was devoted to studying the relationships between parent behavior in the home and parent behavior in analogue situations. By and large, the results of this research indicated no systematic relationships between the two measures. The same general findings for parents' responses to deviant and nondeviant behavior were replicated in the naturalistic and the analogue data, but correlations relating individual parental behavior in one setting with that in the other were generally nonsignificant. We don't know, of course, which, if either, of the measures represents "truth" but this study underlines the importance of seriously questioning the assumptions usually made in any analogue or modified naturalistic research. As Martin (1971) points out, these negative results are very representative of findings in other investigations where naturalistic behavior data has been compared to data collected in more artificial analogue conditions (e.g., see Fawl, 1963; Gump & Kounin, 1960; Chapanis, 1967).

Before closing this section on validity, we would like to briefly take note of the efforts of Cronbach and his associates to reconceptualize the issue of observer agreement, reliability and validity as parts of the

broader concept of generalizability. A full elaboration of generalizability theory goes far beyond the purposes of this chapter and the interested reader may be referred to several primary and secondary sources for a more complete presentation of this model (e.g., Cronbach, Rajaratnam, & Gleser, 1963; Rajaratnam, Cronbach, & Gleser, 1965; Gleser, Cronbach, & Rajaratnam, 1965; Wiggins, 1972). According to this generalizability view, the concerns of observer agreement, reliability and validity all boil down to a concern for the extent to which an obtained score is generalizable to the "universe" to which the researcher wishes the score to apply. Once an investigator is able to specify this "universe," he should be able to specify and test the relevant sources of possible threat to generalizability. In a typical naturalistic observational study, for example, we would usually at least want to know the generalizability of data across a) observers, b) occasions in the same setting, and c) settings. Through the generalizability model, each of these sources of variance could be explored in a factorial design and their contribution analyzed within an analysis-of-variance model. This model is particularly appealing because it provides for simultaneous assessment of the extent of various sources of "error" which could limit generalizability. In spite of the advantages of this factorial model, there are few precedents for its us . This is probably more the result of practical problems rather than a resistance to this intellectually appealing and theoretically sound model. Even if one were to restrict himself to the three sources of variance outlined above, the resulting generalizability study would, for most useful purposes, be a formidable project, indeed. Projects of this kind appear to us, however, to be well worth doing and we can probably expect to see more investigations which employ this generalizability model.

It should be pointed out at this point that the generalizability study outlined above does not really speak to the traditional validity requirement as succinctly defined by Campbell and Fiske (1969): "Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods." As stated earlier, to fulfill this requirement, one must provide evidence of some form of convergent validity by the use of methods other than direct behavioral observation. The generalizability model can, theoretically, handle any factor of this type under the heading of methods or "conditions," but the analysis-of-variance model employed requires a factorial design. Thus, it would seem extremely difficult and sometimes impossible to integrate factorially other methods of testing or rating in a design which encompassed the three variables outlined above: observers, occasions and settings. As a result of these considerations, we question the extent to which one generalizability study, at least in this area of research, can fulfill all the requirements of observer agreement, validity, and reliability which we view as so important. Rather, it is likely that multiple analyses will still be necessary to sufficiently establish all of the methodological requirements we have outlined for naturalistic observational data. These multiple analyses may, of course, involve analyses of variance in a generalizability model or correlational analyses as traditionally employed.

Krantz (1971) points out that the basic controversy over group vs. individual subject designs has contributed largely to the development of the mutual isolation of operant and nonoperant psychology. Since the measurement of reliability and convergent validity is typically based on correlations across a group of subjects, the operant psychologist may feel that these are alien concepts which have no relevance for his research. We would dispute

this view on the following logical grounds. Reliability involves the requirement for consistency in measurement and without some minimal level of such consistency, there can be no demonstration of functional relationships between the dependent variable and the independent variable. Efforts are currently underway to discover statistical procedures for establishing reliability estimates for the single case (e.g., see Jones, 1972). Any operant study which involves repeating manipulative procedures on more than one subject can be used for reliability assessment by traditional methods. Once such reliability is established, either for the individual case or for a group, we can be much more confident in the data and its meaning. Validity involves the requirement of convergence among different methods in measuring the same behavioral dimension. Where the validity of a measurement procedure has been previously established for a group, we can use it with more confidence in each individual case. Where it has not, it is still possible to explore for convergence in a single case. We can simply see, for example, if the child who shows high rates of aggressive behavior is perceived as aggressive by significant others. This procedure may be done with some precision if normative data is available on the measures used in the single case. Thus, with normative data available one can explore the position of the single case on the distribution of each measurement instrument. One could see, for example, if the child who is perceived to be among the top 5% in aggressiveness actually shows aggressive behavior at a rate higher than 95% of his peers. The requirements of reliability and validity are logically sound ones which transcent experimental method and means of calculation.

These methodological issues, like all others presented in this chapter, are highly relevant for behavioral research, even though they may at first

seem alien to it as the products of rival schools of thought. It has been our argument that the requirements of sound methodology transcend "schools" and that the time has come for us to attend to any variables which threaten the quality, generalizability, or meaningfulness of our data. Behavioral data is the most central commonality and critical contribution of all behavior modification research. The behaviorists' contribution to the science of human behavior and to solutions of human problems will largely rest on the quality of this data base.

References

Adkins, D. A., & Johnson, S. M. An empirical approach to identifying deviant behavior in children. Paper presented at the Western Psychological Association Convention, Portland, Oregon, April 1972.

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.

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

Bales, D. F. Interaction Process Analysis. Cambridge: Addison-Wesley, 1950.

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

Barker, R. G. & Wright, H. F. Midwest and its children: The psychological ecology of an American town. New York: Row, Peterson, 1955.

Bechtel, R. B. The study of man: Human movement and architecture. Trans-Action, 1967, 4 (6), 53-56.

Berberich, J. Adult child interactions: I. Correctness of a "child" as a positive reinforcer for the behavior of adults. Unpublished manuscript, University of Washington, 1970.

Bernal, M. E., Gibson, D. M., William, D. E., & Pesses, D. I. A device for recording automatic audio tape recording. Journal of Applied Behavior Analysis, 1971, 4 (2), 151-156.

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

Bolstad, O. D., & Johnson, S. M. Self-regulation in the modification of disruptive classroom behavior. Journal of Applied Behavior Analysis, 1972, in press.

Browning, R. M., & Stover, D. O. Behavior modification in child treatment: An experimental and clinical approach. Chicago: Aldine Atherton, Inc., 1971.

Campbell, D. T., & Fiske, D. Convergent and discriminant validation by the multi-trait, multi-method matrix. Psychological Bulletin, 1959, 56, 81-105.

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

Chapanis, A. The relevance of laboratory studies to preschool situations. Ergonomics, 1967, 10, 557-577.

Clairborn, W. L. Expectancy effects in the classroom: A failure to replicate. Journal of Educational Psychology, 1969, 60, 377-383.

Cobb, J. A. The relationship of observable classroom behaviors to achievement of fourth grade pupils. Unpublished doctoral dissertation, University of Oregon, Eugene, 1969.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generalizability: Liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

DeMaster, B., & Reid, J. B. Effects of feedback procedures in maintaining observer reliability. In preparation, Oregon Research Institute, Eugene, 1972.

D'Heurle, A., Mellinger, J. C., & Haggard, E. A. Personality, intellectual, and achievement patterns in gifted children. Psychological Monographs: General and Applied, 1959, 73 (13, Whole No. 483).

Eyberg, S. An outcome study of child-family intervention: The effects of contingency contracting and order of treated problems. Unpublished doctoral dissertation, University of Oregon, Eugene, 1972.

Fawl, C. Disturbances experienced by children in their natural habitats. In R. G. Barker (Ed.), The stream of behavior. New York: Appleton-Century-Crofts, 1963. Pp. 99-126.

Fode, K. L. The effect of experimenters' and subjects' anxiety and social desirability on experimenter outcome bias. Unpublished doctoral dissertation, University of North Dakota, 1965.

Fode, K. L. The effect of non-visual and non-verbal interaction on experimenter bias. Unpublished master's thesis, University of North Dakota, 1960.

Gelfand, D. M., Gelfand, S., & Dobson, W. R. Unprogrammed reinforcement of patients' behavior in a mental hospital. Behavior Research and Therapy, 1967, 5, 201-207.

Glaser, G. C., Cronbach, L. J., & Rajaratnam, N. Generalizability of scores influenced by multiple sources of variance. Psychometrika, 1965, 30, 395-418.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Gump, R., & Kounin, J. Issues raised by ecological and "classic" research efforts. Merrill-Palmer Quarterly, 1960, 6, 145-152.

Harris, A. Observer effect on family interaction. Unpublished doctoral dissertation, University of Oregon, Eugene, 1969.

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

Horton, G. O., Larson, J. L., & Maser, A. L. The generalized reduction of student teacher disapproval behavior. Unpublished manuscript,

University of Oregon, Eugene, 1972.

Hughes, L. D. A study of the relationship of coping strength to self-control, school achievement, and general anxiety level in sixth grade pupils. Dissertation Abstracts, 1968, 28 (A) (10), 4001.

Johansson, S., Johnson, S. M., Martin, S., & Wahl, G. Compliance and non-compliance in young children: A behavioral analysis. Unpublished manuscript, University of Oregon, Eugene, 1971.

Johnson, S. M., & Brown, R. Producing behavior change in parents of disturbed children. Journal of Child Psychology and Psychiatry, 1969, 10, 107-121.

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

Johnson, S. M., Wahl, G., Martin, S., & Johansson, S. How deviant is the normal child: A behavioral analysis of the preschool child and his family. Advances in Behavior Therapy, 1972, in press.

Jones, R. R. Intraindividual stability of behavioral observations: Implications for evaluating behavior modification treatment programs. Paper presented at the meeting of the Western Psychological Association, Portland, Oregon, April 1972.

Karpowitz, D. Stimulus control in family interaction sequences as observed in the naturalistic setting of the home. Unpublished doctoral dissertation, 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 entitled "Behavior Analysis in Education," University of Kansas, Lawrence, April 1970.

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

Kent, R. The human observer: An imperfect cumulative recorder. Paper presented at the Banff Conference on Behavior Modification, Banff, Canada, March 1972.

Krantz, D. L. The separate worlds of operant and non-operant psychology. Journal of Applied Behavior Analysis, 1971, 4 (1), 61-70.

Lahaderne, H. M. Attitudinal and intellectual correlates of attention: A study of four sixth-grade classrooms. Journal of Educational Psychology, 1968, 59 (5), 320-324.

Littman, R., Pierce-Jones, J., & Stern. T. Child-parent activities in the natural setting of the home: results of a methodological pilot study. Unpublished manuscript, University of Oregon, Eugene, 1957.

Lobitz, G., & Johnson, S. M. Normal versus deviant--Fact or fantasy? Paper presented at the Western Psychological Association Convention, Portland, April 1972.

Martin, M. F., Gelfand, D. M., & Hartmann, D. P. Effects of adult and peer observers on boys' and girls' responses to an aggressive model. Child Development, 1971, 42, 1271-1275.

Martin, S. The comparability of behavioral data in laboratory and natural settings. Unpublished doctoral dissertation, University of Oregon, Eugene, 1971.

Meyers, C. E., Attwell, A. A., & Orpet, R. E. Prediction of fifth grade achievement from kindergarten test and rating data. Educational and Psychological Measurement, 1968, 28 (2), 457-463.

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

O'Conner, R. D. Modification of social withdrawal through symbolic modeling. Journal of Applied Behavior Analysis, 1969, 2, 15-22.

Olson, W. The incidence of nervous habits in children. Journal of Abnormal and Social Psychology, 1930-31, 35, 75-92.

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

Patterson, G. R. & Cobb, J. A. Stimulus control for classes of noxious behaviors. Paper presented at the University of Iowa, May 1971, Symposium, "The control of aggression: Implications from basic research." J. F. Knutson (Ed.). Aldine Publishing, 1972, in press.

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

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

Patterson, G. R., Ray, R. S., & Shaw, D. A. Direct intervention in families of deviant children. Oregon Research Bulletin, 1969, 8.

Patterson, G. R., Ray, R. S., Shaw, D. A., & Cobb, J. A. Manual for coding family interactions, sixth revision, 1969. Available from ASIS National Auxiliary Publications Service, in care of CCM Information Service, Inc., 909 Third Avenue, New York, N. Y. 10022. Document #01234.

Patterson, G. R., & Reid, J. B. Reciprocity and coercion: Two facets of social systems. In C. Neuringer & J. Michael (Eds.), Behavior modification in clinical psychology. New York: Appleton-Century Crofts, 1970.

Patterson, G. R., & Reid, J. B. Family intervention in the homes of aggressive boys: A replication. Paper presented at the American Psychological Association Convention, Washington D. C., 1971.

Peine, H. A. Behavioral recording by parents and its resultant consequences. Unpublished master's thesis, University of Utah, 1970.

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

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

Purcell, K., & Brady, K. Adaptation to the invasion of privacy: Monitoring behavior with a miniature radio transmitter. Merrill-Palmer Quarterly, 1965, 12, 242-254.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. Generalizability of stratified-parallel tests. Psychometrika, 1965, 30, 39-56.

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

Rausch, H. L. Interaction sequences. Journal of Personality and Social Therapy, 1965, 2, 487-499.

Reid, J. B. Reciprocity in family interaction. Unpublished Doctoral Dissertation, University of Oregon, Eugene, 1967.

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

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

Romanczyk, R. G., Kint, R. W., 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, Lawrence, Kansas, May 1971.

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

Rosenthal, R. On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. American Scientist, 1963, 51, 268-283.

Rosenthal, R., & Fode, K. L. The effect of experimenter bias on the performance of the albino rat. Behavior Science, 1963, 8, 183-189.

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

Rosenthal, R., & Jacobsen, L. Teacher's expectancies: Determinants of pupils IQ gains. Psychological Reports, 1966, 19, 115-118.

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.

Rosenthal, R., Persinger, G. W., Vikan-Kline, L. E., & Mulry, R. C. The role of the research assistant in the mediation of experimenter bias. Journal of Personality, 1963, 31, 313-335.

Rothbart, M., & Maccoby, E. Parents' differential reaction to sons and daughters. Journal of Personality and Social Psychology, 1966, 4, 237-243.

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

Goldman and Borstad

Sheffield, F. D., Kaufman, H. ., Rhine, . . A Pr experiment [illegible] starts a controversy. Journal of American Social and Experimental Research, 1952, 46, 111-1. .

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

Smith, E. E. Effects of threat induced by ambiguous role expectations on defensiveness and productivity in small groups. Dissertation Abstracts, 1957, 17, 3104-3105.

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

Soskin, W. F., & John, V. P. The study of spontaneous talk. In R. G. Barker (Ed.), The stream of behavior. New York: Appleton-Century-Crofts, 1963. Pp. 228-281.

Taplin, P. ., & Reid, J. B. Effects of instructional set and experimental influence on observer reliability. In Preparation, Oregon Research Institute, Eugene, 1972.

Thomas, D. R., Becker, W. C., Armstrong, M. Production and elimination of disruptive classroom behavior by systematically varying teacher's behavior. Journal of Applied Behavior Analysis, 1968, 1, 35-45.

Thorndike, R. L. Pygmalion in the classroom: A review. Teacher College Record, 1963, 70, 805-807.

Verplanck, N. S. The control and the content in conversation: Reinforcement of statements of opinion. Journal of Abnormal and Social Psychology, 1955, 51, 668-676.

Wahl, G., Johnson, S. M., Martin, ..., & Johnson, . An operant analysis of child-family interaction. Behavior Therapy, 1972, in press.

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

Walker, H. M., Johnson, S. M., & Hops, H. Generalization and maintenance of classroom treatment effects. Unpublished manuscript, University of Oregon, Eugene, 1972.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. Unobtrusive measures: A survey of non-reactive research in the social sciences. Chicago: Rand McNally, 1966.

Weick, K. E. Systematic observational methods. In G. Lindsey & E. Aronson (Eds.), The handbook of social psychology, 1968, 2, 357-451.

White, G. D. Effects of observer presence on mother and child behavior. Unpublished doctoral dissertation, University of Oregon, September 1972.

Wiggins, J. S. Personality and prediction: Principles of personality assessment. Reading, Mass.: Addison Wesley, 1972, in press.

Footnotes

1. The preparation of this manuscript and the research reported therein was supported by research grant MH 19633-01 from the National Institute of Mental Health. The writers would like to thank their many colleagues who contributed critical reviews of this manuscript: Robyn Dawes, Lewis Goldberg, Richard Jones, Gerald Patterson, John Reid, Carl Skindrud and Geoffry White.

2. The authors would like to credit Lee Sechrest for first suggesting this illustrative example.

3. The authors would like to credit Donald Hartman for clarifying this as the appropriate procedure for establishing the level of agreement to be expected by chance.

4. For additional justification of the use of this statistical procedure for problems of this kind, see Wiggins (1972).

5. Several behaviors which are used in the coding system are not included in the present analysis. The behaviors humiliate and dependency could not be included because they did not occur in the behavioral sample. Repeated noncompliance and temper tantrums were not used on the verbal report scale because they are subsumed in other categories (i.e., tantrums are defined as simultaneous occurrences of three or more of the following--physical negative, destructiveness, crying, yelling, etc.). Nonresponding of the child was excluded post hoc because it was clear that parents were responding to this item as ignoring rather than mere nonresponse to ongoing activity (i.e., it was a poorly-written item).

Table 1

Coded Behaviors as Ranked by Two Methods:

Parental Ratings and Negative Social Consequences*

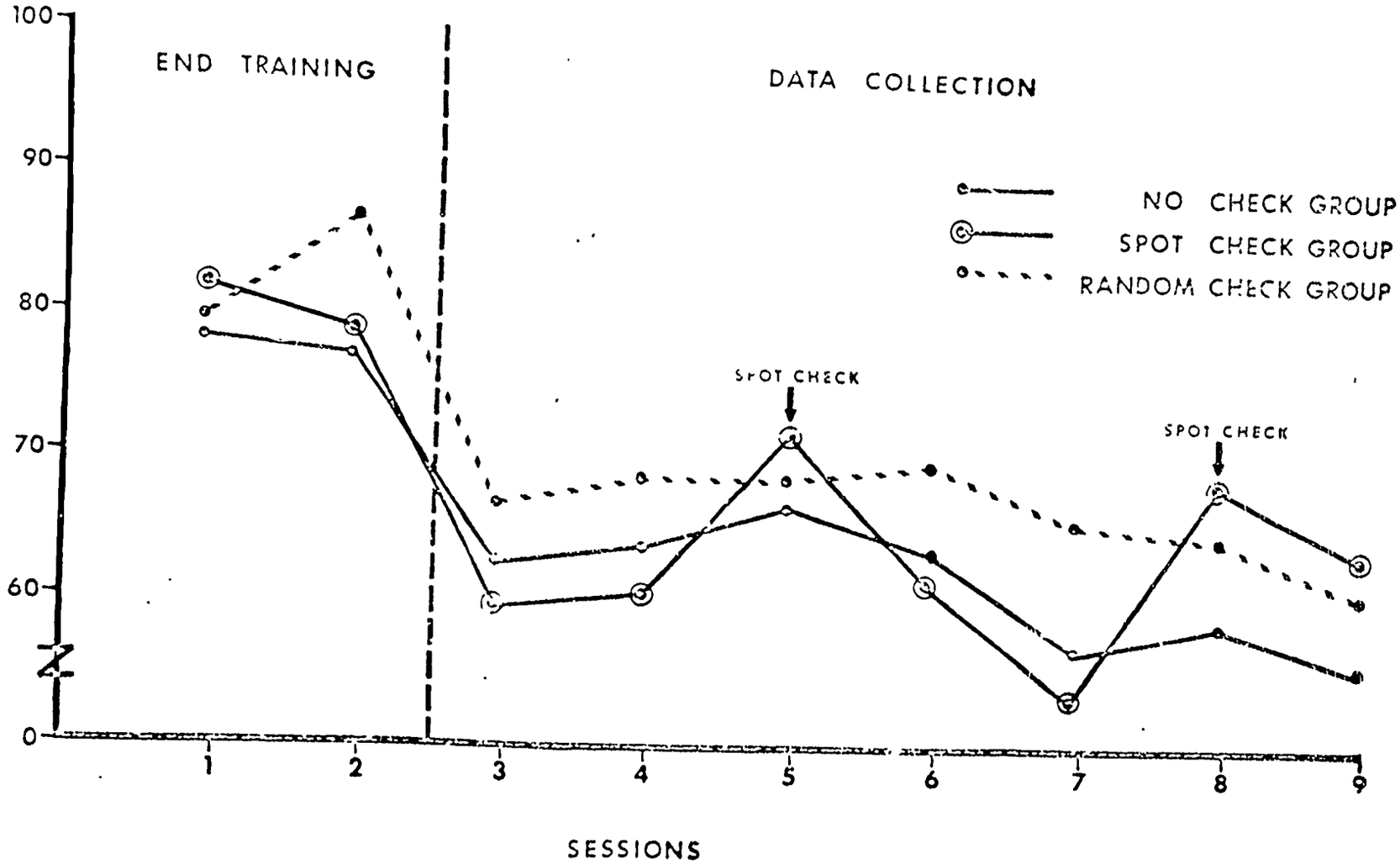
Behavior Rank by Parental Rating		Behavior Rank by Proportion of Negative Consequences	Mean Parent Rating for Behavior	Proportion of Negative Consequences to Behavior
1	Whine	13	1.056	.125
2	Physical Negative	2	1.074	.527
4	Destructive	8	1.204	.352
4	Tease	5	1.204	.382
4	Smart Talk	4	1.204	.390
6	Aversive Command	3	1.208	.428
7	Noncompliance	12	1.278	.175
8	High Rate	16	1.307	.064
9	Ignore	11	1.370	.205
10	Yell	10	1.537	.215
11	Demand Attention	15	1.611	.083
12	Negativism	6	1.685	.375
13	Command Negative	1	1.833	.569
14	Disapproval	9	1.870	.235
15	Cry	14	1.962	.097
16	Indulgence	22	2.093	.027
17	Command Prim	27.5	2.132	.000
18	Receive	18	2.222	.050
19	Talk	23	2.278	.020
20	Command	7	2.296	.355
21	Attention	25	2.556	.013
22	Touch	20	2.648	.043
23	Independent Activity	26	2.704	.005
24	Physical Positive	21	2.741	.034
25	Comply	17	2.759	.053
26	Laugh	19	2.778	.044
27	Nonverbal Interaction	24	2.833	.012
28	Approval	27.5	2.926	.000

*Spearman Rank-order correlation between columns 1 & 2 = .73 ($p < .01$).

Figure 1

[illegible] with [illegible] [illegible] [illegible] [illegible] [illegible] sessions

([illegible] and Bolstad, [illegible], 1973)



Johnson and Bolstad

87

DOCUMENT RESUME

ED 071 749 PS 006 291

AUTHOR	Johnson, Stephen M.; Bolstad, Orin D.
TITLE	Methodological Issues in Naturalistic Observation: Some Problems and Solutions for Field Research. Final Report.
SPONS AGENCY	National Inst. of Mental Health (DHEW), Bethesda, Md.
PUB DATE	Mar 72
NOTE	87p.; Presented at the Banff International Conference on Behavior Modification (4th, March 1972)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Behavioral Science Research; Children; Classification; *Data Analysis; *Family Life; Observation; Problem Solving; *Psychology; *Research Methodology; Standards

ABSTRACT

An attempt at defining and describing those factors which most often jeopardize the validity of naturalistic behavioral data is presented. A number of investigations from many laboratories which demonstrate these methodological problems are reviewed. Next, suggestions, implementations, and testing of effectiveness of various solutions to these dilemmas of methodology are steps taken. Research in the paper involves the observation of both "normal" and "deviant" children and families in the home setting. The observation system employed is a modified form of the code devised by Patterson, Ray, Shaw, and Cobb (1969). The observations are made under certain restrictive conditions: (1) All family members must be present in two adjoining rooms; (2) No interactions with the observer are permitted; (3) The television set may not be on; and (4) No visitors or extended telephone calls are permitted. Other later studies are also reviewed in this paper. (CK)

METHODOLOGICAL ISSUES IN NATURALISTIC OBSERVATION:

SOME PROBLEMS AND SOLUTIONS FOR FIELD RESEARCH

FINAL REPORT

by

Stephen M. Johnson and Orin D. Bolstad

University of Oregon

Presented at
The Fourth Banff International Conference on Behavior Modification
in March 1972.
To be published in the proceedings of the Banff Conference,
Research Press, 1972, in press.

METHODOLOGICAL ISSUES IN NATURALISTIC OBSERVATION: SOME PROBLEMS AND SOLUTIONS FOR FIELD RESEARCH[1]

Stephen M. Johnson and Orin D. Bolstad

University of Oregon

Encapsulated schools of thought have occurred in all sciences at some stage in their development. They appear most frequently during periods where the fundamental assumptions of the science are in question. Manifesto papers, acrimonious controversy, mutual rejection, and isolation of other schools' strategies are hallmarks of such episodes [David L. Krantz, The separate worlds of operant and non-operant psychology. Journal of Applied Behavior Analysis, 1971, 4 (1), p. 61].

History may well reveal that the greatest contribution of behavior modification to the treatment of human problems came with its emphasis on the collection of behavioral data in natural settings. The growth of the field will surely continue to produce greater refinement and proliferation of specific behavior change procedures, but the critical standard for assessing their utility will very likely remain the same. We will always want to know how a given procedure affects the subject's relevant behavior in his "real" world.

If a behaviorist wants to convince someone of the correctness of his approach to treating human problems, he is generally much less likely to rely on logic, authority, or personal testimonials to persuade than are proponents of other schools of psychotherapeutic thought. Rather, it is most likely that he will show his behavioral data with the intimation that this data speaks eloquently for itself. Because he is aware of the research on the low level of generalizability of behavior across settings (e.g., see Mischel, 1968), he is likely to be more confident in this data as it becomes more naturalistic in character (i.e., as it reflects naturally occurring behavior in the subject's usual habitat). As a perusal of the behavior modification literature will indicate, these data are often

extremely persuasive. Yet, the apparent success of behavior modification and the enthusiasm that this success breeds may cause all of us to take an uncritical approach in evaluating the quality of that data on which the claims of success are based. A critical review of the naturalistic data in behavior modification research will reveal that most of it is gathered under circumstances in which a host of confounding influences can operate to yield invalid results. The observers employed are usually aware of the nature, purpose and expected results of the observation. The observed are also usually aware of being watched and often they also know the purpose and expected outcome of the observation. The procedures for gathering and computing data on observer agreement or accuracy are inappropriate or irrelevant to the purposes of the investigation. There is almost never an indication of the reliability of the dependent variable under study, and rarely is there any systematic data on the convergent validity of the dependent measure(s). Thus, by the standards employed in some other areas of psychological research, it can be charged that much behavior modification research data is subject to observer bias, observee reactivity, fakability, demand characteristics, response sets, and decay in instrumentation. In addition, the accuracy, reliability and validity of the data used is often unknown or inadequately established.

But, the purpose of this paper is not to catalogue our mistakes or to argue for the rejection of all but the purest data. If that were the case, we would probably have to conclude with that depressing note which makes so many treatises on methodology so discouraging. Although dressed in more technical language, this purist view often expresses itself as: "You can't get there from here." We can get there, but it's not quite as

simple as perhaps we were first led to believe. The first step in getting there is to define and describe those factors which most often jeopardize the validity of naturalistic behavioral data. To this end, we will review a host of investigations from many laboratories which demonstrate these methodological problems. The second step is more constructive in nature: to suggest, implement, and test the effectiveness of various solutions to these dilemmas of methodology. Because behavioral data has become the primary basis for our approach to diagnosing and treating human problems, the endeavor to improve methodology is perhaps our most critical task for strengthening our contribution to the science of human behavior.

We will argue that the same kinds of methodological considerations which are relevant in other areas of psychology are equally pertinent for behavioral research. At least with respect to the requirements of sound methodology, the time of isolation of behavioral psychology from other areas of the discipline should quickly come to an end.

Throughout this paper, we will rely heavily on the experience of our own research group in meeting, or at least attenuating, these problems. We take this approach to illustrate the problems and their possible solutions more precisely and concretely. Most of our solutions are far from perfect or final, but it is our hope that a report based on real experience and data may be more meaningful than hypothetical solutions which remain untested. Thus, before beginning on the outline of methodological problems and their respective solutions, it will be necessary for the reader to have a general understanding of the purposes and procedures of our research. This research involves the observation of both "normal" and "deviant" children and families in the home setting. The observation

system employed is a modified form of the code devised by Patterson, Ray, Shaw, and Cobb (1969). This revised system utilizes 35 distinct behavior categories to record all of the behaviors of the target child and all behaviors of other family members as they interact with this child. The system is designed for rapid sequential recording of the child's behavior, the responses of family members, the child's ensuing response, etc. Observations are typically done for forty-five minutes per evening during the pre-dinner hour for five consecutive week nights. The observations are made under certain restrictive conditions: a) All family members must be present in two adjoining rooms; b) No interactions with the observer are permitted; c) The television set may not be on; and, d) No visitors or extended telephone calls are permitted. Obviously, this represents a modified naturalistic situation.

On the average, these procedures yield the recording of between 1,800 and 1,900 responses and an approximately equal number of responses of other family agents over this time period of 3 hours and 45 minutes. This data is collected in connection with a number of interrelated projects. These include normative research investigations of the "normal" child (e.g., Johnson, Wahl, Martin & Johansson, 1972); research involving a behavioral analysis of the child and his family (e.g., Wahl, Johnson, Martin & Johansson, 1972; Karpowitz, 1972; Johansson, Johnson, Martin, & Wahl, 1971); outcome research on the effects of behavior modification intervention in families (Eyberg, 1972); comparisons of "normal" and "deviant" child populations (Lobitz & Johnson, 1972); and studies of methodological problems (Johnson & Lobitz, 1972; Adkins & Johnson, 1972; Martin, 1971). These latter studies will be reviewed in detail in the body of this paper. More recently, we have begun

to investigate the generality of children's behavior across school and home settings, and to document the level of generalization of the effects of behavior modification in one setting to behavior in other settings (Walker, Johnson, & Hops, 1972). Research is also in progress to relate naturalistic behavioral data to parental attitudes and behavioral data obtained in more artificial laboratory settings. With all of these objectives in mind, it is most critical that the behavioral data collected is as valid as possible and it is to this end that we explore the complex problems of methodology presented here.

Observer Agreement and Accuracy I: Problems of Calculation and Inference

The most widely recognized requirement of research involving behavioral observations is the establishment of the accuracy of the observers. This is typically done by some form of calculation of agreement between two or more observers in the field. Occasionally, observers are tested for accuracy by comparing their coding of video or audio tape with some previously established criterion coding of the recorded behavior. For convenience, we will refer to the former procedure as calculation of observer agreement and the latter as calculation of observer accuracy. In general, both of these procedures have been labeled observer reliability. We will eschew this terminology because it tends to confuse this simple requirement for observer agreement or accuracy with the concept of the reliability of a test as understood in traditional test theory. As we shall outline in section three, it is quite possible to have perfect observer agreement or accuracy on a given behavioral score with absolutely no reliability or consistency of measurement in the traditional sense. Generally, the classic reliability requirement involves

a demand for consistency in the measurement instrument over time (e.g., test-retest reliability) or over-sampled item sets responded to at roughly the same time (e.g., split-half reliability). An example may help clarify this point. If two computers score the same MMPI protocol identically, there is perfect "observer agreement" but this in no way means that the MMPI is a reliable test which yields consistent scores.[2] Although the question of reliability as traditionally understood has been largely ignored in behavioral research, we will argue in section three that it is a critical methodological requirement which should be clearly distinguished from observer agreement and accuracy.

There is no one established way to assess observer agreement or accuracy and that is as it should be, because the index must be tailored to suit the purposes of each individual investigation. There are three basic decisions which must be made in calculating observer agreement. The first decision involves the stipulation of the unit score on which the index of agreement should be assessed. In other words, what is the dependent variable for which an index of accuracy is required as measured by agreement with other observers or with a criterion? An example from our own research may help clarify this point. We obtain a "total deviant behavior score" for each of the children we observe. This score is based on the sum output of 15 behaviors judged to be deviant in nature. An outline of the rationale and validity of this score will be given in a later section. Suffice it to say, whenever two observers watch the same child for a given period, they each come up with their own deviant behavior score. These scores may then be compared for agreement on overall frequency. It is obvious that the same deviant behaviors need not be observed to get high

indexes of agreement on the total number of deviant behaviors observed. Yet, for many of our purposes, this is not important, since we merely want an index of the overall output of deviant behavior over a given period. The same procedure is, of course, applicable to one behavior only, chains of behavior, etc. The point is that the researcher must decide what unit is of interest to him for his purposes and then compare agreement data on that variable. In complex coding systems, like the one used in our laboratory, it has been customary to get an overall percent agreement figure which reflects the average level of agreement within small time blocks (e.g., 6-10 seconds) over all codes. In general, we would argue that this kind of observer agreement data is relatively meaningless. It has limited meaning because it is based on a combination of codes, some of which are observed with high consensus and some which are not. Furthermore, the figure tends to overweight those high rate behaviors which are usually observed with greater accuracy and underweight those low frequency behaviors which are usually observed with less accuracy. Patterson (personal communication) has reported that the observer agreement on a code correlates .49 with its frequency of use. Since it is often the low base rate behaviors which are of most interest to researchers, this overall index of observer agreement probably overestimates the actual agreement on those variables of most concern.

The second question to be faced involves the time span within which common coding is to be counted as an agreement. For most purposes of our current research, score agreement over the entire 225 minutes of observation is adequate. Thus, when we compute the total deviant behavior score over this period, we do not know that each observer sees the same deviant

behavior at the same time. But, good agreement on the overall score tells us that we have a consensually validated estimate of the child's overall deviancy. For some research purposes, this broad time span for agreement would be totally inadequate. For conditional probability analysis of one behavior (cf. Patterson & Cobb, 1971), for example, one needs to know that two observers saw the same behavior at the same time and (depending on the question) that each observer also saw the same set or chain of antecedents and/or consequences. This latter criterion is extremely stringent, particularly with complex codes where low rate behaviors are involved, but these criteria are necessary for an appropriate accuracy estimate.

Once one has decided on the score to be analyzed and the temporal rules for obtaining this score, one must then face the problem of what to do with these scores to give a numerical index of agreement. The two most common methods of analysis are percent agreement and some form of correlational analysis over the two sets of values. Both methods may, of course, be used for observer agreement calculation within one subject or across a group of subjects. Once again, neither method is always appropriate for every problem and each has its advantages and disadvantages. The most common way of calculating observer agreement involves the following simple formula:

$$\frac{\text{number of agreements}}{\text{number of agreements + disagreements}}$$

What is defined as an agreement or disagreement has already been solved if one has decided on the "score" to be calibrated and the time span involved.

Use of this formula implies, however, that one must be able to discriminate the occurrence of both agreements and disagreements. This can

only be accomplished precisely when the time span covered is relatively small (e.g., 1-15 seconds) so that one can be reasonably sure that two observers agreed or disagreed on the same coding unit. It has been common practice for investigators to compare recorded occurrences of behavior units over much longer time periods and obtain a percent agreement figure between two observers which reflects the following:

$$\frac{\text{smaller number of observed occurrences}}{\text{larger number of observed occurrences}}$$

The present authors would view this as an inappropriate procedure because there is no necessary "agreement" implied by the resulting percent. If one observer sees 10 occurrences of a behavior over a 30-minute period and the other sees 12, there is no assurance that they were ever in agreement. The behavior could have occurred 22 or more times and there could be absolutely no agreement on specific events. The two observers did not necessarily agree 84% of the time. Data of this kind can be more appropriately analyzed by correlational methods if such analysis is consistent with the way in which the data is employed for the question under study. Although the same basic problem mentioned above can, of course, occur, the correlational method is viewed as more appropriate because; a) The correlation is computed over an array of subjects or observation time segments and b) The correlation reflects the level of agreement on the total obtained score and it does not imply any agreement on specific events.

Whenever using the appropriate method of calculating observer agreement percent, (i.e: $\frac{\text{number of agreements}}{\text{number of agreements + disagreements}}$) the investigator should be particularly cognizant of the base rate problem. That is, the obtained percent agreement figure should be compared with the amount of agreement that could be obtained by chance. An example will clarify this point. Suppose two coders are coding on a binary behavior coding system (e.g., appropriate vs. inappropriate behavior). For the sake of illustration, let us suppose that observers have to characterize the subject's behavior as either appropriate or inappropriate every five seconds. Now, let us suppose, as is usually the case, that most of the subject's behavior is appropriate. If the subject's behavior were appropriate 90% of the time,

two observers coding randomly at these base rates (i.e., .90-.10) will obtain 82% agreement by chance alone. Chance agreement is computed by squaring the base rate of each code category and summing these values.[3] In this simple case, the mathematics would be as follows: $.90^2 + .10^2 = .82$. The same procedure may, of course, be used with multi-code systems.

The above .90-.10 split problem may be reconceptualized as one in which the occurrence or nonoccurrence of inappropriate behavior is coded every five seconds. If, for purposes of computing observer agreement, we look at only those blocks in which at least one of two observers coded the occurrence of inappropriate behavior, the chance level agreement is drastically reduced. The probability that two observers would code occurrence in the same block by chance is only $.10^2$ or one percent. It would not be theoretically inappropriate to count agreement on nonoccurrence but, in the present example and in most cases, this procedure is associated with relatively high levels of chance agreement.

Whenever percent agreement data is reported, the base rate chance agreement should also be reported and the difference noted. Statistical tests of that difference can, of course, be computed. As long as the base rate data is reported, the percent agreement figure would always seem to be appropriate. For obvious reasons, however, it becomes less satisfactory as the chance agreement figure approaches 1.0.

The other common method of computing agreement data is by means of a correlation between two sets of observations. The values may be scores from a group of subjects or scores from n observation segments on one subject. This method is particularly useful when one is faced with the high chance agreement problem or where the requirement of simple similarity in ordering subjects on the dependent variable is sufficient for the research. As we shall illustrate, the

correlation is also particularly useful in cases where one has a limited sample of observer agreement data relative to the total amount of observation data. In general, correlations have been used with data scores based on relatively large time samples. In other words, they

tend to be used for summary scores on individuals over periods of 10 minutes to 24 hours. There is no reason why correlation methodology could not be applied to data from smaller time segments (e.g., 5 seconds), but this has rarely been done. So, studies using correlation methods have generally been those in which one cannot be sure that the same behaviors are being jointly observed at the same time. In using correlation methods for estimating agreement, one should be aware of two phenomena. First, it is possible to obtain high coefficients of correlation when one observer consistently overestimates behavioral rates relative to the other observer. This difference can be rather large, but if it is consistently in one direction, the correlation can be quite high. For some purposes this problem would be of little consequence but for other purposes it could be of considerable importance. The data can be examined visually, or in other more systematic ways, to see to what extent this is the case. This problem can be virtually eliminated if one uses many observers and arranges for all of them to calibrate each other for agreement data. Under these circumstances, one will obtain a collection of regular observer figures and a list of mixed calibrator figures for correlation. This procedure should generally correct for systematic individual differences and make a consistent pattern as outlined above extremely unlikely. The second problem to be cognizant of in using correlations is that higher values become more possible as the range on the dependent variable becomes greater. This fact may lead to high indexes of agreement when observers are really quite discrepant with respect to the number of a given behavior they are observing. An illustration may clarify this point. Let us suppose we are observing rates of crying and whining behavior in preschool children over a five-hour period. Some

particularly "good" children may display these behaviors very little and, given a true occurrence score of 7, two observers may obtain scores of 5 and 10 on this behavior class. This would be only 50% agreement. Other children display these behaviors with moderate to very high frequency. For a child with high frequency, we may find our two observers giving us scores of 75 and 125 respectively. This would be equivalent to 60% agreement and, of course, represents a raw discrepancy of 50 occurrences. Yet, if these examples were repeated throughout the distribution of scores and if there were little overlap, a high correlation would be obtained. This would be even more true, of course, if one observer consistently overestimated the rates observed by the other. Yet, even this possibility does not necessarily jeopardize the utility of the method. It must merely be recognized, examined and its implication for the question under study evaluated. In our own research we want to catalogue the deviancy rates of normal children, compare them with deviant children, and observe changes in deviancy rates as a result of behavior modification training with parents. For these purposes, general agreement on levels of deviant responding is quite good enough.

In our research on the normal child, we have had 47 families of the total 77 families observed for the regular five-day period by an assigned observer. On one of these days an additional observer was sent to the family for the purpose of checking observer agreement. The correlation between the deviant behavior scores of the two observers was .80. But, in a purel statistical sense, this figure is an underestimate of what the agreement correlation would be for the full five days of observation. Since we are using a statistic based on five times as much data, we want to know the expected

observer agreement correlation for this extended period. Adding time to an observation period is analogous to adding items to a test. The problem we are faced with here is very similar to that dealt with by traditional test theorists who have sought, for example, to estimate the reliability of an entire test based on the reliability of some portion of the test. In our case, we want to know the expected correlation for the statistic based on five days when we have the correlation based on one day. The well-known Spearman-Brown formula (Guilford, 1954) may be applied to this end (as in Patterson, Cobb, & Ray, 1972; Patterson & Reid, 1970; Reid, 1967).[4]

$$r_{nn} = \frac{nr_{tt}}{1 + (n-1)r_{tt}}$$

where r_{tt} = reliability of the test of unit length

n = length of total test.

With the Spearman-Brown correction, the expected observer agreement correlation for the deviant behavior score is .95. This same procedure has also been applied to other statistics of particular interest in this research including: a) the proportion of the parent's generally "negative" responses (correct agreement = .97), b) the proportion of the parent's generally positive responses (corrected agreement = .98), c) the median agreement coefficient of the 29 behavior codes observed for five or more children (corrected agreement = .91), d) the median corrected agreement of the 11 out of 15 deviant behavior codes used (r = .91), e) the number of parental commands given (corrected agreement = .99), and f) the compliance ratio (i.e., compliances/compliances plus noncompliances) of the child (corrected agreement = .92). As our research is completed, we will be presenting observer agreement data using different statistics, computed in different ways, and evaluated by different criteria.

The primary point of this section is to indicate that there are many ways of calculating observer agreement data and there is no one "right way to do it." The methods differ on three basic dimensions: a) the nature and breadth of the dependent variable unit, b) the time span covered, and c) the method of computing the index. Each investigator must make his own decisions on each of these three points in line with the purposes of his investigation. But, the investigator should be guided by one central prescription--the agreement data should be computed on the score used as the dependent variable. It makes no sense to report overall average agreement data (except perhaps as a bow to tradition) when the dependent variable is "deviant behavior rate." In addition, it makes little sense to make the agreement criteria relative to time span more stringent than necessary. If the dependent variable is overall rate of deviant behavior for a five-day period, then this is the statistic for which agreement should be computed. It is not necessary for this limited purpose that both observers see the same deviant behavior in the same brief time block.

Before closing this section on the computation of observer agreement, we should address the somewhat unanswerable question of the minimum criteria for the acceptability of observer agreement data. In other words, how much agreement is sufficient for moving on to consider the results of a particular study. When using observer agreement percent, it would seem reasonable, at the very minimum, to show that the agreement percent is greater than that which could be expected by chance alone. When dealing with correlation data, one should at least show the obtained correlation to be statistically significant. These criteria are, of course, extremely minimal and certainly far below those criteria commonly used in traditional testing and measurement

to establish reliability (e.g., see Guilford, 1954). Yet, these criteria do provide a reasonable lowest level standard and there are some very good reasons why we should not be overly conservative on this point. In the first place, very complex codes, which may provide us with some of our most interesting findings, are very difficult to use with complete accuracy. On the basis of our experience, and that of G. R. Patterson (personal communication), we see an overall agreement percent of 80% to 85% as traditionally computed as a realistic upper limit for the kind of complex code we are using.

Furthermore, to the extent that less than perfect agreement represents only unsystematic error in the dependent variable, it cannot be considered a confounding variable accounting for positive results. Any positive finding which emerges in spite of a good deal of "noise" or error variance is probably a relatively strong effect.

Low observer agreement does, however, have very important implications for negative results. This gets us back to the fundamental principle that one can never prove the null hypothesis. The more error in the measurement instrument, the greater the chance for failing to discover important phenomena. Thus, just as with traditional test reliability, the lower the observer accuracy, the less confidence one can have in any negative findings from the research.

Observer Agreement and Accuracy II: Generalizability of Observer Agreement Data

All of the preceding discussion on the calculation of observer agreement data relies on the assumption that the obtained estimates of agreement are generalizable to the remainder of the observers' data collection.

In most naturalistic behavioral research, however, this assumption cannot go unchallenged and this brings us to our next, and largely [illegible] methodological problem. To illustrate this problem, let us take the not untypical case of an investigator who trains his observers on a behavioral code until they meet the criterion of two consecutive observation sessions at 80% agreement or better. After completing this training, the investigator embarks on his research with no further assessment of observer agreement. There are three basic problems with this methodology which make the generalizability of this agreement data extremely questionable. These problems are a) the nonrandomness of the selected data points, b) the unrepresentativeness of the selected data points in terms of the time of the assessment, and c) the potential for the observer's reactivity to being checked or watched. The first two problems may be rather easily solved in all naturalistic research, but the third problem represents quite a challenge to some forms of naturalistic observation. Let us explore these problems in more detail. The nonrandomness of selecting the last two "successful" observation sessions in a series for establishing a true estimate of agreement should be very obvious. It is not unlikely that, had the investigator obtained several additional agreement sessions, he would find the average agreement figure to be lower than 80%. It is quite possible that our observers had, by chance, two consecutive "good days" which are highly unrepresentative of the days to come. One can almost visualize our hypothetical investigator, after the first day of highly accurate observation, saying to his observers, "That was really a good one; all we need is one more good session and we can begin the study." But, now we are getting into problems two and three.

The second problem of unrepresentativeness in terms of time has previously been discussed by Campbell and Stanley (1966) and labeled instrument decay. That is, estimates of observer accuracy obtained one week may not be representative of observer accuracy the next week. The longer the research lasts, the greater is the potential problem of instrument decay. In the case of human observers, the decay may result from processes of forgetting, new learning, fatigue, etc. Thus, because of instrument decay, our investigator's estimate of 80% agreement is probably an exaggeration of the true agreement during the study itself. The problem of instrument decay is also often compounded by the fact that during observer training, there is usually a great deal of intense and concentrated work with the code, coupled with extensive training and feedback concerning observer accuracy. This intensity of experience and feedback is usually not maintained throughout the course of the research, and, as a result, the two time periods are characterized by very different sets of experiences for the observers. The third problem of generalizability of this agreement data involves the simple fact that people often do a better, or at least a different, job when they are aware of being watched as opposed to when they are not. Campbell and Stanley (1966) have labeled this problem reactive effects of testing. It is likely that, when observers are being "tested" for accuracy, they will have heightened motivation for accuracy and heightened vigilance for critical behaviors or for the coding peculiarities of their calibrator. This point has been brought home dramatically to us on more than one occasion by the tears of an observer after earning a particularly low agreement rating. Thus, because of the reactivity problem, estimates of observer agreement obtained with the awareness of the observer are likely to over-

estimate the true agreement level which would be obtained if the observer were not aware of such calibration.

Fortunately, all of the preceding logical arguments have been investigated in some recent research largely contributed by John Reid of the Oregon Research Institute. In his first published study on this problem, Reid (1970) designed a study which, from the observer's point of view, was almost identical to the hypothetical example given earlier. In this study, observers were trained for a maximum of eight days on complex behavioral coding of videotape material in a laboratory setting. They were trained until they obtained 70% agreement on two consecutive days for 15 minutes of observation per day. In each training session, feedback was given to the observers concerning their accuracy, but in the immediately following assessment session no such feedback was given. After training was completed, the observers were led to believe that they would be the only ones seeing the new tapes because criterion protocols had not been previously constructed and the tapes would have to be erased immediately after coding so that they might be reused. Observers were urged, therefore, to be very careful in their subsequent coding. On each of four consecutive days after reaching criterion, the observers were asked to complete 30 minutes of observation on similar stimulus materials. The first 5 minutes of each session served as a practice period in which the experimenter gave feedback to the observer when necessary. Although the observers were not aware of it, criterion protocols had been prepared on these videotape materials, thus providing a covert assessment of observer accuracy. Throughout the study, observer accuracy was computed by comparing the data entry with the standard protocol. Both errors of commission and omission were counted. This is the type of overall and generally stringent accuracy criteria which we

have earlier recommended against for most research purposes. In any event, all observers reached the 70% agreement criterion at or before eight days of training and the average agreement during the two criterion days was 75.50%. The average agreement during the following four days, however, was only 51.25%, representing a drop of nearly 25 percentage points. Now it is obvious that, in this first study, all three confounding problems were present--a) unrepresentativeness of the criterion days, b) instrument decay due to passage of time, etc., and c) reactivity to being tested. This is the procedure, however, which is representative of many investigations in the field. Reid's study certainly demonstrates the potentially powerful impact of these confounding influences. Because of the relatively short time span involved in this study, it is likely that most of this effect is attributable to the reactivity problem.

A more recent study by Taplin and Reid (in preparation) sheds greater light on the contribution of each of these confounding influences. Although some of the details of this latter study differed from the earlier investigation, the basic procedures and concepts involved were the same. After training to criterion of 70% agreement, the observers were assigned to one of three groups: a) No-check group--the procedures in this group essentially replicated those of the former study such that the observers were unaware of agreement assessment beyond the criterion period. b) Spot-check group--in this group, observers were told that their accuracy would be checked from time to time during the experimental phase, but that they would be aware of when these checks would be made. This group controls for the problem of unrepresentativeness of criterion day performance as well as instrument decay, but does not control for reactivity to testing. c) Random-check group--in this group, the observers were told that their

accuracy would be checked randomly in 20% of the sessions but that they would not be aware of when these checks occurred. This procedure not only controls for the first two problems, but also controls for the reactivity problem by making it constant. In other words, it is likely that the knowledge of a "potential" assessment of accuracy will increase the observer's motivation and vigilance. Yet, this is not a confounding influence since any random assessment of accuracy in the series will not be unrepresentative of the other points in the series. Thus, whenever practically possible, this strategy of random assessment without observer awareness or assessment on every observation occasion is most consistent with rigorous research requirements. The results of the Taplin and Reid (in preparation) study are presented in Figure 1 and are of particular interest here for the way in which they illustrate the operation of the three confounding

---
Insert Figure 1 about here
---

problems. The discrepancy between percent agreement on the criterion days and on the days of data collection is obvious and significant for all groups. Thus, the use of this kind of previously collected criterion data for generalization to subsequent data collection appears highly inappropriate, regardless of the method of data collection or the rationale provided observers during the data collection period. In all groups, the confounds of unrepresentativeness of criterion day data and instrument decay from the criterion days are obvious. Now, the observer accuracy data produced on the spot check days are not unrepresentative, in the sense that these days were presumably not chosen initially as particularly "good" ones, as were the criterion days. In addition, this procedure controls for the

problems of instrument decay due to passage of time since the check was randomly throughout the period of data collection. It is clear that the "spot check" accuracy assessment produced less overestimation of true accuracy than did the criterion day estimates. The data in this "spot check" group is very clear, however, in illustrating the problem of observer reactivity. The agreement percents in the two spot check periods are clearly overestimates of agreement percents obtained under covert assessment. Thus, the "spot check" method represents a decided improvement over the criterion score method, but it is an imperfect solution.

In some naturalistic settings, the "spot check" method may have to remain the optimal solution for reasons of cost and the potential for the reactive effects inherent in using multiple observers continuously (e.g., in observations in homes or other confined quarters). As previously stated, the methods of total accuracy assessment (e.g., as used by Wahler, 1967, and Browning & Stover, 1971) or random accuracy assessment without awareness (as in Taplin & Reid, in preparation) are always preferable when possible. These methods are, of course, particularly simple to apply with video or audio tape materials or in natural settings where two or more observers are, for whatever reason, employed simultaneously and continuously. In classrooms, for example, it is often the case that two or more observers record the behaviors of two or more children. Under these circumstances, the investigator can arrange the observers' recording schedules so that their observation of subjects overlap at random times. In this way, two observers can record the behavior of the same subject at the same time without either having knowledge of the ongoing calibration for agreement which is occurring at that specific time. This procedure would replicate the "random check

group" of Taplin and Reid (in preparation) in a field setting. It would probably be difficult, if not impossible, to keep the fact of random calibration a secret from the observers for any extended period, but, as stated earlier, this is no real problem, because the randomly collected data without specific awareness is representative of accuracy at other times. The Taplin and Reid (in preparation) data would suggest that the motivational effects of informing observers of the random checks slightly increases the level and stability of their accuracy scores. (Compare the three groups' accuracy level and stability in the data collection period in Figure 1.)

In more recent research, Reid and his colleagues have directed their efforts to finding ways of eliminating the instrument decay or "observer drift" observed in all previous studies regardless of the method of monitoring. In several long-term research projects, including our own (e.g., Johnson, Wahl, Martin & Johansson, 1972), the one directed by G. R. Patterson (e.g., Patterson, Cobb, & Ray, 1972) and the one reported by Browning and Stover (1971), continuous training, discussion of the coding system, and accuracy feedback are provided for the observers. It is possible that this kind of training and feedback could eliminate, or at least attenuate, observers' accuracy drift as well as the problem of the unrepresentativeness of "spot check" accuracy assessments. To test this hypothesis, DeMaster and Reid (in preparation) designed a study in which three levels of feedback and training during data collection were compared on a sample of 28 observers. The observers were divided into 14 pairs and all subsequent procedures were carried out in the context of these fixed pairs. The three experimental groups were as follows: Group I--Total Feedback--In this group observers a) discussed their observation performance together while reviewing their coding of the previous day's video tape, b) discussed their previous

day's observation with the experimenter in terms of their agreement with the criterion coded protocol, and c) received a daily report of their accuracy with respect to the criterion protocol. Group II--Pair Agreement Feedback--In this group, observers were given the opportunity to discuss their performance as in a above and b)were given a daily report on the extent to which each observer's coding protocol agreed with the protocol of the other observer. Subjects in this group were deprived of a discussion or report of their level of agreement with the criterion protocols. Group III--No Feedback--Subjects in this group were deprived of the kinds of feedback given in the previous two conditions and were instructed not to discuss their work among themselves to eliminate a possible "bias of the data." This group was similar in concept to the random-check group in the Taplin and Reid (in preparation) study in that they were told, as were all other subjects, that their accuracy would be checked at random intervals in the data collection period. The dependent variables were a) the agreement scores between pairs of observers and b) the "accuracy" scores reflected by the percent agreement with the criterion protocols. The results showed that the intra-pair observer agreement scores were significantly higher than were scores reflecting agreement with the criterion. These results tend to corroborate the hypothesis forwarded by Baer, Wolf, and Risley (1968) and Bijou, Peterson and Ault (1968) that high intra-pair agreement does not necessarily reflect proper use of the coding system. We shall call this problem "consensual observer drift." It is very important to note, however, that the design of this study which placed observers in fixed and unchanging pairs would tend to maximize this effect. In the field studies referred to above, observers typically meet in larger groups for training and feedback and observers rotate in calibrating each other's observations. Under these

circumstances, the effects of consensual drift would logically be expected to be less potent. Indeed, further data from the DeMaster and Reid (in preparation) study lends support to this argument. On those video-tape materials where more than one pair of observers had coded the sequence the investigators compared the fixed pair agreement with the agreement between observers in other pairs. In all cases, the fixed pair agreed more with one another than they did with the observers in the other pairs. Thus, this idiosyncratic drift of fixed pairs may be greater than drift experienced under currently employed field research procedures. Yet, a recent study by Romanczyk, Kent, Diament, and O'Leary (1971) showed that during overt agreement assessment observers would change their coding behavior to more closely approximate the differential coding styles of their calibrators. Thus, it is possible for observers to produce one kind of consensual drift with some calibrators and an opposite consensual drift with others to yield artifically high observer agreement data.

The manipulations in the Romanczyk et al. (1971) study were quite powerful, however, and one can question the generalizability of these artificially induced conditions to real field studies. Nevertheless, this study does demonstrate the potential for powerful and differential consensual drift. In spite of these considerations, one must realize that it is impossible in an ongoing field observation to have a "pure" criterion protocol, since one cannot arbitrarily designate one observer's protocol as the "true" criterion and the other as the imperfect approximate. But, one can attenuate this problem considerably by having frequent training sessions with observers on pre-coded video-tape material or on pre-coded behavioral scripts which may be acted out live by paid subjects. The importance of

this recommendation is underlined by DeMaster and Reid's (in preparation) second important finding. Analysis of the data indicated a significant main effect for feedback conditions, with the total feedback group doing best, followed by the intra-pair feedback group and the no feedback group, respectively.

It may be of interest to review briefly how our own project stacks up with regard to these considerations and to suggest ways in which it and similar projects might be improved in this area. Initial observer training in our laboratory consists of the following program: a) reading and study of the observation manual, b) completion of programmed instruction materials involving precoded interactions, c) participation in daily intensive training sessions which include discussion of the system and coding of precoded scripts which are acted out live by paid but nonprofessional actors, d) field training with a more experienced observer followed immediately by agreement checks. Currently, when an observer obtains five sessions with an average overall percent agreement of 70% or better, she may begin regular observation without constant monitoring. All observers continue to participate in continuous training and are subject to continuous checking with feedback. This is accomplished in two ways. First, each observer is subject to one spot-check calibration for each family she observes. This calibration may come on any one of the regular five days of observation. Both observers figure their percent agreement in the traditional way immediately after the session and discuss their disagreements at this time. If they cannot resolve their disagreement on a particular or idiosyncratic problem, they call the observer trainer immediately who serves as sort of an imperfect criterion coder. From time to time, idiosyncratic problems arise which cannot be resolved by the coding manual alone. Decisions on how to

code these special cases are made by the group and the trainer and are entered in a "decision log" which is periodically studied by all observers. These special circumstances are unfortunate and provide an opportunity for consensual drift, but are part of the reality with which we must deal. The "decision log" helps attenuate the drift problem on these decisions, and most of them tend to be idiosyncratic to one or two families. The second aspect of continual training involves a minimum of one 90-minute training session per week for all observers involving discussion and live coding experience. We have been negligent in our procedures in not retaining our precoded scripts over time and recoding these from month to month and year to year. On the basis of our review of Reid's excellent work, we have now begun to correct this error by retaining these scripts and subjecting them to recoding periodically to check the problem of "consensual observer drift." As will be obvious, we use the imperfect method of "spot check" calibration for observer agreement, but Reid's data is encouraging in that it indicates that the kind of intensive and continual training outlined here may attenuate the problems associated with this method. Furthermore, our observers are convinced that calibration scores obtained on a single day of observation are probably lower than would be obtained over two or more days of observation. The reason for this belief is that the calibrator would logically have more difficulty in adapting to each new home environment and identifying the subjects of observation on the first day in the home than on subsequent days. Unfortunately, we have no hard data to prove this hypothesis, but we have begun to do more than one day of calibration on families in order to test it.

The problem of consensual drift is also attenuated in this project by

the practice of having each observer calibrate all other observers. We recently began to employ only one calibrator for reasons of convenience and cost, but this review has persuaded us to return, at least partially, to multiple calibration among all observers.

As stated earlier, the problems associated with reactivity to testing for observer agreement could largely be solved by procedures which involved coding of audio or video tapes. This is true because one could arrange calibration on a random basis without observer awareness. Because procedures of this kind could also solve or attenuate problems of observer bias and subject reactivity, we are beginning to consider procedures of this type more seriously for future research and are now involved in pilot work on the feasibility of these methods. Short of this, we must be content with the "spot check" method as outlined and attempt to attenuate the problems associated with this method by use of extensive training and feedback as suggested by DeMaster and Reid (in preparation).

Reliability of Naturalistic Behavioral Data

One must look long and hard through the behavior modification literature to find even an example of reliability data on naturalistic behavior rate scores. In classical test theory, the concept of reliability involves the consistency with which a test measures a given attribute or yields a consistent score on a given dimension. Theoretically, a test of intelligence, for example, is reliable if it consistently yields highly similar scores for the same individual relative to other individuals in the sample. There are several approaches to measuring reliability including split-half measures, equivalent forms, test-retest methods, etc. Each method has a somewhat different meaning, but the basic objective of each is an estimate

of the consistency of measurement. It is difficult to tell whether behaviorists have simply neglected, or deliberately rejected, the reliability requirement for their own research. The concept comes out of classical test theory and is obviously allied to trait concepts of personality. Behaviorists may feel that the concept is irrelevant to their purposes. After all, we know that there is often very little proven consistency in human behavior over time and stimulus situations (e.g., see Mischel, 1968), so why should we require a consistency in our measurement instruments that is not present in real life? Behaviorists may feel that reliability is an outmoded concept and belongs exclusively to the era of trait psychology. If this is, in fact, the reason for the neglect of the reliability issue in behavioral research, it represents a serious conceptual error and a clear misapplication of the meaning of the data on the lack of behavioral consistency so eloquently summarized by Mischel (1968). It is true, of course, that behaviorists employ more restricted definitions of the topography of the relevant response dimensions (e.g., hitting vs. aggression) and that they often include more restrictive stimulus events in defining these dimensions (e.g., child noncompliance to mother's commands vs. child negativism). Yet, the fact remains that we are still dealing with <u>scores</u> that reflect behavioral dimensions. If the word "trait" offends, then another label will do as well. Furthermore, the scores are obtained for the same purposes that trait scores are obtained--to correlate with some other variable. Generally, behavior modifiers "correlate" these scores with the presence or absence of some treatment procedure but certainly our data is not limited to this one objective. In our own research, for example, we are currently comparing children's deviant behavior rates in their homes with their deviancy in the school classroom (Walker, Johnson, & Hops, 1972) and comparing the deviancy

rates of normal children with those observed in referred or "deviant" children (Lobitz & Johnson, 1972). The most elementary knowledge of the concept of reliability tells us that some minimal level of behavior score reliability is necessary before we can ever hope to obtain any significant relationship between our behavioral score and any external variable. Thus, the requirement of score reliability is just as important in research employing behavioral assessment as it is in more traditional forms of psychological assessment, but with only a few exceptions (e.g., Cobb, 1969; Harris, 1969; Olson, 1930-31; Patterson, Cobb, & Ray, 1972) behaviorists have ignored this important issue.

As a consequence of the reasoning presented above, we have been particularly cognizant of the reliability of the scores used in our research. We were quite encouraged to find, for example, that the odd-even-split-half reliability of our "total deviant behavior score" in a sample of 33 "normal" children was .72. This reliability was computed by correlating the total deviant behavior score obtained on the first, third, and first half of the fifth day with the same score obtained from the remainder of the period. After applying the Spearman-Brown correction formula, we found that the reliability of this score for the entire five-day observation period was .83. This relatively high level of reliability indicate that this score should, at least in a statistical sense, be quite sensitive to manipulation or to true relationships with other external variables (e.g., social class, or educational level of the parents). Other behavioral scores which are important to our research include: a) the proportion of generally negative responses of the parents (corrected reliability = 90), b) the proportion of generally positive responses of parents (corrected reliability = .87),

c) the median reliability of the 35 individual codes (r = .69), d) the corrected median reliability of the deviant codes = .66, 3) the number of parental commands during the observation (corrected reliability = .85), and f) the compliance ratio (i.e., compliances/compliances + noncompliance) of the child (corrected reliability = .49). The reliability of the compliance ratio is not as high as we might have wished, but it may still be high enough to be sensitive enough for powerful manipulations. We have been less fortunate in obtaining good reliability scores on some other statistics important to our research efforts. For example, the compliance ratios to specific agents (i.e., to mothers or fathers) have yielded rather low reliabilities. The reasons for this are two-fold: First, ratio scores are always less reliable than are their componant raw scores, because they combine the error variance of both components. Second, and of more general importance, these scores are based on relatively few occurrences. On the average, for example, fathers give only 36 commands over the five-day period. These occurrences must then be divided for the compliances and noncompliances and further split in half for the odd-even reliability estimate. By the time this erosion takes place, there are few data points on which to base reliability estimates. This problem is even more profound when we use one day of compliance ratio data to compute observer agreement on this statistic, since, on the average, fathers give only 7.2 commands per day. Thus, when we are dealing with behavioral events of fairly low base rate, observer agreement correlations and reliability coefficients may often not be "fairly" computed because there is simply not enough data. In classical test theory terminology, there may often not be enough "items" on the behavioral test to permit an accurate estimation of the reliability of the

score. What should we do with cases of this kind? A methodological purist might argue that we should throw out this data and use only scores with proven high reliability and observer agreement. We would argue that this course would be a particularly unfortunate solution for several reasons. First, low base rate behaviors are often those of special importance in clinical work. Second, if low reliability reflects nothing more than random, unsystematic error in the measurement instrument, it cannot jeopardize or provide a confounding influence on positive results (i.e., it cannot contribute to the commission of Type I errors). But, either low reliability or low observer agreement does have profound implications for the meaning of negative results (i.e., the commission of Type II errors). Fortunately, the effects of many behavior modification procedures are so dramatic that they will emerge significant in spite of relatively low reliability or observer accuracy.

In one of the other few examples of reliability data in the behavior modification literature, Cobb (1969) found that the average odd-even reliability of relevant behavioral codes used in the school setting was only .72. Yet, Cobb (1969) found that the rates of certain coded behaviors showed strong relationships to achievement in arithmetic. Thus, relatively low reliability or observer agreement jeopardizes very little the meaning of positive results, but leaves negative results with little meaning. There is, however, one very critical qualifying point to this argument. It is that the error expressed in low reliability or observer accuracy must be random, unsystematic, and unbiased. With this consideration in mind, we now move to what are perhaps the most important methodological issues in naturalistic research--observer bias and observee reactivity to the observation process.

The Problem of Observer Bias in Naturalistic Observation

Shortly after the turn of the century, O. Pfungst became intrigued with a mysteriously clever horse named Hans. By tapping his foot, "Clever Hans" was able to add, subtract, multiply and divide and to spell, read, and solve problems of musical harmony (Pfungst, 1911). Hans' owner, a Mr. von Osten, was a German mathematics teacher who, unlike the vaudeville trainers of show animals, did not profit from the horse's peculiar talents. He insisted that he did not cue the animal and, as proof, he permitted others to question Hans without his being present. Pfungst remained incredulous and began a program of systematic study to unravel the mystery of Hans' talents.

Pfungst soon discovered that, if the horse could not see the questioner, Hans could not even answer the simplest of questions. Neither would Hans respond if the questioner himself did not know the answer. Pfungst next observed that a forward inclination of the questioner's head was sufficient to start the horse tapping, and raising the head was sufficient to terminate the tapping. This was true even for very slight motions of the head, as well as the lowering and raising of the eyebrows and the dilation and contraction of the questioner's nostrils.

Pfungst reasoned and demonstrated that Hans' questioners, even the skeptical ones, expected the horse to give correct responses. Unwittingly, their expectations were reflected in their head movements and glances to and from the horse's hooves. When the correct number of hoof taps was reached, the questioners almost always looked up, thereby signaling Hans to stop (Rosenthal, 1966).

Some fifty years later, Robert Rosenthal began to investigate the importance of the expectations of experimenters in psychological research.

In his now classical article, Rosenthal(1963) presented evidence suggesting that the experimenter's knowledge of the hypothesis could serve as an unintended source of variance in experimental results. In a prototypical study, Rosenthal and Fode (1963) had naive rats randomly assigned to two groups of undergraduate experimenters in a maze-learning task. One group of experimenters was told that they were working with maze-bright animals and the other group was told that their rats were maze-dull. The group of experimenters which was led to believe that their rats were maze-bright reported faster learning times for their subjects than the group which was told their animals were maze-dull. An extension of this finding to the classroom was offered by Rosenthal and Jacobson (1966). Teachers were led to believe that certain, randomly selected students in their classrooms were "late bloomers" with unrealized academic potential. Pre- and post-testing in the fall and spring suggested that children in the experimental group (late bloomers) had a greater increase in IQ than did the controls.

The purpose of this section will be to examine the problem of experimenter-observer bias with regard to naturalistic observational procedures. The amount of literature which deals directly with observer bias in naturalistic observation is sparse (Kass & O'Leary, 1970; Skindrud, 1972; Kent, 1972). However, Rosenthal has written an extensive review of experimenter bias in behavioral and social psychological research (Rosenthal, 1966). In spite of failures to replicate many of Rosenthal's findings (Barber & Silver, 1968; Clairborn, 1969) and extensive criticisms of Rosenthal's methodology (Snow, 1969; Thorndike, 1969, Barber & Silver, 1968), the massive body of literature compiled and summarized by Rosenthal (1966) remains the

best available resource for conceptualizing the phenomenon of observer bias and for isolating possible sources of bias relevant to naturalistic observation. A brief review of this literature follows with a focus on integrating implications from this literature with naturalistic observational procedures. In addition, we will give consideration to the few experiments which have directly investigated observer bias in naturalistic observation and further consider some proposals for dxperiments yet to be conducted. Finally, suggestions for minimizing observer bias will be outlined and data on this problem from our laboratory will be presented.

Conceptualization of Observer Bias

Rosenthal (1966) has defined experimenter bias "as the extent to which experimenter effect or error is asymmetrically distributed about the 'correct' or 'true' value." Observer errors or effects are generally assumed to be randomly distributed around a "true" or "criterion" value. Observer bias, on the other hand, tends to be unidirectional and thereby confounding.

Sources of Observer Bias

An important distinction should be drawn between observer error and observer effect on subjects. Invalid results may be contributed solely by systematic or "biased" errors in recording by observers. Or, invalid findings may be realized as a result of the effect that the observer has on his subjects (Rosenthal, 1966). First we will consider recording error as a source of observer bias.

Kennedy and Uphoff (1939) illustrate the problem of recording errors in an experiment in extrasensory perception. The observers' task was simply to record the investigator's guesses as to the kind of symbol being "trans-

mitted" by the observer. Since the investigators guesses for the observers had been programmed, it was possible to count the number of recording errors. In all, 126 recording errors out of 11,125 guesses were accumulated among 28 observers. The analysis of errors revealed that believers in telepathy made 71.5 percent more errors increasing telepathy scores than did non-believers. Disbelievers made 100 percent more errors decreasing the telepathy scores than did their counterparts. Sheffield and Kaufman (1952) found similar biases in recording errors among believers and nonbelievers in psychokinesis on tallying the results of the fall of dice. Computational errors in summing recorded rates have also been documented by Rosenthal in an experiment on the perception of people (Rosenthal, Friedman, Johnson, Fode, Schill, White, & Vikan-Kline, 1964).

It is doubtful that these recording and computational errors were intentional. However, as Rosenthal (1966, p. 31-32) notes, data fabrication or intentional cheating is not absent in psychological research, especially where undergraduate student experimenters are employed as data collectors. Rosenthal points out that these students "have usually not identified to a great extent with the scientific values of their instructors." Students may fear that a poor grade will be the result of an accurately observed and recorded event which is incompatible with the expected event. Of two experiments by Rosenthal which were designed to examine intentional erring by students in a laboratory course in animal learning, one revealed a clear instance of data fabrication (Rosenthal & Lawson, 1964) and the other showed no evidence of intentional erring but did show some deviations from the prescribed procedure (Rosenthal & Fode, 1963). Another study employing student experimenters by Azrin, Holz, Ulrich, and Goldiamond (1961) replicated

Verplanck's (1955) verbal conditioning experiment. However, an informal post-experimental check revealed that data had been fabricated by the student experimenters. Later, the authors employed advanced graduate students as experimenters and found that Verplanck's results were not replicated.

The implications for naturalistic observation are obvious. Observer error, whether it be unintentional or intentional, incurred during recording or during computation, must be guarded against by accuracy checks and by carefully concealing the experimenter's hypotheses. Although observer agreement checks do not rule out the possibility of bias among the observers whose data is compared, it at least arouses suspicion where agreement figures are low and disagreements are consistent. Ideally, observers should not be made responsible for the tallying of their own data. Computations should be made by a nonobserver who is removed from knowledge of the observations. Observers should be selected on the basis of their identification with scientific integrity and admonitions against possible biasing effects should be repeated during the course of the experiment. Finally, observers should be encouraged to disclose to the experimenter both the nature and sources of any information they receive that might be relevant to the objectivity of their observations. A questionnaire, filled out after observation sessions, can facilitate this disclosure.

The other source of observer bias, which Rosenthal discusses (Rosenthal, 1966), is the effect of the observer's expectancy on the subject. If an observer has an hypothesis about a subject's behavior, he may be able to communicate his expectations and thereby influence the behavior.

Expectancy effects have previously been alluded to in Rosenthal's study with animal laboratory experimenters (Rosenthal & Fode, 1963) and

teachers in the classroom (Rosenthal & Jacobson, 1966). Rosenthal's first major study in expectancy effects is instructive in its simplicity. Rosenthal and Fode (1963) had 10 experimenters obtain ratings from 206 subjects on the photo person-perception task. All 10 experimenters received identical instructions except that five experimenters were informed that their subjects would probably average a +5 success rating on the ten neutral photos while the other five experimenters were led to expect a -5 failure average. The results revealed that the group given the +5 expectation obtained an average of +.40 vs. the -5 expectation group which yielded a -.08 score. These differences were highly significant and subsequent replications have supported these findings (Fode, 1960; Fode, 1965).

The implications for naturalistic observational procedures of the expectancy effect on the subject's behavior are most discomforting. If, as in the Rosenthal laboratory studies, observers in the natural setting can communicate their expectancies to their subjects such that the subject's behavior falls in line with these expectations, a serious threat to internal validity is posed. Assuming that humans are no less sensitive to subtle cues than Mr. von Osten's Clever Hans, it seems reasonable to infer that observer expectancy effects are operative in the natural setting. Consider the not atypical case of an observer who records selected deviant behaviors of a child in a classroom before, during, and after treatment. Seldom is it not obvious to the observer when treatment begins and ends. Assuming that an observer might infer the expectations of the experimenter in such a setting, how might he communicate these expectations to his subjects? One way of influencing the targeted child is by nonverbal expressive cues. Expressions of amusement by the observer during baseline might inflate deviant behaviors. During intervention, expressions of disapproval or

caution by the observer might reduce the subject's deviant rate. These biasing effects may be systematic and confounding.

Although few studies have systematically assessed the effects of observer bias in the natural setting, many field investigators have taken note of the expectancy phenomenon, and have included procedures to minimize its effect. One such technique is to mask changes in experimental conditions (e.g., Thomas, Becker, & Armstrong, 1968). Another is to keep observers unaware of assignment of subjects to various treatment or control conditions (e.g., O'Conner, 1969). The addition of new observers in the last phase of a study who are naive to previous manipulations is another approach (e.g., Bolstad & Johnson, 1972).

Three studies in the natural setting shed further light on expectancy effects with naturalistic observational procedures. Rapp (1966) had eight pairs of untrained observers describe a child in a nursery school for a period of one minute. One member of each observer pair was subtly informed that the child under observation was feeling "under par" that day and the other that the child was "above par." In fact, all eight children showed no such behaviors. Seven of the eight pairs of observers evidenced significant discrepancies between partners in their description of the nursery children in the direction of their respective expectations. Both recording errors and expectancy effects on the subjects' behavior may have contributed to this demonstration of observer bias.

A second study by Azrin et al. (1961) employed untrained undergraduate observers who were asked to count opinion statements of adults when they spoke to them. The observations of those who had been exposed to an operant interpretation of the verbal conditioning phenomenon under study were the exact opposite of those given a psychodynamic interpretation.

Again, both the expectancy effects of the observer on the subject and recording errors may have accounted for the observer bias. Post experimental inquiries by an accomplice student revealed that recording errors were the main factor. The accomplice learned that 12 of the 19 undergraduates questioned intentionally fabricated their data to meet their expectations.

A third study by Scott, Burton and Yarrow (1967) allows a comparison between the simultaneous observations of hypothesis informed (Scott herself) and uninformed observers. The observers coded behavior into positive and negative acts from an audio-tape recording of the targeted child and his peers. The informed observer's data differed significantly from the others' in the direction of the experimenters' hypothesis.

These three studies strongly suggest that data collected by relatively untrained observers are influenced by observer expectations. Do these findings generalize to the observations of professional observers who are highly trained in the use of sophisticated multivariate behavior codes? As indicated earlier, the amount of available research which directly pertains to this question is limited and somewhat equivocal.

Kass and O'Leary (1970) conducted the first systematic attempt to manipulate observer expectations in a simulated field-experimental situation. Three groups of female undergraduates observed identical videotaped recordings of two disruptive children in a simulated classroom. The observers were trained in nine category codes of disruptive behavior. Group I was then given the expectation that soft reprimands from the teacher would increase the rate of disruptive behavior. Group II was told that soft reprimands would decrease disruptive behavior. And, Group III was given no expectation at all about the effects of soft reprimands. Rationales were

given each group explaining the reasons for each specific expectation. The effects of these expectations were assessed by having the observers watch four days of baseline and five days of treatment data. The interaction between the mean rate of disruptive behavior in the three conditions and the two treatment conditions was significant at the .005 level, indicating the presence of observer bias. Ronald Kent (1972) has suggested that these reported effects of expectation bias were confounded with observer drift in the accuracy of recording. When different groups of raters, who are interreliable within groups, fail to frequently compute agreement between groups, they may "drift" apart in their application of the behavioral code. However, it should be noted that when this drift, comprised of recording errors, is alligned asymmetrically in the direction of the expectation, then the drift is, by definition, observer bias.

Skindrud (1972) attempted to replicate the findings of Kass and O'Leary (1970). Observers were divided into three groups, each group given a different expectation about video-taped family interactions. The first group was given the expectation that when the father was absent there would be more child deviant behaviors than when the father was present. A second group was given the opposite expectation. Appropriate rationales were provided for each of these two groups. An additional control group was added with no expectations provided regarding father-present or father-absent tapes. All observers were checked at the end of training on the rates of deviant behaviors they recorded and subsequently matched on this variable when assigned to conditions. Throughout the study, observer agreement data was collected randomly. During training, reliability was checked daily, and the average observer agreement prior to the beginning of the manipulation was 64%. The results of the study gave no evidence for observer bias. There were no

significant differences between groups and no significant interaction effects. There was little drift in the accuracy with which the code was used. Sequential reliabilities were computed for the increase, decrease, and control groups with average observer accuracy of 58.5%, 57.6%, and 58.4%, respectively. These accuracy figures were computed by comparisons with previously coded criterion protocols. The relatively small and consistent decline in accuracy is consistent with the failure to find bias.

A similar unsuccessful attempt to replicate Kass and O'Leary (1970) was reported by Kent (1972). Kent found that knowledge of predicted results was not sufficient to produce an observer bias effect. However, when the experimenter reacted positively to data which was consistent with the given predictions and negatively to nonconsistent data, a significant observer bias effect was obtained.

The available literature dealing with observer bias in naturalistic observation is both sparse and contradictory. Furthermore, the few studies available have focused exclusively on only one source of observer bias, namely, recording errors or errors of apprehension. Thus far, no one has systematically investigated the effects of the observer's expectancies on the subjects' behavior. In the three studies reported above, all observations were made from video-taped recordings. There were no opportunities for the observers to communicate their expectancies to their subjects. Yet, in most studies employing naturalistic observational procedures, observers do have that opportunity.

An important study which needs to be conducted is one which examines the observer's expectancy effects on the subject. First, it would be interesting to determine if observers could nonverbally communicate their

expectancies to subjects such that the subject's behavior changes in the direction of the expectancy. The next step, of course, would be to replicate this same design without specifically asking observers to attempt to influence subjects, but merely to give them an expectation.

Perhaps the most important test of observer bias effects will be the one which combines recording errors and effects of observer expectancy on subjects in the naturalistic setting. One can question the generalizability of highly controlled laboratory studies to live observations and to research projects in which the observers are more invested in the outcome of the research. The generalizability of studies which employ only taped versions of a subject's behavior is further limited by excluding the possible effects of an observer's expectancy on his subject's behavior.

Another variable which seems crucial to observer bias in the naturalistic setting is the observer's responsiveness to admonitions to remain scientific, objective, and impartial in the collection of data. Rosenthal (1966) stresses the importance of the experimenter-observer's identification with science and objectivity. He cites evidence suggesting that graduate students obtain less biased data than undergraduates and interprets this difference as a function of identification with science. Perhaps observers who are repeatedly reminded to be impartial might be less susceptible to the influence of biasing information than observers not given these admonitions.

A dimension which seems important in considering observer bias is the specificity of the code. In most of the Rosenthal literature, the dependent variable is scaled between such global poles as success and failure. Intuitively, it seems logical that the more ambiguous the dependent measure, the greater the possibility for bias. A multivariate coding system, with

well-defined behavioral codes might be expected to restrict interpretive bias. This is an empirical question worthy of examination.

Another variable which might greatly affect observer bias is observer agreement. The greater the observer agreement, the less likely is observer bias, even among observers with the same expectancy.

Until more information is available on observer bias effects in naturalistic observation, it seems very critical to do everything possible to minimize the potential for these effects. Whenever possible, observers should not have access to information that may give rise to confounding consequences and encouraged to reveal the nature and source of any information they do receive. In our research, we are currently observing both families in clinical treatment and "normal" or nontreated families. Knowledge of a family's status might seriously affect the observer's data. Also, knowledge about treatment stages (baseline, mid-treatment, post-treatment, and follow-up) might effect the observers' data. After each observation, it is our policy to have observers fill out a questionnaire concerning the nature and source of any biasing information. Thus far, of 75 observations of referred families, observers have considered themselves informed only 36% of the time. And, in all of these cases, their information was correct. This information usually comes from a member of the family being observed (56%). Other sources of information include information leaks from the therapists (12%), the Child Study Center Clinic generally (16%), and other sources (16%). Of the observer considering themselves informed as to the clinic vs. "normal" status of the families, 29% also considered themselves informed as to treatment stage, but only two-thirds of these observers were correct in their discrimination. In only 20% of the cases did the observer actually know

the status of the case (i.e., clinic vs. normal) and the treatment stage (baseline vs. after baseline). Of the observers considering themselves completely uninformed of the families' status, their guessing rate (clinical or "normal") barely exceeded chance at 51%. Their guesses as to the four stages of treatment were 36% correct and 80% correct on the discrimination between baseline and after baseline.

Of the "normal" families seen, observers have considered themselves informed as to family status only 17% of the time. However, in only 45% of these cases were the observers actually correct in making the discrimination. In the uninformed observations, however, observers were able to guess the family's status correctly 75% of the time.

Not only are these questionnaires beneficial in gauging the amount of potentially biasing information that observers discover, but they are helpful in two other ways as well. First, by revealing sources of information leakage, steps can be made to eliminate these sources. Second, questionnaires, given after each family is observed, serve as a regular reminder for the importance of unbiased, objective recording of behavior.

It is difficult to make any firm conclusions about the presence or absence of observer bias in naturalistic observation. Clearly, more research is needed on this question. However, it should also be clear that the potentially confounding influence of observer bias cannot be ignored and that steps can and should be taken to minimize its possible effect.

The Issue of Reactivity in Naturalistic Observation

In the previous section, we have considered the effects of an observer's bias in naturalistic observation. In this section, we will discuss the effect of the observer's presence on the subjects being observed. Whereas observer bias can potentially invalidate comparisons by confounding influences, the reactive effects to being observed primarily constitute a

threat to the generalizability of the findings. That is, subjects' observed behavior in the natural setting may not generalize to their unobserved behavior. Webb, Campbell, Schwartz, and Sechrest (1966) have defined reactivity in terms of measurement procedures which influence and thereby change the behavior of the subject. Weick (1968) has also referred to reactivity as "interference" or the intrusiveness of the observer himself upon the behavior being observed. Clearly, situations which are highly reactive in terms of "observer effects" are not likely to be generalizable to situations in which such effects are absent.

Reactive effects have been studied with two basic paradigms: a) by the study of behavioral stability over time and b) by comparison of the effects of various levels of obtrusiveness of the observation procedure. In employing the first method, investigators have typically examined behavioral data for change over time in the median level and variance of the dependent variable. In general, it has been assumed that change reflects initial reactivity and progressive adaptation to being observed. This interpretation is particularly persuasive if there is an obvious stability in the data after some initial period of change or high variability. While this is a viable way of checking for reactivity effects, it is a highly indirect method and relies on assumptions concerning the causes of observed change. It is obvious that other processes could account for such change. Furthermore, the lack of change certainly does not indicate a lack of reactive effects. The second method, comparing obtrusive levels of observation, appears less inferential than the first method. The problem with this method is that it only provides a picture of relative degrees of reactivity between obtrusiveness levels; it does not provide a measure of the degree of reactivity relative to the true, unobserved behavior. However, this problem can

be remedied if one of the observational treatments in the comparison is totally unobtrusive or concealed.

To what extent does reactivity occur in naturalistic observation? The literature addressing this question is commonly reported in reviews to be contradictory (Wiggins, 1970; Weick, 1968; Patterson & Harris, 1968). Several studies have been cited as providing evidence for the position that reactive effects may be quite minimal. Others have been cited which suggest that reactive effects are quite pronounced. The purpose of this review is to: a) reconsider the contradictions in the literature on reactivity, b) tease out those factors which seem to account for reactivity, and c) propose further investigations which isolate these factors.

In a number of reviews on reactivity, several studies have been consistently cited which support the position that reactivity does not constitute a major threat to generalizability. One study frequently cited is the timely investigation of a Midwest community by Barker and Wright (1955). In this admirable study, careful naturalistic observations were made of children under ten years of age and their daily interactions with peers and parents. The authors assumed that reactive effects were short lived and that the adults and other members of the families quickly habituated to the presence of the observers. In addition, it was reported that, with the younger subjects in the sample, reactive effects were slight. However, these findings should be interpreted with much caution. What is easily lost sight of in the summaries of this work is that the observers in this study were free to interact with the subjects in a friendly but nondirective manner. In fact, the basis for the authors' conclusion that reactive effects were not pronounced was the

finding that "only" 20% of the children's behavioral interactions were with the observer. Allowing the observer to interact with the subject must certainly have increased the intrusiveness of the observer and provided the opportunity for the observer to influence the subject's behavior. The authors' other conclusion that reactivity, as measured by frequency of interactions, positively correlated with age is also suspect in that children below the age of five were not always informed that they were being observed, whereas children above this age were.

Another study commonly cited in support of the minimal reactivity position is that of Bales (1950). In this controlled laboratory investigation, the behavior of a discussion group was not found to be changed by three levels of observer conspicuousness. This finding, however, may be limited to the laboratory setting.

Two additional studies, frequently mentioned as supportive of the minimal reactivity argument, made use of radio transmitter recording in the naturalistic environment. Foskin and John (1963) had a married couple wear a transmitter the entire time they were on a two-week vacation. Purcell and Brady (1965) outfitted adolescents in a treatment center with a similar recording device for one hour a day. When the protocols in both studies were examined for the frequency of comments about being observed or listened to, it was found that such references declined to a zero level either during the first or second day of recording. This is not to say, of course, that these subjects were not still aware of, and affected by, the recording device; the results only indicate that the subjects talked about the device less after the first day.

A recent investigation by Martin, Gelfand, and Hartmann (1971) can also be interpreted as providing evidence for low levels of reactivity to

observation. This study involved 100 elemntary school children, ages [illegible] 7. Equal numbers of male and female subjects were assigned to five observation conditions following exposure to an aggressive model: a) observer absent, b) female adult observer present, c) male adult absorver present, d) female peer observer present, and e) male peer observer present. During the free-play session, the subjects' aggressive behavior was recorded by observers behind a one-way mirror. No significant differences in aggressive behaviors were obtained between the observer-present and observer-absent conditions. The absence of differences between these two levels of obtrusiveness in observation suggests little or no reactivity to the presence of an observer. Within the observer-present condition, however, it was found that peer observers significantly facilitated imitative aggressive responding in both boys and girls compared to adult observers. Also, there was more imitative aggression when the observer was the same sex as the subject. The girls, but not the boys, showed significant increases in aggressive output over time when the observer was present but not when the observer was absent. This latter finding suggests that girls manifest initial reactivity to the presence of an observer but later habituate to the observer's presence. It is interesting that both paradigms for measuring reactivity were used in this investigation and that each method supports different conclusions about the degree of reactivity. In considering the generalizability of these findings to naturalistic observation procedures, it should be noted that observers in this study were instructed to not pay attention to the subjects and were either seated facing away from the subjects (adult observers) or given a coloring task to complete (peer observers). With naturalistic observation procedures, on the other hand, observers typically pay very close attention to their subjects.

for the most part, the [illegible] which have been purported to provide evidence for minimal levels of reactivity to observation have been based on data of questionable meaning and/or are restricted to highly specific circumstances (e.g., Bales, 1950; Martin et al., 1971)

Many other studies have been cited as demonstrating considerable reactive effects of observation in naturalistic settings. One such study is that of Polansky, Freeman, Norwitz, Irwin, Lamiri, Rappaport, and Whaley (1949). These investigators observed delinquent children in a study of group emotional contagion phenomena. The children were informed that the observers were studying their reactions to various aspects of the summer-camp program. The authors report that during the first week of observations, the children essentially ignored the presence of the coders. But, during the second week, many "blow-ups" or outbursts were directed against the coders, especially by the older children. The authors speculate that the aggressiveness of the children can be explained as a result of resistance to being observed. They also conceptualized their resistance hypothesis was confounded by "the second week syndrome" which they describe as an increasing anti-adult aggressiveness that typically evolves after the children have adjusted to the camp, peaking in the second week. It is unclear as to what to conclude from this study about reactivity. Was reactivity most prevalent when children were aggressive toward the observers in the second week? Or, was reactivity most intensive during the first week when the delinquent children were "suppressing" aggression prior to habituating to the unfamiliar environment. And, more importantly, how much of this pattern of cooperation in the first week and anti-adult aggressiveness in the second would have occurred in the absence of observers? This question is left unanswered by the Polansky et al. (1949) study.

A more sophisiticated study illustrating considerable reactive effects is that of Roberts and Renzaglia (1965). In this study, eight therapists saw two "clients" (students solicited from an introductory psychology course) for three sessions. The three conditions were: a) with a tape recorder in the room, b) microphone only, c) with a concealed microphone and a concealed recorder. While no differences were found as to the number of minutes that subjects talked in each condition, it was found that the clients made more favorable self-references when they knew they were being recorded (conditions a and b, and more unfavorable self-references when they did not realize they were being recorded (condition c). While the generalizability of this study is limited by the small n and structured situation employed, it does provide evidence for reactivity effects.

The Bechtel (1967) study has also been presented in reviews as a study demonstrating considerable reactive effects. Bechtel studied the movements of people in an art museum with an automated recording device, called an hodometer. In one condition, subjects were requested to go through the art exhibit in one room and rank the prints according to preference. Another group was given the same instructions and also informed that they were being observed. For the first group, subjects were found to leave the room with an average latency of 313 seconds, covering an average area of 76 square feet. The second group left after 220 seconds and covered only 58 square feet. Bechtel concludes that being observed constituted an aversive situation which led to escaping the room faster. No statistics were provided to determine whether or not the reported averages were significantly different, nor were the number of subjects in each condition reported. Unfortunately, this kind of study does not provide any information about habituation effects, since subjects were allowed to leave the setting.

[illegible] concerned [illegible] with [illegible] environment. In point of fact, [illegible] reported, [illegible] environment [illegible] activity level of the [illegible] activity [illegible] of the laboratory setting.

Few [illegible] reactivity [illegible] natural setting [illegible], and [illegible] (19[illegible]7) provide evidence for [illegible]. They found that [illegible] in their [illegible] was [illegible]. However, for one of [illegible] families, [illegible] even after six hours of [illegible].

A study by Patterson and Harris [illegible] cited as evidence for considerable reactive effects of observation in a naturalistic environ-

ment. This article is the only study available which was designed specifically to manipulate and measure observer effects in the homes of the families observed. In this study, data obtained from mothers on their own families were compared with the data on the same family collected by an outside observer. There were three conditions, with five families per condition: a) mothers collected the first five ten-minute sessions of observational data and an outside observer collected the second five sessions of data on the child and father only (M-O), b) the observer collected all ten sessions as a test for habituation effects (O-O), and c) the mothers collected all ten sessions as a control for habituation effects (M-M). The dependent variables were the rates of total behaviors and the rate of deviant behaviors. A problem in the research design of this study should be noted. The mother was present in the family as a participant in the second condition (O-O) and the second half of condition a (M-O), but was not a participant when she was an observer in condition c and the first half of condition a. These comparisons are confounded by mother presence and absence. In spite of this confound, which would probably bias in favor of showing group differences, no main effects for groups were found in analysis of variance for either the rate of total interactions or deviant behaviors. Thus, on the initially selected dependent variables, no reactive effects were apparent.

Patterson and Harris also divided their groups into high and low rate interactors on the basis of the first five sessions. On the frequency of total interactions measure, high rate interactors in the first five sessions showed significant reductions in rate during the last five sessions. The authors describe this decline as a "structuring effect" in that the subjects appeared to program some activity together in the first five sessions.

Conversely, the low rate interactors in the first five sessions showed slight increases in rates during the last five sessions. The authors describe this transition as an habituation effect in that subjects initially involved themselves in solitary activities or attempted to escape the observational situation but later adjusted to it and interacted more. In general, there were no changes in deviant behavior from the first set of five observations to the last set of five. The only significant finding was that subjects who displayed low rates of deviant behavior in the first five sessions (under the M-O condition) increased their rate in the last five sessions. However, it is possible that the mothers were recording less deviant behaviors and more positive behaviors in the first five sessions than were the observers in the second five sessions, thus contributing differentially to main trials effects. An observational study by Rosenthal (1966) supports such a thesis. He found that parents tended to code more positive changes in their children than were actually present. And, Peine (1970) found that parents were less observant of their children's deviant behaviors than were nonparent observers.

Patterson and Harris conclude that "generalization about 'observer effects' should probably be limited to special classes of behavior " (p. 16). A more recent study by Patterson and Cobb (1971) analyzed the stability of each of the 29 behavior codes used in their coding system. If it is assumed that individuals adapt to the presence of an observer over time, then a repeated measures analysis of variance should reveal differences in the mean level of various behaviors. Patterson and Cobb analyzed data for 31 children from problem and nonproblem families over seven baseline sessions. None of the changes in mean level for the codes produced a significant effect over time. The investigators conclude that the observation data were

fairly stable for most code categories. It is possible, of course, that had observations continued over a longer period of time, significant changes in mean level for some behaviors would have been discovered. Given that families were rarely observed on consecutive days by the same observer, it is possible that different observers could have resensitized the families each day, thereby extending the period required for adaptation.

In summary, there are a few well-designed studies which have discovered reactive effects (e.g., Roberts and Renzaglia, 1965; Bechtel, 1967; White, 1972), but there are several others where the meaning of the results is unclear. There can be little doubt that the entire question has been inadequately researched. Any general conclusions about the extent of reactivity in naturalistic observation would seem premature at this time.

As White (1972) points out, the finding of reactive effects seems to depend on many factors, including the setting (e.g., home, school, laboratory), the length of observation, and the constraints placed on subjects by the conditions of observation (e.g., no television during observations, remain within two adjacent rooms, etc.). Furthermore, it should be realized that reactivity may or may not be discovered depending upon what paradigm of measurement is used (e.g., Patterson & Harris, 1968; Martin et al., 1971) and what variables are analyzed as dependent variables (e.g., Roberts & Renzaglia, 1965; White, 1972). Unless these factors are controlled for in comparing experiments on reactivity, both contradictions and consistencies as to the relative presence or absence of reactivity may falsely appear.

Assuming that reactivity to being observed in naturalistic settings does occur, even if only to some minimal degree, the critical task is to localize the sources of interference so that they can be dealt with more

directly. Four such sources will be discussed and experiments will be proposed to measure the extent of their instrusiveness.

Factor 1: Conspicuousness of the Observer

The literature points to the level of conspicuousness or intrusiveness of the observer as an important factor contributing to reactivity. Presumably, the more novel and conspicuous the agent of observation, the more distracting are the effects upon the individuals being observed. It would also follow that longer habituation periods would be required for more distracting observational agents in order to achieve stability of data.

Bernal, Gibson, William, and Pesses (1971) compared two observation procedures which would presumably vary on obtrusiveness. These investigators compared data collected by an observer with that collected by means of an audio tape recorder which was switched on by an automatic timing device. The family members involved in this study were aware of the presence of the recorder but were unaware of the exact time of its operation. The primary purpose of this study was to explore the feasibility of the audio tape method and to explore the relationship of data collected by the two methods rather than to study reactivity per se. The results indicated that, during the same time interval, there was a high relationship between the mother's command rate as coded by the observer and from the tape (r = .86) but that the observer coded more commands. Similar results were obtained when the observer's data was compared with data based on coding of the audio tapes from different time intervals. The question arises as to how much of this latter discrepancy was due to differences in levels of reactivity and how much was due to differences associated with the source of coding. The authors point out, for example, that the observer could code gestural

commands while the coder using the tape could not. Since the discrepancies at the same time and at different times were of the same general order of magnitude, it is likely that most of the observed difference across time was due to the material on which coding was based rather than to differences in subject reactivity. To study the impact of reactivity effects separately, one might design such a study so that the same stimulus materials would be used for coding.

We are currently completing a study on reactivity which employs this strategy to compare reactivity associated with an observer present in the home carrying a tape recorder vs. the tape recorder alone. This study involves six days of observation for 45 minutes per day with single-child families. The two conditions are alternated so that the observer is present one evening and not present the next. The observer is actually a "bogus" observer. All behavioral coding is done on the basis of the tapes. It is our suspicion that reactivity to the tape recorder will be short lived and minimal compared to the reactivity associated with the observer present.

If these hypotheses are substantiated in this and other research, alternatives to having an observer present in the home should be explored. One solution to be seriously considered would be extended use of portable video or audio tape recording equipment. These recording devices could remain in the homes over an extended observation period to facilitate habituation effects. In addition, the devices could be preprogrammed to turn on and off at different times during the day so that the observed would not know when they are in operation (as in Bernal et al., 1971). This solution, which would, of course, require full knowledge and consent of the parties involved, appears to be a promising one for attenuating reactivity effects as well as solving problems of observer bias.

Factor 2: Individual Differences of the Subjects

Some people might be expected to manifest more reactivity to the presence of an observer than others. A "personality" variable such as guardedness might be correlated with degree of reactivity. For example, scores on the K scale of the MMPI (or other comparable tests) might be related to the effects of being observed in a natural setting.

The literature also suggests that age is correlated with reactivity. Several authors (Barker & Wright, 1955; Polansky et al., 1949) have suggested that younger children are less self-conscious and thereby less subject to reactive effects than older children. The Martin et al. (1971) study also suggests that sex might be an important factor accounting for different levels of reactivity. Experiments are needed which compare these individual difference variables in the natural setting with naturalistic observation procedures.

Factor 3: Personal Attributes of the Observer

Evidence from semi-structured interviews suggests that reactive effects may also be contributed by the unique attributes of the observer. Different attributes of the observer may elicit different roles on the part of the subject, depending upon what might be appropriate given the observer's attribute. Rosenthal (1966) reports several such attributes that have been demonstrated to yield differential effects, including the age of the observer,

sex, race, socio-economic class, and the observer's professional status (i.e., undergraduate observer vs. Ph.D. therapist). Martin et al. (1971) also discovered that both the factors of age and sex of the observer had differential effects on the subjects being observed. Varying any of these dimensions parametrically would be relatively simple in investigating this problem in the natural setting.

Factor 4: Rationale for Observation

Another factor that may be important in accounting for reactivity is the amount of rationale given subjects for being observed. Whereas the Bales (1950) study found no differential reactivity of three levels of observer conspicuousness in a group-discussion setting, Smith (1957) found that nonparticipant observers aroused hostility and uncertainty among participating group members. Weick (1968) suggests that this discrepency may have been a function of different amounts of rationale for the presence of an observer. We hypothesize that a thorough rationale for being observed might be expected to reduce guardedness, anxiety, etc., and thereby reduce the reactivity.

Observer reactivity is a problem that cannot be easily dismissed for naturalistic observation. There is sufficient evidence to suggest that observer reactivity can seriously limit the generalizability of naturalistic observation data. Clearly, factors accounting for reactivity need to be investigated and solutions derived to minimze the effects of the observer on the observed. In the next section, we will describe how reactivity, in addition to posing a problem for generalizability, can also interact with and confound the dependent variable.

Observee Bias:

Demand Characteristics, Response Sets and Fakability

Reactivity to observation will always be a problem for naturalistic research, but it would be a relatively manageable one if we could assume it to be a relatively constant, noninteractive effect. That is, if we knew that the presence of an observer reliably reduced activity level or deviant behavior by 30%, for example, the problem would not be too damaging to research investigations involving groups of subjects. But, what if the observee's reactivity to being observed interacts with the dependent variable under study.

Let us take the example of a treatment study on deviant children in which observations are taken prior to and after treatment. Prior to treatment, the appropriate thing for involved parents or teachers to do is to make their referred child appear to be deviant in order to justify treatment. The appropriate response at the end of treatment, on the other hand, is to make the child appear improved in order to justify the termination, please the therapist, etc. These are the demand characteristics of the situation. In this case, the reactivity to being observed is not constant or unidirectional, but interacts with and confounds the dependent variable. It is possible that any improvement we see in the children's behavior is simply the result of differential reactivity as a consequence of the demand characteristics of the situation. Now, let us suppose we employ a wait list control group and collect observational data twice before beginning treatment and at the same interval as used for the treated group. This procedure provides an excellent pretest-post-test control for our treated group. But, what of the demand characteristics of this procedure? On the first assessment, the involved

parents or teachers will probably behave in the same general way as their counterparts in the treated group, but by the second observation they may be more desperate for help and even more concerned to present their child as highly deviant. Thus, simply as a result of the demand characteristics involved, we might expect our treatment group to show improvement while the control groups would show some deterioration.

We also may wish to compare our referred children with children who are presumably "normal" or at least not referred for psychological treatment. Once again, however, we might anticipate that parents recruited for "normative" research on "typical" families would be more inclined than our parents of referred children to present their wards as nondeviant or good. In other words, a response set of social desirability could be operative with this sample making them less directly comparable to the referred sample.

These arguments would, of course, be even more persuasive if we were dealing with the observed behavior of the adults themselves. The foregoing observations on children assume, however, that the involved adults are capable of influencing children to appear relatively "deviant" or "normal" if they wish to do so (i.e., that observational data on children is potentially fakable by adult manipulation).

We have just completed a study (Johnson & Lobitz, 1972) which was directed at testing this assumption. Twelve sets of parents with four- or five-year-old children were instructed to do everything in their power to make their children look "bad" or "deviant" on three days of a six-day home observation and to make their children look "good" or "nondeviant" on the remaining three days. Parents alternated from "good" to "bad" days in a counterbalanced design.

Four predictions were made regarding the behavior of both children and parents. During the "fake bad" periods, it was anticipated that, relative to the "fake good" periods, there would be:

a) more deviant child behaviors,

b) a lower ratio of compliance to parental commands,

c) more "negative" responses on the part of parents, and

d) more parental commands.

Predictions a, c, and d were confirmed at or beyond the .01 level of confidence. Only the child's compliance ratio failed to be responsive to the manipulation. It will be recalled from the section on reliability that this statistic is by far the least reliable and thus the least sensitive (statistically) to manipulation. These results which demonstrate the fakability of naturalistic behavioral data indicate that this kind of data may potentially be confounded by demand characteristics and/or response sets.

We are aware of only one other study involving naturalistic observation which helps demonstrate this problem (Horton, Larson, & Maser, 1972). This study involved one teacher who was under the instruction of a "master" teacher for the purpose of raising her classroom approval behavior. She was observed, without her knowledge, by students in the class. The results clearly showed that her approval behavior was at a much higher rate when she was being observed by the "master" teacher than when she was not being observed. Generalization from overtly observed periods to periods of covert observation was very minimal indeed. More generalization was found when the "master" teacher's presence in the classroom was put on a more random schedule. This study is not completely analogous to most naturalistic research because, in this case, the observer and trainer were the same person

and the study is limited in generalizability because of the N = 1 design. Yet, in most cases, the observed are aware that the collected observational data will be seen by the involved therapist, teacher, or researcher, and if the problem exists for one subject, it is a potential problem for all subjects. Observee bias is really a special case of subject reactivity to observation. Thus, the potential solutions outlined in the previous section apply here as well. In general, we suspect that observation procedures which are relatively unobtrusive and which allow for relatively long periods of adaptation will yield less reactivity and observee bias.

Validity of Naturalistic Behavioral Data

Just as behaviorists have ignored the requirement of classical reliability in their data, they have also neglected to give any systematic attention to the concept of validity. Most research investigations in the behavior modification literature which have employed observational methods have relied on behavior sampling in only one narrowly circumscribed situation with no evidence that the observed behavior was representative of the subject's action in other stimulus situations. In addition, behaviorists have largely failed to show that the obtained scores on behavioral dimensions bear any relationship to scores obtained or the same dimensions by different measurement procedures. This fact calls into serious question the validity of any of this research where the purpose has been to generalize beyond the peculiar circumstances of the narrowly defined assessment situation. Of course, the methodological problems we have presented thus far all pose threats to the validity of the behavioral scores obtained. But, we would argue that even if all these problems could somehow be magically solved, the requirement for some form of convergent validity would still be essential.

As with reliability, there are many different methods of validation, but as Campbell and Fiske (1959) point out:

> Validation is typically convergent; a confirmation by independent measurement procedures. Independence of methods is a common denominator among the major types of validity (excepting content validity) insofar as they are to be distinguished from reliability. . . . Reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods.

Thus, convergent validity is established when two dissimilar methods of measuring the same variable yield similar or correlated results. Predictive validity is established when the measure of a behavioral dimension correlates with a criterion established by a dissimilar measurement instrument.

With only a few exceptions, behaviorists have restricted themselves to face or content validity. And, of course, it must be admitted that the face validity of narrowly-derined behavioral variables is often quite persuasive. This is particularly true in cases where the behavioral dimension under study has very narrow breadth or "band width." After all, a behaviorist might argue, what can be a more valid measure of the rate of a child's hitting in the classroom than a straight-forward, accurate count of that hitting. While this argument is persuasive, two counter arguments must be considered. First, because of all of the methodological problems which we have presented thus far, we can never be certain that the observed rates during a limited observation period are completely valid or generalizable even to very similar stimulus stiuations. While many of the problems we have outlined can be solved and othersattenuated, it is unlikely that all will ever be completely eliminated. Second, is it not still of consequence to know whether our behavior rate estimates have any relationship to other important and logically related external variables? Is it not important,

for example, to know whether or not the teacher and classmates of an observed high-rate hitter perceive this child as a hitter? It does seem important to us, particularly for practical clinical purposes, since we know that people's perceptions of others' behavior often have more to do with the way they treat them than does the subject's actual behavior. The need for establishing some form of convergent validation becomes even more profound as the behavioral dimensions we deal with increase in band width. As we begin to talk about such broad categories as appropriate vs. inappropriate behavior (e.g., Gelfand, Gelfand, & Dobson, 1967), deviant vs. nondeviant behaviors in children (e.g., Patterson, Ray, & Shaw, 1969; Johnson et al., 1972), or friendly vs. unfriendly behaviors (e.g., Raush, 1965), we are labeling broader behavioral dimensions. At this level, we are dealing with constructs, whether we like to admit it or not, and the importance of establishing the validity of these constructs becomes crucial. In most cases, these broad behavior categories have been made up of a collection of more discrete behavior categories and, in general, the investigators involved have simply divided behaviors into appropriate-inappropriate or deviant-nondeviant on a purely a priori basis. While the categorizations often make a good deal of sense (i.e., have face validity), this hardly seems a completely satisfactory procedure for the development of a science of behavior.

We have had to face this problem in our own research, where we have sought to combine the observed rates of certain coded behaviors and come up with scores reflecting certain behavioral dimensions. The most central dimension in this research has been the "total deviant behavior score" to which we have repeatedly referred in this chapter. Let us outline here the procedures we have used to explore the validity of this score. Although

we had a pretty good idea of which child behaviors would be viewed as "deviant" or "bad" in this culture, we attempted to enhance the consensual face validity of this score by asking parents of the "normal" children we observed to rate the relative deviancy of each of the codes we use in our research. Thus, in our sample of 33 families of four- and five-year-old children, we asked each parent to read a simplified version of our coding manual and characterize each behavior on a three-point scale from "clearly deviant" to "clearly nondeviant and pleasing." We established an arbitrary cut-off score and characterized any behavior above this cut-off as deviant. This resulted in a list of 15 deviant behaviors out of a total of 35 codes. The second step in validating this score and our implicit deviant-nondeviant dimension was presented in a study by Atkins and Johnson (1972). We had already divided our 35 codes into positive, negative, and neutral consequences. This categorization was done on a purely a priori basis with a little help from the data provided by Patterson and Cobb (1971) on the function of some of these codes for eliciting and maintaining children's behavior. We reasoned that behaviors which parents viewed as more deviant would receive relatively more negative consequences than would behaviors viewed as less deviant. To test this hypothesis, we simply rank ordered each behavior, first by the mean parental verbal report score obtained and second by the mean proportion of negative consequences the behavior received from family members. The results of this procedure are presented in Table 1. Not all 35 behaviors are

---
Insert Table 1 about here
---

included in this analysis, but the complex reasons for this outcome can more parsimoniously be explained in a footnote.[5] In any case, the Spearman Rank Order Correlation between the two methods of characterizing behaviors

on the deviant-nondeviant dimension was .73. This was an encouraging finding, but we noticed that the most dramatic exceptions to a more perfect agreement between the two methods involved the reasonable command codes (command positive, and command negative). These codes are used when the child reasonably asks someone to do something (positive command) or not to do something (negative command). Naturally, most parents felt that these innocuous responses were nondeviant. But, behaviorally, people don't always do what they are asked to by a four- or five-year-old child, and since noncompliance was coded as a negative consequence, it seemed that this artifact of our characterization might have artificially lowered this coefficient. By eliminating these two command categories from the calculation, the correlation coefficient was raised to .81.

The third piece of evidence for the validity of the deviant behavior score comes from the Johnson and Lobitz (1972) study already reviewed in the previous section. In this study, parents were asked to make their children look "good" and "nondeviant" for half of the observations and "bad" or "deviant" on the other half. They were <u>not</u> told how to accomplish this, nor were they told what behaviors were considered "bad" or "deviant." The fact that the deviant behavior score was consistently and significantly higher on the "bad" days lends further evidence for the construct validity of the score.

While evidence for the convergent or predictive validity of behavioral data is difficult to find in the literature, there are some encouraging exceptions to this general lack of data. Patterson and Reid (1971), for example, found an average correlation of .63 ($p < .05$) between parents' observations of their children's low rate referral symptoms on a given day and

the trained observer's tally of target deviant behaviors on that day. Several studies have found significant relationships between behavioral ratings of children in the classroom and academic achievement (Meyers, Attwell, & Orpet, 1968; D'Heurle, Millinger, & Haggard, 1959; Hughes, 1968). The data base of these studies is somewhat different from that currently employed by most behaviorists because they involve ratings by observers on relatively broad dimensions, as opposed to behavior rate counts. For example, dimensions used in these studies included "coping strength," defined as ability to attend to reading tests while being subjected to delayed auditory feedback (Hughes, 1968), or "persistence," defined as ". . . uses time constructively and to good purpose; stays with work until finished" (D'Heurle, Mellinger, & Haggard, 1959). Nevertheless, these studies demonstrate the potential for behavior observation data to provide evidence of predictive validity. Two other studies (Cobb, 1969; Lahaderne, 1968) yield similar predictive validity findings based on behavioral rate data. Lahaderne (1968) found that attending behavior as observed over a two-month period, provided correlations ranging from .39 to .51 with various standard tests of achievement. Even with intelligence level controlled, significant correlations between attentive behavior and achievement were found. Cobb (1969) obtained similar results in correlating various behavior rate scores with arithmatic achievement, but found no significant relationship between these behavior scores and achievement in spelling and reading. These predictive validity studies are very important to the development of the field as they suggest that manipulation of these behavioral variables may well result in productive changes in academic achievement.

In our own laboratory, we are exploring the convergent validity of naturalistic behavioral data by relating it to measures on similar dimensions

in the laboratory which include: a) parent and child interaction behavior in standard stimulus situations similar to those employed by Wahler (1967) and Johnson and Brown (1969), b) parent behavior in response to standard stimulus audio tapes similar in design to those used by Rothbart and Maccoby (1966) and parent behavior in standardized tasks similar to those used by Berberich (1970), and c) parent attitude and behavior rating measures on their children. Unfortunately, at this writing, most of this data has not been completely analyzed, but an overall report of this research will be forthcoming. A recent dissertation by Martin (1971), however, was devoted to studying the relationships between parent behavior in the home and parent behavior in analogue situations. By and large, the results of this research indicated no systematic relationships between the two measures. The same general findings for parents' responses to deviant and nondeviant behavior were replicated in the naturalistic and the analogue data, but correlations relating individual parental behavior in one setting with that in the other were generally nonsignificant. We don't know, of course, which, if either, of the measures represents "truth" but this study underlines the importance of seriously questioning the assumptions usually made in any analogue or modified naturalistic research. As Martin (1971) points out, these negative results are very representative of findings in other investigations where naturalistic behavior data has been compared to data collected in more artificial analogue conditions (e.g., see Fawl, 1963; Gump & Kounin, 1960; Chapanis, 1967).

Before closing this section on validity, we would like to briefly take note of the efforts of Cronbach and his associates to reconceptualize the issue of observer agreement, reliability and validity as parts of the

broader concept of generalizability. A full elaboration of generalizability theory goes far beyond the purposes of this chapter and the interested reader may be referred to several primary and secondary sources for a more complete presentation of this model (e.g., Cronbach, Rajaratnam, & Gleser, 1963; Rajaratnam, Cronbach, & Gleser, 1965; Gleser, Cronbach, & Rajaratnam, 1965; Wiggins, 1972). According to this generalizability view, the concerns of observer agreement, reliability and validity all boil down to a concern for the extent to which an obtained score is generalizable to the "universe" to which the researcher wishes the score to apply. Once an investigator is able to specify this "universe," he should be able to specify and test the relevant sources of possible threat to generalizability. In a typical naturalistic observational study, for example, we would usually at least want to know the generalizability of data across a) observers, b) occasions in the same setting, and c) settings. Through the generalizability model, each of these sources of variance could be explored in a factorial design and their contribution analyzed within an analysis-of-variance model. This model is particularly appealing because it provides for simultaneous assessment of the extent of various sources of "error" which could limit generalizability. In spite of the advantages of this factorial model, there are few precedents for its use. This is probably more the result of practical problems rather than a resistance to this intellectually appealing and theoretically sound model. Even if one were to restrict himself to the three sources of variance outlined above, the resulting generalizability study would, for most useful purposes, be a formidable project, indeed. Projects of this kind appear to us, however, to be well worth doing and we can probably expect to see more investigations which employ this generalizability model.

It should be pointed out at this point that the generalizability study outlined above does not really speak to the traditional validity requirement as succinctly defined by Campbell and Fiske (1969): "Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods." As stated earlier, to fulfill this requirement, one must provide evidence of some form of convergent validity by the use of methods other than direct behavioral observation. The generalizability model can, theoretically, handle any factor of this type under the heading of methods or "conditions," but the analysis-of-variance model employed requires a factorial design. Thus, it would seem extremely difficult and sometimes impossible to integrate factorially other methods of testing or rating in a design which encompassed the three variables outlined above: observers, occasions and settings. As a result of these considerations, we question the extent to which one generalizability study, at least in this area of research, can fulfill all the requirements of observer agreement, validity, and reliability which we view as so important. Rather, it is likely that multiple analyses will still be necessary to sufficiently establish all of the methodological requirements we have outlined for naturalistic observational data. These multiple analyses may, of course, involve analyses of variance in a generalizability model or correlational analyses as traditionally employed.

Krantz (1971) points out that the basic controversy over group vs. individual subject designs has contributed largely to the development of the mutual isolation of operant and nonoperant psychology. Since the measurement of reliability and convergent validity is typically based on correlations across a group of subjects, the operant psychologist may feel that these are alien concepts which have no relevance for his research. We would dispute

this view on the following logical grounds. Reliability involves the requirement for consistency in measurement and without some minimal level of such consistency, there can be no demonstration of functional relationships between the dependent variable and the independent variable. Efforts are currently underway to discover statistical procedures for establishing reliability estimates for the single case (e.g., see Jones, 1972). Any operant study which involves repeating manipulative procedures on more than one subject can be used for reliability assessment by traditional methods. Once such reliability is established, either for the individual case or for a group, we can be much more confident in the data and its meaning. Validity involves the requirement of convergence among different methods in measuring the same behavioral dimension. Where the validity of a measurement procedure has been previously established for a group, we can use it with more confidence in each individual case. Where it has not, it is still possible to explore for convergence in a single case. We can simply see, for example, if the child who shows high rates of aggressive behavior is perceived as aggressive by significant others. This procedure may be done with some precision if normative data is available on the measures used in the single case. Thus, with normative data available one can explore the position of the single case on the distribution of each measurement instrument. One could see, for example, if the child who is perceived to be among the top 5% in aggressiveness actually shows aggressive behavior at a rate higher than 95% of his peers. The requirements of reliability and validity are logically sound ones which transcent experimental method and means of calculation.

These methodological issues, like all others presented in this chapter, are highly relevant for behavioral research, even though they may at first

seem alien to it as the products of rival schools of thought. It has been our argument that the requirements of sound methodology transcend "schools" and that the time has come for us to attend to any variables which threaten the quality, generalizability, or meaningfulness of our data. Behavioral data is the most central commonality and critical contribution of all behavior modification research. The behaviorists' contribution to the science of human behavior and to solutions of human problems will largely rest on the quality of this data base.

References

Adkins, D. A., & Johnson, S. M. An empirical approach to identifying deviant behavior in children. Paper presented at the Western Psychological Association Convention, Portland, Oregon, April 1972.

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.

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

Bales, D. F. Interaction Process Analysis. Cambridge: Addison-Wesley, 1950.

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

Barker, R. G. & Wright, H. F. Midwest and its children: The psychological ecology of an American town. New York: Row, Peterson, 1955.

Bechtel, R. B. The study of man: Human movement and architecture. Trans-Action, 1967, 4 (6), 53-56.

Berberich, J. Adult child interactions: I. Correctness of a "child" as a positive reinforcer for the behavior of adults. Unpublished manuscript, University of Washington, 1970.

Bernal, M. E., Gibson, D. M., William, D. E., & Pesses, D. I. A device for recording automatic audio tape recording. Journal of Applied Behavior Analysis, 1971, 4 (2), 151-156.

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

Bolstad, O. D., & Johnson, S. M. Self-regulation in the modification of disruptive classroom behavior. Journal of Applied Behavior Analysis, 1972, in press.

Browning, R. M., & Stover, D. O. Behavior modification in child treatment: An experimental and clinical approach. Chicago: Aldine Atherton, Inc., 1971.

Campbell, D. T., & Fiske, D. Convergent and discriminant validation by the multi-trait, multi-method matrix. Psychological Bulletin, 1959, 56, 81-105.

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

Chapanis, A. The relevance of laboratory studies to preschool situations. Ergonomics, 1967, 10, 557-577.

Clairborn, W. L. Expectancy effects in the classroom: A failure to replicate. Journal of Educational Psychology, 1969, 60, 377-383.

Cobb, J. A. The relationship of observable classroom behaviors to achievement of fourth grade pupils. Unpublished doctoral dissertation, University of Oregon, Eugene, 1969.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generalizability: Liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

DeMaster, B., & Reid, J. B. Effects of feedback procedures in maintaining observer reliability. In preparation, Oregon Research Institute, Eugene, 1972.

D'Heurle, A., Mellinger, J. C., & Haggard, E. A. Personality, intellectual, and achievement patterns in gifted children. Psychological Monographs: General and Applied, 1959, 73 (13, Whole No. 483).

Eyberg, S. An outcome study of child-family intervention: The effects of contingency contracting and order of treated problems. Unpublished doctoral dissertation, University of Oregon, Eugene, 1972.

Fawl, C. Disturbances experienced by children in their natural habitats. In R. G. Barker (Ed.), The stream of behavior. New York: Appleton-Century-Crofts, 1963. Pp. 99-126.

Fode, K. L. The effect of experimenters' and subjects' anxiety and social desirability on experimenter outcome bias. Unpublished doctoral dissertation, University of North Dakota, 1965.

Fode, K. L. The effect of non-visual and non-verbal interaction on experimenter bias. Unpublished master's thesis, University of North Dakota, 1960.

Gelfand, D. M., Gelfand, S., & Dobson, W. R. Unprogrammed reinforcement of patients' behavior in a mental hospital. Behavior Research and Therapy, 1967, 5, 201-207.

Glaser, G. C., Cronbach, L. J., & Rajaratnam, N. Generalizability of scores influenced by multiple sources of variance. Psychometrika, 1965, 30, 395-418.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Gump, R., & Kounin, J. Issues raised by ecological and "classic" research efforts. Merrill-Palmer Quarterly, 1960, 6, 145-152.

Harris, A. Observer effect on family interaction. Unpublished doctoral dissertation, University of Oregon, Eugene, 1969.

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

Horton, G. O., Larson, J. L., & Maser, A. L. The generalized reduction of student teacher disapproval behavior. Unpublished manuscript,

University of Oregon, Eugene, 1972.

Hughes, L. D. A study of the relationship of coping strength to self-control, school achievement, and general anxiety level in sixth grade pupils. Dissertation Abstracts, 1968, 28 (A) (10), 4001.

Johansson, S., Johnson, S. M., Martin, S., & Wahl, G. Compliance and non-compliance in young children: A behavioral analysis. Unpublished manuscript, University of Oregon, Eugene, 1971.

Johnson, S. M., & Brown, R. Producing behavior change in parents of disturbed children. Journal of Child Psychology and Psychiatry, 1969, 10, 107-121.

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

Johnson, S. M., Wahl, G., Martin, S., & Johansson, S. How deviant is the normal child: A behavioral analysis of the preschool child and his family. Advances in Behavior Therapy, 1972, in press.

Jones, R. R. Intraindividual stability of behavioral observations: Implications for evaluating behavior modification treatment programs. Paper presented at the meeting of the Western Psychological Association, Portland, Oregon, April 1972.

Karpowitz, D. Stimulus control in family interaction sequences as observed in the naturalistic setting of the home. Unpublished doctoral dissertation, 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 entitled "Behavior Analysis in Education," University of Kansas, Lawrence, April 1970.

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

Kent, R. The human observer: An imperfect cumulative recorder. Paper presented at the Banff Conference on Behavior Modification, Banff, Canada, March 1972.

Krantz, D. L. The separate worlds of operant and non-operant psychology. Journal of Applied Behavior Analysis, 1971, 4 (1), 61-70.

Lahaderne, H. M. Attitudinal and intellectual correlates of attention: A study of four sixth-grade classrooms. Journal of Educational Psychology, 1968, 59 (5), 320-324.

Littman, R., Pierce-Jones, J., & Stern. T. Child-parent activities in the natural setting of the home: Results of a methodological pilot study. Unpublished manuscript, University of Oregon, Eugene, 1957.

Lobitz, G., & Johnson, S. M. Normal versus deviant--Fact or fantasy? Paper presented at the Western Psychological Association Convention, Portland, April 1972.

Martin, M. F., Gelfand, D. M., & Hartmann, D. P. Effects of adult and peer observers on boys' and girls' responses to an aggressive model. Child Development, 1971, 42, 1271-1275.

Martin, S. The comparability of behavioral data in laboratory and natural settings. Unpublished doctoral dissertation, University of Oregon, Eugene, 1971.

Meyers, C. E., Attwell, A. A., & Orpet, R. E. Prediction of fifth grade achievement from kindergarten test and rating data. Educational and Psychological Measurement, 1968, 28 (2), 457-463.

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

O'Conner, R. D. Modification of social withdrawal through symbolic modeling. Journal of Applied Behavior Analysis, 1969, 2, 15-22.

Olson, W. The incidence of nervous habits in children. Journal of Abnormal and Social Psychology, 1930-31, 35, 75-92.

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

Patterson, G. R. & Cobb, J. A. Stimulus control for classes of noxious behaviors. Paper presented at the University of Iowa, May 1971, Symposium, "The control of aggression: Implications from basic research." J. F. Knutson (Ed.). Aldine Publishing, 1972, in press.

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

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

Patterson, G. R., Ray, R. S., & Shaw, D. A. Direct intervention in families of deviant children. Oregon Research Bulletin, 1969, 8.

Patterson, G. R., Ray, R. S., Shaw, D. A., & Cobb, J. A. Manual for coding family interactions, sixth revision, 1969. Available from ASIS National Auxiliary Publications Service, in care of CCM Information Service, Inc., 909 Third Avenue, New York, N. Y. 10022. Document #01234.

Patterson, G. R., & Reid, J. B. Reciprocity and coercion: Two facets of social systems. In C. Neuringer & J. Michael (Eds.), Behavior modification in clinical psychology. New York: Appleton-Century Crofts, 1970.

Patterson, G. R., & Reid, J. B. Family intervention in the homes of aggressive boys: A replication. Paper presented at the American Psychological Association Convention, Washington D. C., 1971.

Peine, H. A. Behavioral recording by parents and its resultant consequences. Unpublished master's thesis, University of Utah, 1970.

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

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

Purcell, K., & Brady, K. Adaptation to the invasion of privacy: Monitoring behavior with a miniature radio transmitter. Merrill-Palmer Quarterly, 1965, 12, 242-254.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. Generalizability of stratified-parallel tests. Psychometrika, 1965, 30, 39-56.

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

Rausch, H. L. Interaction sequences. Journal of Personality and Social Therapy, 1965, 2, 487-499.

Reid, J. B. Reciprocity in family interaction. Unpublished Doctoral Dissertation, University of Oregon, Eugene, 1967.

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

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

Romanczyk, R. G., Kint, R. W., 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, Lawrence, Kansas, May 1971.

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

Rosenthal, R. On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. American Scientist, 1963, 51, 268-283.

Rosenthal, R., & Fode, K. L. The effect of experimenter bias on the performance of the albino rat. Behavior Science, 1963, 8, 183-189.

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

Rosenthal, R., & Jacobsen, L. Teacher expectancies: Determinants of pupils IQ gains. Psychological Reports, 1966, 19, 115-118.

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.

Rosenthal, R., Persinger, G. W., Vikan-Kline, L. E., & Mulry, R. C. The role of the research assistant in the mediation of experimenter bias. Journal of Personality, 1963, 31, 313-335.

Rothbart, M., & Maccoby, E. Parents' differential reaction to sons and daughters. Journal of Personality and Social Psychology, 1966, 4, 237-243.

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

...th L. and Bolstad

Sheffield, F. D., Kaufman, E. ..., Raines, ... A Pre-experiment ... starts a controversy. Journal of Applied Social and Psychological Research, 1952, 46, 111-117.

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

Smith, E. E. Effects of threat induced by ambiguous role expectations on defensiveness and productivity in small groups. Dissertation Abstracts, 1957, 17, 3104-3105.

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

Soskin, W. F., & John, V. P. The study of spontaneous talk. In R. G. Barker (Ed.), The Stream of Behavior. New York: Appleton-Century-Crofts, 1963. Pp. 228-281.

Taplin, P. S., & Reid, J. B. Effects of instructional set and experimental influence on observer reliability. In Preparation, Oregon Research Institute, Eugene, 1972.

Thomas, D. R., Becker, W. C., & Armstrong, M. Production and elimination of disruptive classroom behavior by systematically varying teacher's behavior. Journal of Applied Behavior Analysis, 1968, 1, 35-45.

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

Verplanck, N. S. The control and the content in conversation: Reinforcement of statements of opinion. Journal of Abnormal and Social Psychology, 1955, 51, 668-676.

Wahl, G., Johnson, C. M., Martin, ., & Thomson, . An operant analysis of child-family interaction. Behavior Therapy, 1972, in press.

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

Walker, H. M., Johnson, S. M., & Hops, H. Generalization and maintenance of classroom treatment effects. Unpublished manuscript, University of Oregon, Eugene, 1972.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. Unobtrusive measures: A survey of non-reactive research in the social sciences. Chicago: Rand McNally, 1966.

Weick, K. E. Systematic observational methods. In G. Lindsey & E. Aronson (Eds.), The handbook of social psychology, 1968, 2, 357-451.

White, G. D. Effects of observer presence on mother and child behavior. Unpublished doctoral dissertation, University of Oregon, September 1972.

Wiggins, J. S. Personality and prediction: Principles of personality assessment. Reading, Mass.: Addison Wesley, 1972, in press.

Footnotes

1. The preparation of this manuscript and the research reported therein was supported by research grant MH 19633-01 from the National Institute of Mental Health. The writers would like to thank their many colleagues who contributed critical reviews of this manuscript: Robyn Dawes, Lewis Goldberg, Richard Jones, Gerald Patterson, John Reid, Carl Skindrud and Geoffry White.

2. The authors would like to credit Lee Sechrest for first suggesting this illustrative example.

3. The authors would like to credit Donald Hartman for clarifying this as the appropriate procedure for establishing the level of agreement to be expected by chance.

4. For additional justification of the use of this statistical procedure for problems of this kind, see Wiggins (1972).

5. Several behaviors which are used in the coding system are not included in the present analysis. The behaviors humiliate and dependency could not be included because they did not occur in the behavioral sample. Repeated noncompliance and temper tantrums were not used on the verbal report scale because they are subsumed in other categories (i.e., tantrums are defined as simultaneous occurrences of three or more of the following--physical negative, destructiveness, crying, yelling, etc.). Nonresponding of the child was excluded post hoc because it was clear that parents were responding to this item as ignoring rather than mere nonresponse to ongoing activity (i.e., it was a poorly-written item).

Table 1

Coded Behaviors as Ranked by Two Methods:

Parental Ratings and Negative Social Consequences*

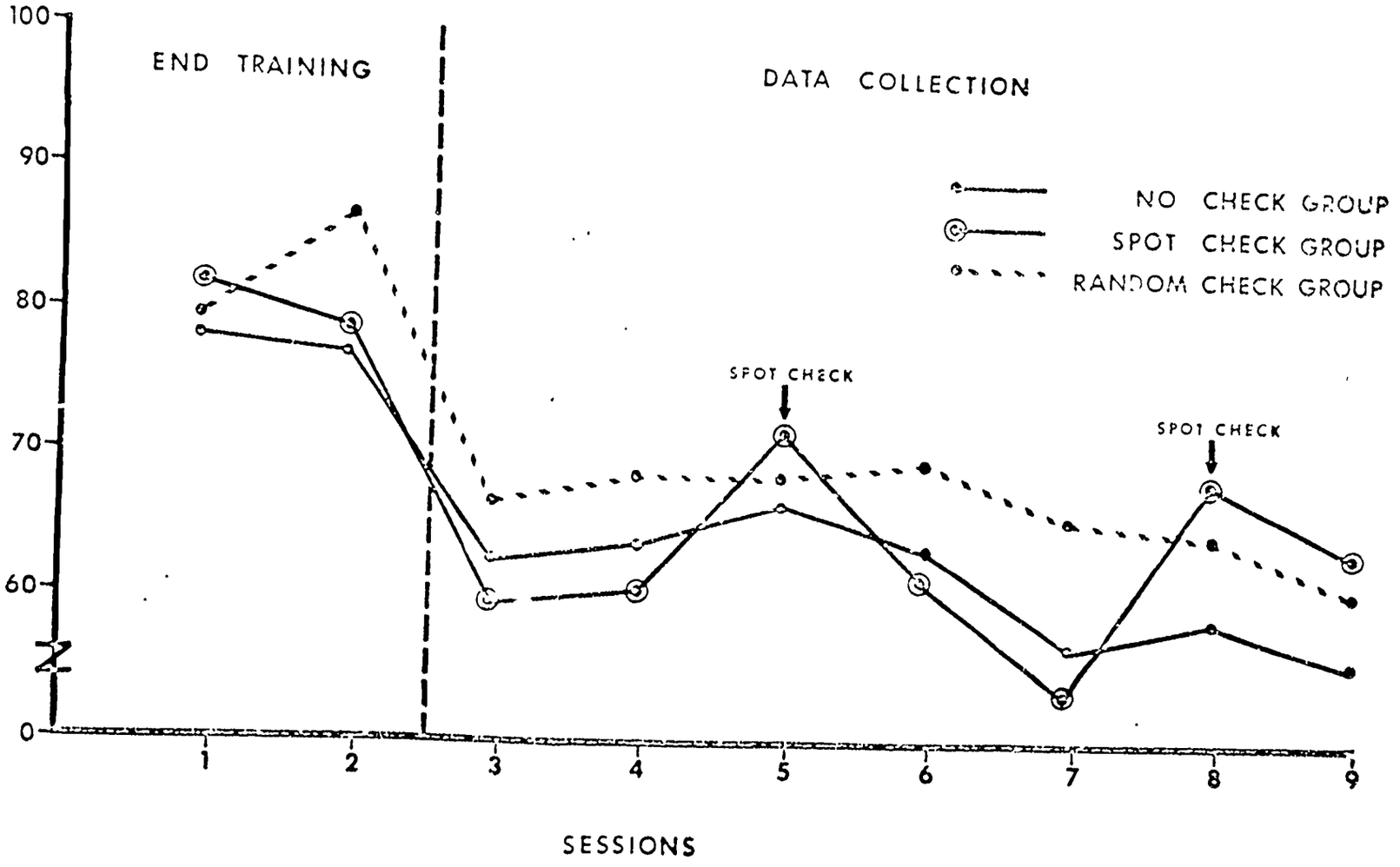
Behavior Rank by Parental Rating		Behavior Rank by Proportion of Negative Consequences	Mean Parent Rating for Behavior	Proportion of Negative Consequences to Behavior
1	Whine	13	1.056	.125
2	Physical Negative	2	1.074	.527
4	Destructive	8	1.204	.352
4	Tease	5	1.204	.382
4	Smart Talk	4	1.204	.390
6	Aversive Command	3	1.208	.428
7	Noncompliance	12	1.278	.175
8	High Rate	16	1.307	.664
9	Ignore	11	1.379	.205
10	Yell	10	1.537	.215
11	Demand Attention	15	1.611	.083
12	Negativism	6	1.635	.375
13	Command Negative	1	1.835	.669
14	Disapproval	9	1.870	.235
15	Cry	14	1.962	.097
16	Indulgence	22	2.093	.027
17	Command Prime	27.5	2.132	.000
18	Receive	18	2.222	.050
19	Talk	23	2.278	.020
20	Command	7	2.296	.355
21	Attention	25	2.556	.013
22	Touch	20	2.648	.043
23	Independent Activity	26	2.704	.005
24	Physical Positive	21	2.741	.034
25	Comply	17	2.759	.053
26	Laugh	19	2.778	.044
27	Nonverbal Interaction	24	2.833	.012
28	Approval	27.5	2.926	.000

*Spearman Rank-order correlation between columns 1 & 2 = .73 ($p < .01$).

Figure 1

[illegible] with [illegible] [illegible] [illegible] [illegible]

(Taplin and Reid, [illegible])



Johnson and Bolstad